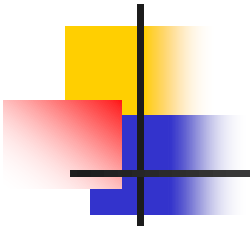
ARM (Association Rule Mining, 연관규칙탐사)

R을 활용한 연관규칙탐사 실습



Mining Associations with R

- Using "arules" and "arulesViz" package
- Related functions
  - read.transactions() - arules package
  - as() - arules package
  - image() - arules package
  - transactionInfo() - arules package
  - inspect() - arules package
  - itemFrequency() / itemFrequencyPlot() - arules package
  - apriori() - arules package
  - summary() - arules package
  - write() - arules package
  - plot() - arulesViz package

Mining Associations using arules

```
install.packages("arules")
library(arules); library(dplyr)

tr <- read.delim("dataTransactions.tab", stringsAsFactors=FALSE)
head(tr)
```

	datetime	custid	store	product	brand	corner	import	amount	installment
1	2000-05-01 10:43	18313	신촌점	4.104840e+12	샤넬	화장품	1	113000	3
2	2000-05-01 11:00	18313	신촌점	2.700000e+12	식품	일반식품	0	91950	3
3	2000-05-01 11:33	27222	신촌점	4.545371e+12	까사미아	가구	0	598000	3
4	2000-05-01 11:43	27222	신촌점	4.500860e+12	대아통상	기타	0	20100	1
5	2000-05-01 11:53	27222	신촌점	4.538130e+12	토이플러스	문화완구	0	24000	1
6	2000-05-01 12:00	27222	신촌점	4.406010e+12	베베	유아동복	0	28000	1

```
tr.filter <- tr %>%
      filter(!(corner %in% c("일반식품","화장품"))) %>%
      distinct(custid, corner)

trans <- as(split(tr.filter$corner, tr.filter$custid), "transactions")
trans
```

```
transactions in sparse format with
 487 transactions (rows) and
 24 items (columns)
```

```
# trans <- read.transactions("dataTransactions.tab", format = "single",
                             sep="\t", cols = c(2,6), skip=1)
```

Mining Associations using arules

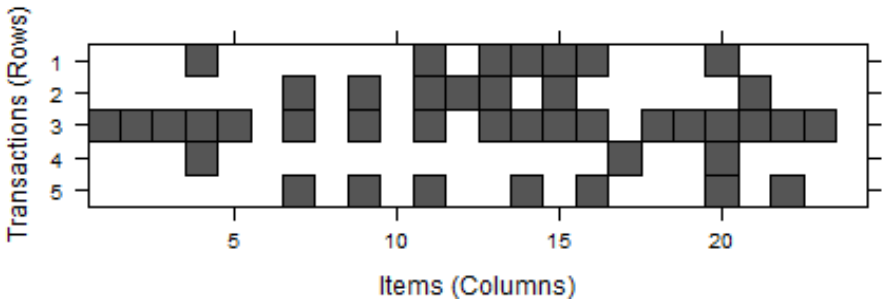
- `inspect(trans[1:2])`

```
    items                                                                transactionID
[1] {니트단품,스포츠,영캐주얼,유니캐주얼,유아동복,장신구,캐릭터캐주얼} 10070
[2] {문화완구,섬유,스포츠,엘레강스캐주얼,영캐주얼,유아동복,타운모피}   10139
```

- `transactionInfo(trans[size(trans) > 20])`

```
    transactionID
84          15968
420         42322
```

