Wk14-2 : 연관규칙 분석 II

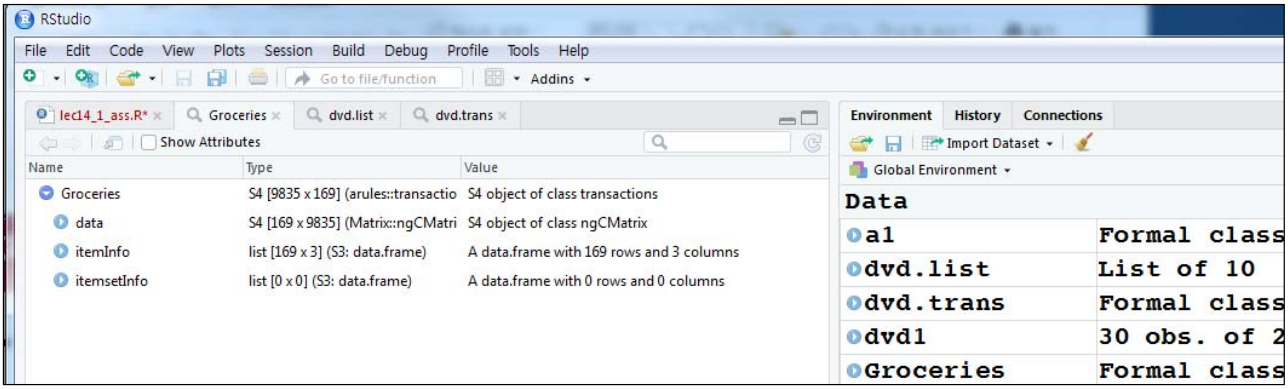
1. 연관규칙 - 데이터 설명 (Groceries)

Groceries data ("arules" package에 탑재되있는 데이터)

- data("Groceries")으로 불러옴
- 실제 식료품점에서 1개월(30일)치의 transaction 데이터
- 9835트랜잭션 / 169항목
- 밀도가 0.026%라고 되어 있는데, 9835*169 cell 중에서 0.026%의 cell에 거래가 발생해 숫자가 차 있다는 뜻임
- Element(itemset/transaction) length distribution : 하나의 거래 장바구니(row 1개 당)에 item의 개수 별로 몇번의 거래가 있었는지 나타냄

1. 연관규칙 - 데이터 설명 (Groceries)

- **Groceries data – transaction 데이터**



- **transaction 9835개**
- **items수 169개**

14-2 연관규칙 분석 II

1. 연관규칙 - 데이터 설명 (Groceries)

- **Groceries (데이터이름, "arules"package에 탑재되있는 데이터)**

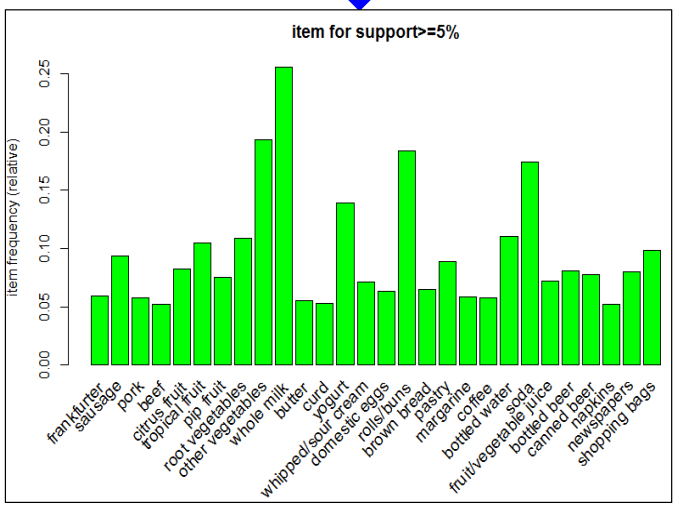
```
> summary(Groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt          (Other)
            2513             1903             1809             1715             1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14   14    9   11
  22   23   24   26   27   28   29   32
   4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

가장 많이 거래된 항목

2. 연관규칙 – visualization (지지도)

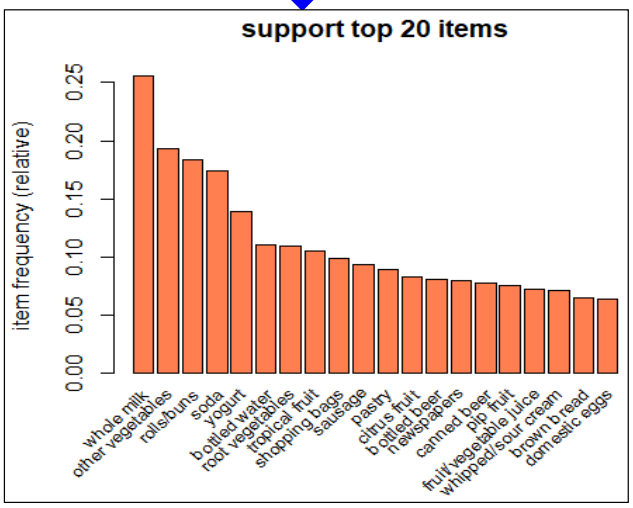
- 그래프로 표현한 연관규칙 (지지도)

