

	<b>단위별 학습내용 (Week14)</b>
<b>wk14-1</b>	<b>연관규칙 분석 I</b>
<b>wk14-2</b>	<b>연관규칙 분석 II</b>
<b>wk14-3</b>	<b>순차적 패턴분석</b>

## Wk14-1 : 연관규칙 분석 I (Association Rule Analysis)

# 1. 연관규칙

14-1 연관규칙 분석 I

## • 연관규칙 (Association Rule)

- 대용량 데이터베이스의 트랜잭션에서 빈번하게 발생하는 패턴을 발견
- 거래간의 상호 관련성을 분석



## • 연관규칙 예시

- 신발을 구매하는 고객의 10%는 양말을 동시에 구입한다.
- 빵과 우유를 구매한 고객의 50%가 주스도 함께 구매한다.

# 1. 연관규칙

14-1 연관규칙 분석 I

- 시장바구니(market basket) : 고객이 구매한 물품에 대한 정보 (구매 시기, 지불 방법, 매장정보 포함)
- 트랜잭션(transaction) : 고객이 거래한 정보를 하나의 트랜잭션
- 시장바구니 분석(market basket analysis) : 시장바구니 데이터로부터 연관규칙을 탐색 분석

Ex) Grocery Point-Of-Sale Transactions

customer	items
1	orange juice, banana
2	orange juice, milk
3	detergent, window cleaner



## 2. 연관규칙 평가 척도

14-1 연관규칙 분석 I

- 연관규칙을 평가하기 위해 지지도(support), 신뢰도(Confidence), 향상도(Lift)를 사용

지지도 (Support)	$\frac{\text{A와 B를 동시에 포함하는 거래수}}{\text{전체 거래수}}$
신뢰도 (Confidence)	$\frac{\text{A와 B를 동시에 포함하는 거래 수}}{\text{A를 포함하는 거래수}}$
향상도 (Lift)	$\frac{\text{A와 B를 동시에 포함하는 거래수}}{\text{A를 포함하는 거래수} \times \text{B를 포함하는 거래수}}$
<ul style="list-style-type: none"> <li>- 지지도가 어느 정도 수준에 도달해야만 한다.</li> <li>- 신뢰도가 높을 경우에는 두 항목 A→B에서 항목 B의 확률이 커야지 연관규칙이 의미가 있다.</li> <li>- 향상도가 1보다 큰 값을 주어야 유용한 정보를 준다고 볼 수 있다.</li> </ul>	

## 2. 연관규칙 평가 척도

14-1 연관규칙 분석 I

### ■ 향상도(lift)

- A가 거래된 경우, 그 거래가 B를 포함하는 경우와 B가 임의로 거래되는 경우의 비율

$$\text{향상도 Lift(R)} = \frac{p(A \cap B)}{p(A)p(B)} = \frac{p(B|A)}{p(B)} = \frac{\text{conf(R)}}{p(B)}$$

향상도	의미
1	두 항목의 거래 발생이 독립적인 관계
< 1	두 항목의 거래 발생이 서로 음의 상관 관계
> 1	두 항목의 거래 발생이 서로 양의 상관 관계

- 각 항목의 구매가 상호 관련이 없다면 P(B|A)와 P(B)와 같게 되어 향상도는 1이 됨
- 1보다 크면 결과 예측에 대하여 우연적 기회(random chance)보다 우수함을 의미
- 향상도의 값이 클수록 A의 거래 여부가 B의 거래 여부에 큰 영향을 미침

### 3. 연관규칙 : 거래데이터 예제

14-1 연관규칙 분석 I

#### ▪ 식료품점 shopping cart

고객	항목들
1	Orange juice, soda
2	Milk, Orange juice, window cleaner
3	Orange juice, detergent
4	Orange juice, detergent, soda
5	Window cleaner, soda

transaction

#### • 동시발생행렬

	Orange juice	Window cleaner	Milk	Soda	Detergent
Orange juice	4 (0.8)	1 (0.2)	1 (0.2)	2 (0.4)	1 (0.2)
Window cleaner	1 (0.2)	2 (0.4)	1 (0.2)	1 (0.2)	0
Milk	1 (0.2)	1 (0.2)	1 (0.2)	0	0
Soda	2 (0.4)	1 (0.2)	0	3 (0.6)	1 (0.2)
Detergent	1 (0.2)	0	0	1 (0.2)	2 (0.4)