- `image(trans[1:5])`



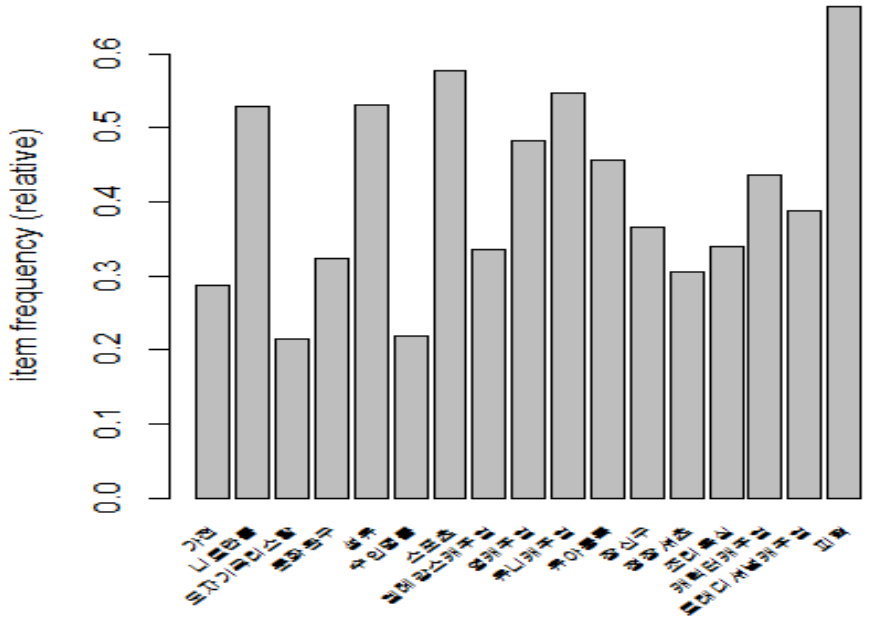
- `image(sample(trans, 100, replace=FALSE), main="matrix diagram")`

Mining Associations using arules

- `itemFrequency(trans, type="absolute")`
- `itemFrequency(trans)[order(itemFrequency(trans), decreasing = TRUE)]`

```
        피혁         스포츠      유니캐주얼          섬유       니트단품      영캐주얼
0.663244353 0.577002053 0.546201232 0.531827515 0.529774127 0.482546201
    유아동복 캐릭터캐주얼 트래디셔널캐주얼      장신구 조리욕실 엘레강스캐주얼
0.455852156 0.435318275 0.388090349 0.365503080 0.338809035 0.334702259
    문화완구      정장셔츠          가전      수입명품 도자기크리스탈          기타
0.324435318 0.305954825 0.287474333 0.219712526 0.215605749 0.195071869
디자이너부띠끄      침구수예          가구      타운모피      행사장 생활용품
0.151950719 0.151950719 0.084188912 0.067761807 0.026694045 0.006160164
```

- `itemFrequencyPlot(trans, support=0.2, cex.names=0.8)`



Mining Associations using arules

```
rules <- apriori(trans, parameter=list(support=0.2, confidence=0.8))
summary(rules)
```

```
set of 70 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
 1 40 29

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     3.0     3.0     3.4     4.0     4.0

summary of quality measures:
    support         confidence          lift
 Min.   :0.2012   Min.   :0.8000   Min.   :1.233
 1st Qu.:0.2115   1st Qu.:0.8182   1st Qu.:1.283
 Median :0.2259   Median :0.8413   Median :1.353
 Mean   :0.2341   Mean   :0.8444   Mean   :1.383
 3rd Qu.:0.2464   3rd Qu.:0.8624   3rd Qu.:1.463
 Max.   :0.3265   Max.   :0.9160   Max.   :1.696

mining info:
  data ntransactions support confidence
 trans           487     0.2        0.8
```

```
# rules <- apriori(trans, parameter=list(support=0.2, confidence=0.8),
                   appearance=list(rhs="스포츠",default="lhs"))
```

Mining Associations using arules

Apriori only creates rules with one item in the RHS (Consequent)!