```
#지지도 5%이상의 item 막대 그래프
itemFrequencyPlot(Groceries,support=0.05,main="items for support>= 5%", col="green")
```



상위-> 하위 정렬

```
#지지도 상위 20개 막대 그래프
itemFrequencyPlot(Groceries,topN=20,main="support top 20 items", col="coral")
```



3. 연관규칙 분석결과 – Groceries 데이터

- 연관규칙분석

```
#association rule analysis
Grocery_rule<-apriori(data=Groceries,
                      parameter = list(support=0.05,
                                       confidence = 0.20,
                                       minlen = 2))
```

```
> Grocery_rule<-apriori(data=Groceries,
+                       parameter = list(support=0.05,
+                                        confidence = 0.20,
+                                        minlen = 2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.2    0.1    1 none FALSE            TRUE       5    0.05      2
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 491

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [28 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [6 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> Grocery_rule
set of 6 rules
```

- 위의 support, confidence 와 length는 minimum값으로 너무 높게 잡으면 연관규칙이 분석이 안됨

3. 연관규칙 분석결과 – Groceries 데이터

- 연관규칙 조회 및 평가

```
#analyzing result
summary(Grocery_rule)
inspect(Grocery_rule)
```

```
> inspect(Grocery_rule)
    lhs                  rhs                support    confidence lift     count
[1] {yogurt}          => {whole milk}       0.05602440 0.4016035  1.571735 551
[2] {whole milk}      => {yogurt}           0.05602440 0.2192598  1.571735 551
[3] {rolls/buns}      => {whole milk}       0.05663447 0.3079049  1.205032 557
[4] {whole milk}      => {rolls/buns}       0.05663447 0.2216474  1.205032 557
[5] {other vegetables} => {whole milk}      0.07483477 0.3867578  1.513634 736
[6] {whole milk}      => {other vegetables} 0.07483477 0.2928770  1.513634 736
```

```
> summary(Grocery_rule)
set of 6 rules

rule length distribution (lhs + rhs):sizes
2
6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.05602   Min.   :0.2193   Min.   :1.205   Min.   :551.0
 1st Qu.:0.05618   1st Qu.:0.2395   1st Qu.:1.282   1st Qu.:552.5
 Median :0.05663   Median :0.3004   Median :1.514   Median :557.0
 Mean   :0.06250   Mean   :0.3050   Mean   :1.430   Mean   :614.7
 3rd Qu.:0.07028   3rd Qu.:0.3670   3rd Qu.:1.557   3rd Qu.:691.2
 Max.   :0.07483   Max.   :0.4016   Max.   :1.572   Max.   :736.0

mining info:
      data ntransactions support confidence
 Groceries          9835    0.05        0.2
```

- 향상도 최소값이 1보다 큰 것을 알 수 있음
- 6개의 rule이 item 2개로 구성되어 있음

3. 연관규칙 분석결과 – Groceries 데이터

- 연관규칙-향상도(Lift)순서로 정렬