4 /5 transactions

### 4. 연관규칙 수행 패키지

14-1 연관규칙 분석 I

#### • 연관규칙 수행을 위한 패키지 : arules

```
# lec14_1_ass.r
# Association Rule
# Market basket analysis

# set working directory
setwd("D:/tempstore/moocr/wk14")

# association rule analysis package
install.packages("arules")
library(arules)
```

연관규칙분석 수행을 위한 패키지 설치 (arules)  
라이브러리 설정

## 5. 연관규칙분석을 위한 데이터

14-1 연관규칙 분석 I

### ▪ Data frame vs Transaction data

Dataframe				
ID	V1	V2	...	Vn
1	1	F	...	20
2	0	M	...	40
⋮	...	...	...	⋮
m	1	F	...	140

Transaction data		
Transaction ID	Items	Time Stamp
1	{A,B}	yyyy-mm-dd
2	{B}	yyyy-mm-dd
⋮	...	...
m	{A,K}	yyyy-mm-dd

Dataframe 형태와 transaction data의 구조가 다르다!!

## 5. 연관규칙분석을 위한 데이터변환

14-1 연관규칙 분석 I

```
# data import-> make transaction data
dvd1<-read.csv("dvdtrans.csv")
dvd1
dvd.list<-split(dvd1$Item,dvd1$ID)
dvd.list
dvd.trans<-as(dvd.list,"transactions")
dvd.trans
inspect(dvd.trans)
```

- arules package를 통해 transaction 데이터 변환과 연관 규칙 분석을 함
- Split을 통해 id별로 item들을 as함수를 통해 transaction 데이터로 변환

> dvd1		
ID	Item	
1	Sixth Sense	
2	LOTR1	
3	Harry Potter1	
4	Green Mile	
5	LOTR2	
6	Gladiator	
7	Patriot	
8	Braveheart	
9	LOTR1	
10	LOTR2	
11	Gladiator	
12	Patriot	



transaction data		
	items	transactionID
[1]	{Green Mile,Harry Potter1,LOTR1,LOTR2,Sixth Sense}	1
[2]	{Braveheart,Gladiator,Patriot}	2
[3]	{LOTR1,LOTR2}	3
[4]	{Gladiator,Patriot,Sixth Sense}	4
[5]	{Gladiator,Patriot,Sixth Sense}	5
[6]	{Gladiator,Patriot,Sixth Sense}	6
[7]	{Harry Potter1,Harry Potter2}	7
[8]	{Gladiator,Patriot}	8
[9]	{Gladiator,Patriot,Sixth Sense}	9
[10]	{Gladiator,Green Mile,LOTR,Sixth Sense}	10