- `inspect(rules)`
- `inspect(sort(rules, by="lift")[1:30])`

```
      lhs                                  rhs          support    confidence lift
[1]   {니트단품,섬유,유니캐주얼}           => {영캐주얼}   0.2032854 0.8181818  1.695551
[2]   {니트단품,스포츠,유니캐주얼}         => {영캐주얼}   0.2114990 0.8174603  1.694056
[3]   {영캐주얼,트래디셔널캐주얼}          => {니트단품}   0.2094456 0.8429752  1.591197
[4]   {섬유,영캐주얼,피혁}                 => {니트단품}   0.2258727 0.8396947  1.585005
[5]   {섬유,영캐주얼,유니캐주얼}           => {니트단품}   0.2032854 0.8250000  1.557267
[6]   {스포츠,영캐주얼,피혁}               => {유니캐주얼} 0.2381930 0.8467153  1.550189
[7]   {니트단품,스포츠,영캐주얼}           => {유니캐주얼} 0.2114990 0.8442623  1.545698
[8]   {섬유,스포츠,영캐주얼}               => {니트단품}   0.2012320 0.8099174  1.528797
[9]   {니트단품,장신구}                    => {섬유}       0.2032854 0.8114754  1.525824
[10]  {니트단품,트래디셔널캐주얼}          => {섬유}       0.2135524 0.8062016  1.515908
[11]  {영캐주얼,트래디셔널캐주얼}          => {유니캐주얼} 0.2053388 0.8264463  1.513080
[12]  {니트단품,스포츠,영캐주얼}           => {섬유}       0.2012320 0.8032787  1.510412
[13]  {섬유,영캐주얼}                      => {니트단품}   0.2546201 0.8000000  1.510078
[14]  {니트단품,유니캐주얼,피혁}           => {섬유}       0.2258727 0.8029197  1.509737
[15]  {스포츠,영캐주얼}                    => {유니캐주얼} 0.2710472 0.8198758  1.501051
[16]  {섬유,스포츠,영캐주얼}               => {유니캐주얼} 0.2032854 0.8181818  1.497949
[17]  {섬유,영캐주얼,피혁}                 => {유니캐주얼} 0.2197125 0.8167939  1.495408
[18]  {유니캐주얼,트래디셔널캐주얼,피혁} => {스포츠}       0.2094456 0.8500000  1.473132
[19]  {영캐주얼,트래디셔널캐주얼}          => {스포츠}     0.2053388 0.8264463  1.432311
[20]  {섬유,영캐주얼,유니캐주얼}           => {스포츠}     0.2032854 0.8250000  1.429804
[21]  {니트단품,영캐주얼,유니캐주얼}       => {스포츠}     0.2114990 0.8240000  1.428071
[22]  {조리욕실,피혁}                      => {스포츠}     0.2053388 0.8196721  1.420571
[23]  {섬유,영캐주얼,피혁}                 => {스포츠}     0.2197125 0.8167939  1.415582
[24]  {섬유,유니캐주얼,피혁}               => {스포츠}     0.2361396 0.8156028  1.413518
[25]  {니트단품,트래디셔널캐주얼}          => {스포츠}     0.2156057 0.8139535  1.410660
[26]  {유니캐주얼,트래디셔널캐주얼}        => {스포츠}     0.2299795 0.8115942  1.406571
[27]  {니트단품,유니캐주얼,피혁}           => {스포츠}     0.2279261 0.8102190  1.404187
[28]  {영캐주얼,유니캐주얼,피혁}           => {스포츠}     0.2381930 0.8055556  1.396105
[29]  {트래디셔널캐주얼,피혁}              => {스포츠}     0.2628337 0.8050314  1.395197
[30]  {섬유,트래디셔널캐주얼}              => {스포츠}     0.2217659 0.8000000  1.386477
```