```
#sorting result by lift
inspect(sort(Grocery_rule,by="lift"))
```

```
> inspect(sort(Grocery_rule,by="lift"))
    lhs                  rhs
[1] {yogurt}          => {whole milk}
[2] {whole milk}      => {yogurt}
[3] {other vegetables} => {whole milk}
[4] {whole milk}      => {other vegetables}
[5] {rolls/buns}      => {whole milk}
[6] {whole milk}      => {rolls/buns}
    support    confidence lift     count
[1] 0.05602440 0.4016035  1.571735 551
[2] 0.05602440 0.2192598  1.571735 551
[3] 0.07483477 0.3867578  1.513634 736
[4] 0.07483477 0.2928770  1.513634 736
[5] 0.05663447 0.3079049  1.205032 557
[6] 0.05663447 0.2216474  1.205032 557
```

- Sort()함수를 통해 분석가가 보고자 하는 기준으로 정렬하여 보는 것도 가능

3. 연관규칙 분석결과 – Groceries 데이터

- 연관규칙-품목별 연관성 탐색

```
rule_interest<-subset(Grocery_rule, items %in% c("yogurt","whole milk"))
inspect(rule_interest)
```

```
> inspect(rule_interest)
    lhs                    rhs
[1] {yogurt}            => {whole milk}
[2] {whole milk}        => {yogurt}
[3] {rolls/buns}        => {whole milk}
[4] {whole milk}        => {rolls/buns}
[5] {other vegetables}  => {whole milk}
[6] {whole milk}        => {other vegetables}
    support     confidence lift     count
[1] 0.05602440 0.4016035  1.571735 551
[2] 0.05602440 0.2192598  1.571735 551
[3] 0.05663447 0.3079049  1.205032 557
[4] 0.05663447 0.2216474  1.205032 557
[5] 0.07483477 0.3867578  1.513634 736
[6] 0.07483477 0.2928770  1.513634 736
```

- Sort()함수를 통해 분석가가 보고자 하는 기준으로 정렬하여 보는 것도 가능
- Subset()함수를 통해 원하는 item이 포함된 연관규칙만 선별해서 보는 것도 가능
- %in%, %pin%, %ain%을 이용해 다양한 결과 도출

4. 연관규칙 분석결과 저장

- 연관규칙결과를 data.frame으로 저장

```
# save as dataframe
Grocery_rule_df<-as(Grocery_rule,"data.frame")
Grocery_rule_df
```

```
> # save as dataframe
> Grocery_rule_df<-as(Grocery_rule,"data.frame")
> Grocery_rule_df
                              rules     support confidence     lift count
1           {yogurt} => {whole milk} 0.05602440  0.4016035 1.571735   551
2           {whole milk} => {yogurt} 0.05602440  0.2192598 1.571735   551
3       {rolls/buns} => {whole milk} 0.05663447  0.3079049 1.205032   557
4       {whole milk} => {rolls/buns} 0.05663447  0.2216474 1.205032   557
5 {other vegetables} => {whole milk} 0.07483477  0.3867578 1.513634   736
6 {whole milk} => {other vegetables} 0.07483477  0.2928770 1.513634   736
```

4. 연관규칙 분석결과 저장

- 연관규칙결과 저장

```
#saving results as csv file
write(Grocery_rule, file="Grocery_rule.csv",
      sep=",",
      quote=TRUE,
      row.names=FALSE)
```

	A	B	C	D	E
1	rules	support	confidence	lift	count
2	{yogurt} => {whole milk}	0.056024	0.401603	1.571735	551
3	{whole milk} => {yogurt}	0.056024	0.21926	1.571735	551
4	{rolls/buns} => {whole milk}	0.056634	0.307905	1.205032	557
5	{whole milk} => {rolls/buns}	0.056634	0.221647	1.205032	557
6	{other vegetables} => {whole milk}	0.074835	0.386758	1.513634	736
7	{whole milk} => {other vegetables}	0.074835	0.292877	1.513634	736