## 5. 연관규칙분석을 위한 데이터변환

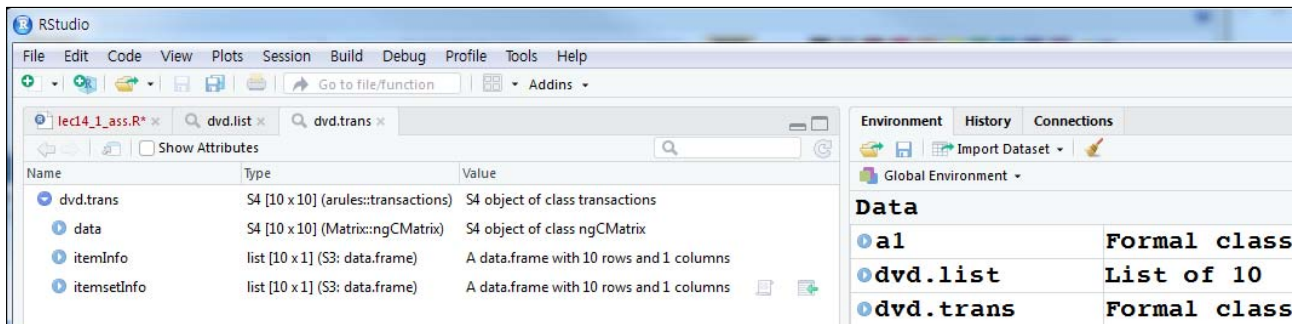
14-1 연관규칙 분석 I

### • transaction데이터로 변환된 'dvd.trans'

```
> dvd.trans
```

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

- transaction 10개
- items수 10개



Name	Type	Value
dvd.trans	S4 [10 x 10] (arules::transactions)	S4 object of class transactions
data	S4 [10 x 10] (Matrix::ngCMatrix)	S4 object of class ngCMatrix
itemInfo	list [10 x 1] (S3: data.frame)	A data.frame with 10 rows and 1 columns
itemsetInfo	list [10 x 1] (S3: data.frame)	A data.frame with 10 rows and 1 columns

Data	
a1	Formal class
dvd.list	List of 10
dvd.trans	Formal class

## 5. 연관규칙분석을 위한 데이터변환

14-1 연관규칙 분석 I

### • transaction 데이터의 요약

```
> summary(dvd.trans)
```

```
transactions as itemMatrix in sparse format with  
10 rows (elements/itemsets/transactions) and  
10 columns (items) and a density of 0.3
```

```
most frequent items:
```

```
Gladiator    Patriot    Sixth Sense    Green Mile    Harry Potter1  
      7         6         6         2         2  
(Other)  
      7
```



- 10트랜잭션 / 10 항목
- 밀도가 0.3 라고 되어 있는데, 10\*10 cell 중에서 30%의 cell에 거래가 발생해 숫자가 차 있다는 뜻
- 거래항목 중 Gladiator=7번, Patriot=6번, Six Sense=6번 순으로 나왔음을 의미

- 연관규칙 함수 : `apriori(transaction, parameter=list(support=0.0#, confidence=0.##))`

```
# for running dvdtrans data
dvd_rule<-apriori(dvd.trans,
                  parameter = list(support=0.2,
                                   confidence = 0.20,
                                   minlen = 2))
dvd_rule
```



```
> dvd_rule
set of 13 rules
```

support=0.2, confidence=0.2이상인 13개의 연관규칙 생성됨

```
# for running dvdtrans data
dvd_rule<-apriori(dvd.trans,
                  parameter = list(support=0.2,confidence = 0.20,minlen = 2))
dvd_rule

# same code with short command
# dvd_rule<-apriori(dvd.trans, parameter = list(supp=0.2,conf= 0.20,minlen = 2))
```

## 7. 연관규칙 수행결과 – dvdtrans 데이터

- 연관규칙 수행 컨솔창

```
> dvd_rule<-apriori(dvd.trans,
+                   parameter = list(support=0.2,
+                                   confidence = 0.20,
+                                   minlen = 2))
+
Apriori

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

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

Absolute minimum support count: 2

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10 item(s), 10 transaction(s)] done [0.00s].
sorting and recoding items ... [7 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [13 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

## 7. 연관규칙 수행결과 - dvdtrans 데이터

14-1 연관규칙 분석 I

### • 연관규칙 수행결과

```
# for running dvdtrans data
dvd_rule<-apriori(dvd.trans,
                  parameter = list(support=0.2,
                                   confidence = 0.20,
                                   minlen = 2))
```

dvd\_rule