Mining Associations using arules

```
> rules.target <- subset(rules, rhs %in% "스포츠" & lift > 1.4)
> inspect(sort(rules.target,by="confidence"))
```

```
     lhs                                  rhs        support   confidence lift
[1]  {유니캐주얼,트래디셔널캐주얼,피혁} => {스포츠} 0.2094456 0.8500000  1.473132
[2]  {영캐주얼,트래디셔널캐주얼}         => {스포츠} 0.2053388 0.8264463  1.432311
[3]  {섬유,영캐주얼,유니캐주얼}         => {스포츠} 0.2032854 0.8250000  1.429804
[4]  {니트단품,영캐주얼,유니캐주얼}     => {스포츠} 0.2114990 0.8240000  1.428071
[5]  {조리욕실,피혁}                     => {스포츠} 0.2053388 0.8196721  1.420571
[6]  {섬유,영캐주얼,피혁}                => {스포츠} 0.2197125 0.8167939  1.415582
[7]  {섬유,유니캐주얼,피혁}              => {스포츠} 0.2361396 0.8156028  1.413518
[8]  {니트단품,트래디셔널캐주얼}         => {스포츠} 0.2156057 0.8139535  1.410660
[9]  {유니캐주얼,트래디셔널캐주얼}       => {스포츠} 0.2299795 0.8115942  1.406571
[10] {니트단품,유니캐주얼,피혁}          => {스포츠} 0.2279261 0.8102190  1.404187
```

```
> rule.interest <- subset(rules, items %in% c("장신구", "섬유"))
> inspect(rule.interest[1:10)
```

```
     lhs                                  rhs        support   confidence lift
[1]  {영캐주얼,장신구}                   => {피혁}   0.2012320 0.8596491  1.296127
[2]  {유니캐주얼,장신구}                 => {피혁}   0.2012320 0.8521739  1.284857
[3]  {니트단품,장신구}                   => {섬유}   0.2032854 0.8114754  1.525824
[4]  {니트단품,장신구}                   => {피혁}   0.2114990 0.8442623  1.272928
[5]  {섬유,장신구}                       => {피혁}   0.2156057 0.8333333  1.256450
[6]  {스포츠,장신구}                     => {피혁}   0.2114990 0.8306452  1.252397
[7]  {니트단품,트래디셔널캐주얼} => {섬유}   0.2135524 0.8062016  1.515908
[8]  {섬유,트래디셔널캐주얼}     => {스포츠} 0.2217659 0.8000000  1.386477
[9]  {섬유,트래디셔널캐주얼}     => {피혁}   0.2340862 0.8444444  1.273203
[10] {섬유,유아동복}                     => {피혁}   0.2114990 0.8174603  1.232518
```

