

Wk14-3 : 순차적 패턴분석 (Sequence Pattern Analysis)

1. 순차적 패턴 분석

- 순차적 패턴(sequence pattern) : 고객들의 시간에 따른 구매 행태
ex) 컴퓨터를 구입한 사람들 중 25%는 그 다음달에 레이저 프린터를 구매한다.
- 시퀀스(sequence) : 항목집합을 순차적으로 나열시킨 리스트
- 시퀀스의 길이: 시퀀스에 포함된 항목집합 수, 길이가 k이면 k-시퀀스라 함

	연관규칙분석	순차적 패턴 분석
분석 주안점	동시 발생 사건	시간/순서에 따른 사건
필요 데이터	거래데이터	거래 집합 셋, Identity information, Time stamp
유용성 평가 척도	Support, confidence, lift	support

2. 순차적 패턴분석 수행 패키지

• 순차적 패턴분석 수행을 위한 패키지 : arulesSequences

```
# lec14_3_ass.r
# Association Rule
# Sequence Pattern Analysis

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

# Sequence Pattern analysis package
install.packages("arulesSequences")
library(arulesSequences)
```

} 순차적 패턴분석 수행을 위한 패키지 (arulesSequences)
라이브러리 설정

```
data(zaki)
summary(zaki)
```

} 탑재된 데이터 불러옴

3. 순차적 패턴 분석- 데이터 설명

▪ zaki 데이터

- M. J. Zaki가 시퀀스 마이닝의 개념을 설명하기 위해 만든 데이터
- Item 항목들과 sequence ID, event ID, size로 구성되어 있다.

Name	Type	Value
zaki	S4 [10 x 8] (arules::transactions)	S4 object of class transactions
data	S4 [8 x 10] (Matrix::ngCMatrix)	S4 object of class ngCMatrix
itemInfo	list [8 x 1] (S3: data.frame)	A data.frame with 8 rows and 1 columns
itemsetInfo	list [10 x 3] (S3: data.frame)	A data.frame with 10 rows and 3 columns

```
> summary(zaki)
transactions as itemMatrix in sparse format with
10 rows (elements/itemsets/transactions) and
8 columns (items) and a density of 0.3375

most frequent items:
  A      B      F      C      D (Other)
6      5      5      3      3          5
```

3. 순차적 패턴 분석- 데이터 설명

14-3 순차적 패턴분석

- zaki 데이터 – item, sequence ID, event ID, size로 구성

```
data(zaki)
summary(zaki)
# to see data frame
as(zaki,"data.frame")
```

```
> as(zaki,"data.frame")
  items sequenceID eventID SIZE
1 {C,D}          1      10    2
2 {A,B,C}         1      15    3
3 {A,B,F}         1      20    3
4 {A,C,D,F}        1      25    4
5 {A,B,F}         2      15    3
6 {E}             2      20    1
7 {A,B,F}         3      10    3
8 {D,G,H}         4      10    3
9 {B,F}           4      20    2
10 {A,G,H}        4      25    3
```

4. 순차적 패턴 분석 함수

14-3 순차적 패턴분석

- 순차적 패턴분석 함수 : cspade (데이터, parameter=list(supp=0.0#, , .))

```
# to see data frame
as(zaki,"data.frame")

seq_rule1<-cspade(zaki,
  parameter = list(support = 0.3,
                   maxsize = 5, maxlen =4),
  control = list(verbose=TRUE))
```

```
> seq_rule1<-cspade(zaki,
+   parameter = list(support = 0.3,
+                   maxsize = 5, maxlen =4),
+   control = list(verbose=TRUE))
```

```
parameter specification:
support : 0.3
maxsize : 5
maxlen : 4
```

```
algorithmic control:
bfstype : FALSE
verbose : TRUE
summary : FALSE
tidLists : FALSE
```

```
preprocessing ... 1 partition(s), 0 MB [0.18s]
mining transactions ... 0 MB [0.08s]
reading sequences ... [0.03s]
```

```
total elapsed time: 0.29s
```

- support : 시퀀스 최소 지지도(default 0.1, range [0,1])
- maxsize : 시퀀스 최대 항목 수 (default 10, range > 0)
- maxlen : 시퀀스 최대 길이 (default 10, range > 0)
- maxwin : 시퀀스의 두 항목 사이의 최대 시간 차이 (default none, range >= 0)

5. 순차적 패턴 분석결과

14-3 순차적 패턴분석

■ 순차적 패턴 결과

```
#analyzing results
summary(seq_rule1)
as(seq_rule1, "data.frame")
```

```
> as(seq_rule1, "data.frame")
  sequence support
1    <{A}>    1.00
2    <{B}>    1.00
3    <{D}>    0.50
4    <{F}>    1.00
5    <{A,F}>    0.75
6    <{B,F}>    1.00
7    <{D},{F}>    0.50
8    <{D},{B,F}>    0.50
9    <{A,B,F}>    0.75
10   <{A,B}>    0.75
11   <{D},{B}>    0.50
12   <{B},{A}>    0.50
13   <{D},{A}>    0.50
14   <{F},{A}>    0.50
15   <{D},{F},{A}>    0.50
16   <{B,F},{A}>    0.50
17   <{D},{B,F},{A}>    0.50
18   <{D},{B},{A}>    0.50
```

- 시퀀스 규칙에서 많이 나오는 항목이 A이다.
- Size가 1인 시퀀스 규칙이 8개이다.

```
> summary(seq_rule1)
set of 18 sequences with

most frequent items:
  A      B      F      D (Other)
11     10     10      8     28

most frequent elements:
  {A}    {D}    {B}    {F}    {B,F} (Other)
  8      8      4      4      4      3

element (sequence) size distribution:
sizes
1 2 3
8 7 3
```

5. 순차적 패턴 분석결과

14-3 순차적 패턴분석

■ 순차적 규칙 선정

```
#selecting rules(size>2)
seq_rule1_df<-as(seq_rule1,"data.frame")
seq_rule1_size<-size(seq_rule1)

seq_rule1_df<-cbind(seq_rule1_df,seq_rule1_size)
seq_rule1_df_size2<-subset(seq_rule1_df,
                           subset=seq_rule1_size>=2)
seq_rule1_df_size2
```

```
> seq_rule1_df_size2
  sequence support seq_rule1_size
7    <{D},{F}>    0.5             2
8    <{D},{B,F}>    0.5             2
11   <{D},{B}>    0.5             2
12   <{B},{A}>    0.5             2
13   <{D},{A}>    0.5             2
14   <{F},{A}>    0.5             2
15   <{D},{F},{A}>    0.5             3
16   <{B,F},{A}>    0.5             2
17   <{D},{B,F},{A}>    0.5             3
18   <{D},{B},{A}>    0.5             3
```