```
summary(dvd_rule)
inspect(dvd_rule)
```

	lhs	rhs	support	confidence	lift	count
[1]	{LOTR1}	=> {LOTR2}	0.2	1.0000000	5.000000	2
[2]	{LOTR2}	=> {LOTR1}	0.2	1.0000000	5.000000	2
[3]	{Green Mile}	=> {Sixth Sense}	0.2	1.0000000	1.666667	2
[4]	{Sixth Sense}	=> {Green Mile}	0.2	0.3333333	1.666667	2
[5]	{Patriot}	=> {Sixth Sense}	0.4	0.6666667	1.111111	4
[6]	{Sixth Sense}	=> {Patriot}	0.4	0.6666667	1.111111	4
[7]	{Patriot}	=> {Gladiator}	0.6	1.0000000	1.428571	6
[8]	{Gladiator}	=> {Patriot}	0.6	0.8571429	1.428571	6
[9]	{Sixth Sense}	=> {Gladiator}	0.5	0.8333333	1.190476	5
[10]	{Gladiator}	=> {Sixth Sense}	0.5	0.7142857	1.190476	5
[11]	{Patriot,Sixth Sense}	=> {Gladiator}	0.4	1.0000000	1.428571	4
[12]	{Gladiator,Patriot}	=> {Sixth Sense}	0.4	0.6666667	1.111111	4
[13]	{Gladiator,Sixth Sense}	=> {Patriot}	0.4	0.8000000	1.333333	4

## 7. 연관규칙 수행결과 - dvdtrans 데이터

14-1 연관규칙 분석 I

### • 연관규칙의 해석

	lhs	rhs	support	confidence	lift	count
[1]	{LOTR1}	=> {LOTR2}	0.2	1.0000000	5.000000	2
[2]	{LOTR2}	=> {LOTR1}	0.2	1.0000000	5.000000	2
[3]	{Green Mile}	=> {Sixth Sense}	0.2	1.0000000	1.666667	2
[4]	{Sixth Sense}	=> {Green Mile}	0.2	0.3333333	1.666667	2
[5]	{Patriot}	=> {Sixth Sense}	0.4			
[6]	{Sixth Sense}	=> {Patriot}	0.4			
[7]	{Patriot}	=> {Gladiator}	0.6			
[8]	{Gladiator}	=> {Patriot}	0.6			
[9]	{Sixth Sense}	=> {Gladiator}	0.5			
[10]	{Gladiator}	=> {Sixth Sense}	0.5			
[11]	{Patriot,Sixth Sense}	=> {Gladiator}	0.4			
[12]	{Gladiator,Patriot}	=> {Sixth Sense}	0.4			
[13]	{Gladiator,Sixth Sense}	=> {Patriot}	0.4			

	items	tran
[1]	{Green Mile,Harry Potter1,LOTR1,LOTR2,Sixth Sense}	1
[2]	{Braveheart,Gladiator,Patriot}	2
[3]	{LOTR1,LOTR2}	3
[4]	{Gladiator,Patriot,Sixth Sense}	4
[5]	{Gladiator,Patriot,Sixth Sense}	5
[6]	{Gladiator,Patriot,Sixth Sense}	6
[7]	{Harry Potter1,Harry Potter2}	7
[8]	{Gladiator,Patriot}	8
[9]	{Gladiator,Patriot,Sixth Sense}	9
[10]	{Gladiator,Green Mile,LOTR,Sixth sense}	10

지지도 : Green Mile과 Sixth Sense를 동시에 구매할 확률 : 20%

신뢰도 : Green Mile을 구매한 경우는 모두 Sixth Sense를 구매 : 100%

향상도 : Green Mile을 구매하면 Six Sense의 구매비율이 1.667배 향상됨을 의미

(계산 : 신뢰도(R)=1, Six Sense구매비율=6/10=> 향상도 =1/0.6=1.667)



## 7. 연관규칙 수행결과 – dvdtrans 데이터

- 그래프로 표현한 연관규칙 – 지지도  $\geq 0.2$  이상의 항목들의 상대빈도

```
# Bar chart for support>0.2
itemFrequencyPlot(dvd.trans,support=0.2,main="item for support>0.2", col="green")
```