Mining Associations using arules

- `write(rules.target, file="arules.csv", sep=",", row.name=F)`

- `install.packages("pmml")`
- `library(pmml)`

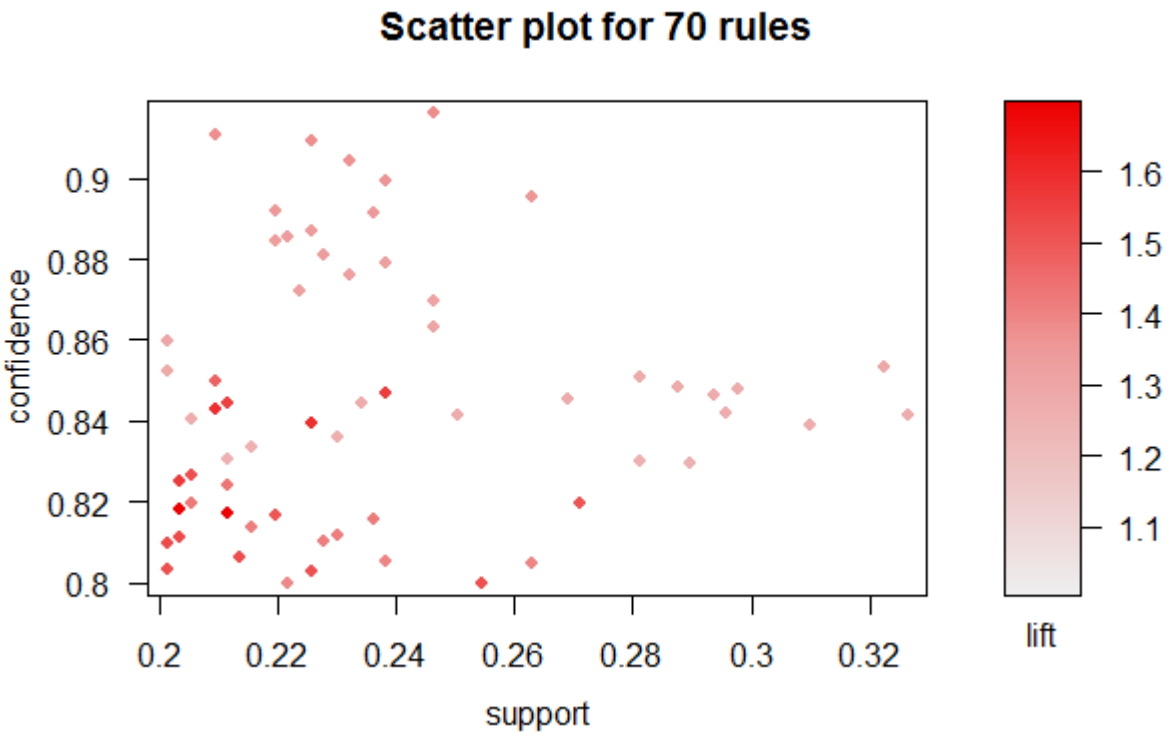
- `write.PMML(rules.target, file = "arules.xml")`

```
<Itemset id="6" numberOfItems="3">
 <ItemRef itemRef="4"/>
 <ItemRef itemRef="13"/>
 <ItemRef itemRef="14"/>
</Itemset>
<Itemset id="7" numberOfItems="3">
 <ItemRef itemRef="9"/>
 <ItemRef itemRef="13"/>
 <ItemRef itemRef="14"/>
</Itemset>
<Itemset id="8" numberOfItems="3">
 <ItemRef itemRef="9"/>
 <ItemRef itemRef="13"/>
 <ItemRef itemRef="23"/>
</Itemset>
<Itemset id="9" numberOfItems="3">
 <ItemRef itemRef="4"/>
 <ItemRef itemRef="14"/>
 <ItemRef itemRef="23"/>
</Itemset>
<Itemset id="10" numberOfItems="3">
 <ItemRef itemRef="9"/>
 <ItemRef itemRef="14"/>
 <ItemRef itemRef="23"/>
</Itemset>
<Itemset id="11" numberOfItems="1">
 <ItemRef itemRef="11"/>
</Itemset>
<AssociationRule support="0.205338809034908" confidence="0.819672131147541" lift="1.42057056181086" antecedent="1" consequent="11"/>
<AssociationRule support="0.205338809034908" confidence="0.826446280991736" lift="1.43231081438781" antecedent="2" consequent="11"/>
<AssociationRule support="0.229979466119097" confidence="0.811594202898551" lift="1.40657073598432" antecedent="3" consequent="11"/>
<AssociationRule support="0.215605749486653" confidence="0.813953488372093" lift="1.41065960440288" antecedent="4" consequent="11"/>
<AssociationRule support="0.209445585215606" confidence="0.85" lift="1.47313167259786" antecedent="5" consequent="11"/>
<AssociationRule support="0.211498973305955" confidence="0.824" lift="1.42807117437722" antecedent="6" consequent="11"/>
<AssociationRule support="0.203285420944559" confidence="0.825" lift="1.42980427046263" antecedent="7" consequent="11"/>
<AssociationRule support="0.219712525667351" confidence="0.816793893129771" lift="1.41558229876939" antecedent="8" consequent="11"/>
<AssociationRule support="0.227926078028747" confidence="0.81021897810219" lift="1.40418733927319" antecedent="9" consequent="11"/>
<AssociationRule support="0.236139630390144" confidence="0.815602836879433" lift="1.41351808384443" antecedent="10" consequent="11"/>
</AssociationModel>
</PMML>
```

- `# rules <- read.PMML('arules.xml')`

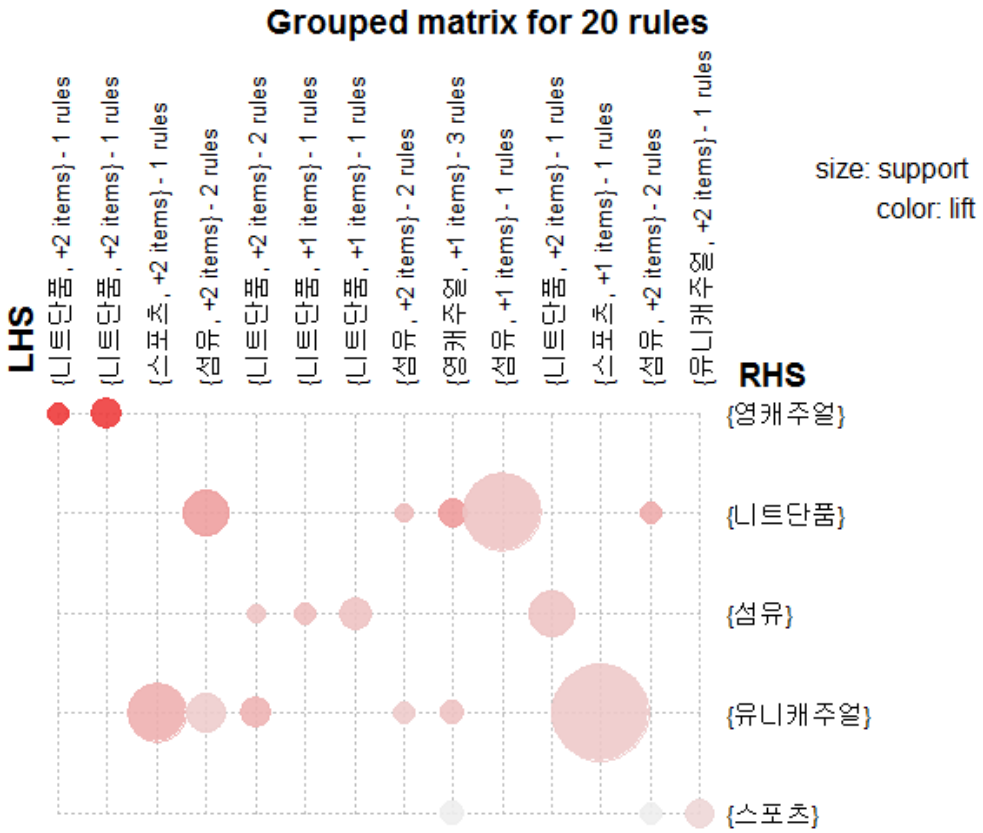
Visualizing Rules using arulesViz

- `install.packages("arulesViz")`
- `library(arulesViz)`
- `plot(rules)`



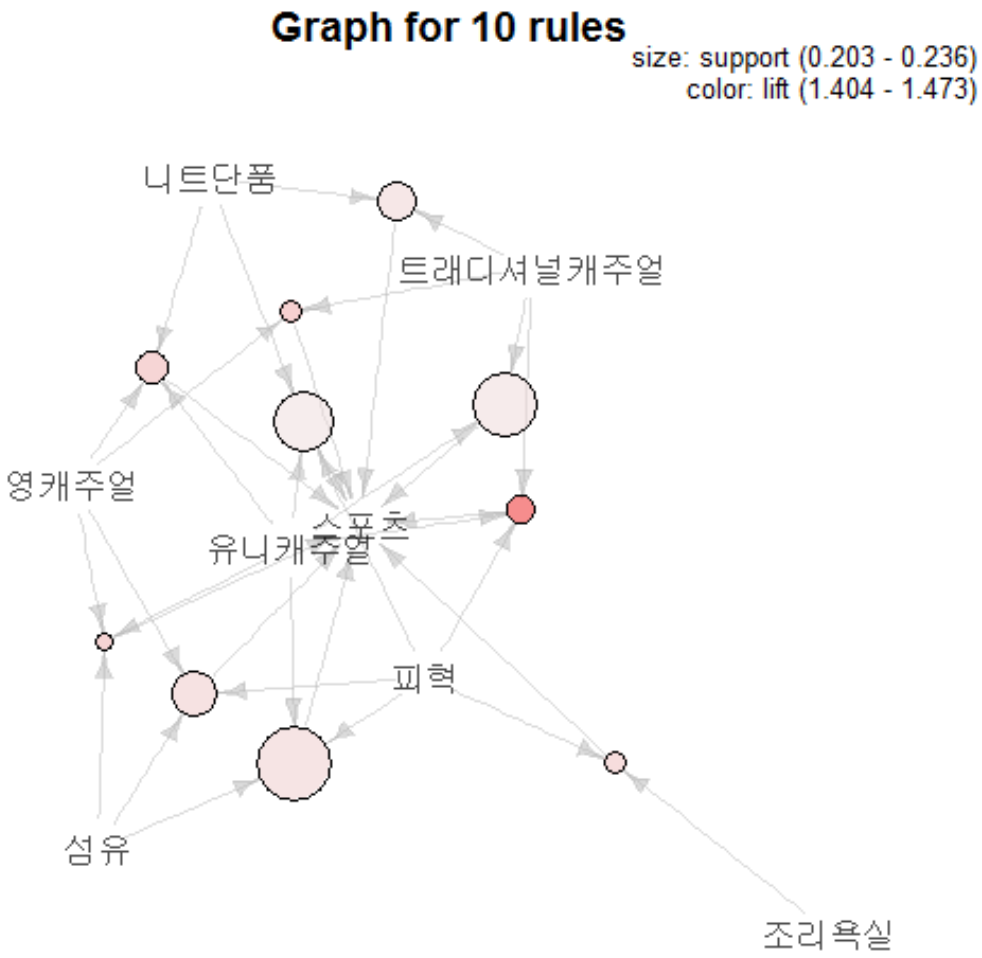
Visualizing Rules using arulesViz

- `plot(sort(rules, by = "lift")[1:20], method = "grouped")`



Visualizing Rules using arulesViz

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



Exercise

- Problem: 여성쇼핑몰 C사는 반응률이 높은 교차판매전략을 기획하기 위해 고객의 구매패턴(장바구니)을 R의 arules 패키지를 이용하여 분석하고자 한다.
- Data: shoppingmall.txt - 여성쇼핑몰 C사 고객 786명의 10가지 구매품목에 대한 거래이력

No.	변수 이름	변수 설명	변수 유형
1	ID	고객 고유번호	수치형
2	Heel	해당 상품 구매 여부 (1, 0)	범주형
3	Tee		
4	Skirt		
5	Knit		
6	Jacket		
7	Jewelry		
8	Coat		
9	Flat		
10	Shorts		
11	Blous		

- Hint:

```
data <- read.delim("shoppingmall.txt", stringsAsFactors=FALSE)
st <- as.matrix(data[,-1])
trans <- as(st, "transactions")
```

개인과제 #2 – 12월03일 제출

❖ 과제내용

- 로또(lotto)복권은 복권에 1에서부터 보통 49까지 숫자 중에서 6개를 골라 써놓고 추첨을 통해 당첨번호(역시 6개)와 일치하는 개수에 따라 등수를 정하는 복권이다. lottoData.csv에는 1회차부터 591회차까지의 로또복권 당첨번호가 아래와 같은 형식으로 저장되어있다. (730회차까지의 데이터를 반드시 추가해야 함)

lottoData.csv		
필드명	데이터형식	설명
seq	numeric	회차
N1 ~ N6	numeric	6개 당첨번호

- 위의 데이터를 사용하여 연관규칙탐사를 수행하고 분석결과를 기반으로 당첨번호를 예측하기 위한 구체적인 방안을 제시하시오.
- Hint: apriori(trans, parameter=list(support=???, target="frequent itemsets"))

❖ 제출방법

- 가상대학 과제관리를 통해 제출해야 함.
- 분석보고서(*.PPT 또는 *.PDF)와 분석코드(*.R)를 같이 제출할 것.
- 각 화일명은 본인의 이름으로 할 것.