Association Rules for Book-Crossing Data

# 1. Book-Crossing 데이터

* 2004년에 Institut f"{u}r Informatik, University of Freiburg에서 수행된 연구에서 심수집된 데이터( Book-Crossing이라는 책 커뮤니티)
* 변수: token(제목이 동일한 다른 ISBN을 갖는 책들을 구분), userid, title(책제목:unique하므로 책 ID로 사용), rating( 사용안됨)

<https://github.com/WinVector/zmPDSwR/raw/master/Bookdata/> 에서 bookdata.tsv.gz를 다운 받을 수 있음

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

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

# 데이터 읽기  
load("d:/bookbasket.RData")  
#bookbaskets <- read.transactions("~/R codes/bookdata.tsv.gz", format="single",   
# sep="\t",   
# cols=c("userid", "title"),  
# rm.duplicates=T)

# 2. 연관규칙분석

데이터 살펴보기

class(bookbaskets)

## [1] "transactions"  
## attr(,"package")  
## [1] "arules"

bookbaskets

## transactions in sparse format with  
## 92108 transactions (rows) and  
## 220447 items (columns)

dim(bookbaskets)

## [1] 92108 220447

colnames(bookbaskets)[1:5]

## [1] " A Light in the Storm: The Civil War Diary of Amelia Martin, Fenwick Island, Delaware, 1861"  
## [2] " Always Have Popsicles"   
## [3] " Apple Magic"   
## [4] " Ask Lily"   
## [5] " Beyond IBM: Leadership Marketing and Finance for the 1990s"

rownames(bookbaskets)[1:5]

## [1] "10" "1000" "100001" "100002" "100004"

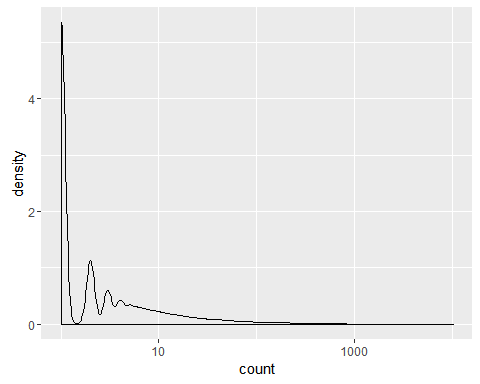
# 트랜잭션 사이즈 분포  
basketSizes <- size(bookbaskets)  
summary(basketSizes)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 1.0 11.1 4.0 10250.0

quantile(basketSizes, probs=seq(0,1,0.1))

## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## 1 1 1 1 1 1 2 3 5 13 10253

library(ggplot2)   
ggplot(data.frame(count=basketSizes)) +  
 geom\_density(aes(x=count)) +  
 scale\_x\_log10()



90%의 고객이 15권 미만의 책에 관심, 99%의 고객이 179권 이하의 책에 관심

어떤 책을 읽는가?

# 데이터에서 각 책의 상대도수  
bookFreq <- itemFrequency(bookbaskets)  
  
bookCount <- (bookFreq/sum(bookFreq))\*sum(basketSizes) # 책의 절대도수  
summary(bookCount)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 4.637 3.000 2502.000

# 가장 인기있는 책 10권  
orderedBooks <- sort(bookCount, decreasing=T)   
orderedBooks[1:10]

## Wild Animus   
## 2502   
## The Lovely Bones: A Novel   
## 1295   
## She's Come Undone   
## 934   
## The Da Vinci Code   
## 905   
## Harry Potter and the Sorcerer's Stone   
## 832   
## The Nanny Diaries: A Novel   
## 821   
## A Painted House   
## 819   
## Bridget Jones's Diary   
## 772   
## The Secret Life of Bees   
## 762   
## Divine Secrets of the Ya-Ya Sisterhood: A Novel   
## 737

orderedBooks[1]/dim(bookbaskets)[1] # 가장 인기있는 10권의 비율

## Wild Animus   
## 0.02716376

가장 인기있는 10권의 비율이 3%도 되지 않음(최소지지도값 결정시 고려)

절반정도 1권에만 관심을 표명하였음. 2권 이상인 경우로 제한

bookbaskets\_use <- bookbaskets[basketSizes > 1]  
dim(bookbaskets\_use)

## [1] 40822 220447

# 연관규칙 찾기  
rules <- apriori(bookbaskets\_use,   
 parameter =list(support = 0.002, confidence=0.75))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport support minlen maxlen  
## 0.75 0.1 1 none FALSE TRUE 0.002 1 10  
## target ext  
## rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 81   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[216031 item(s), 40822 transaction(s)] done [1.23s].  
## sorting and recoding items ... [1256 item(s)] done [0.06s].  
## creating transaction tree ... done [0.03s].  
## checking subsets of size 1 2 3 4 5 done [0.07s].  
## writing ... [191 rule(s)] done [0.00s].  
## creating S4 object ... done [0.05s].

summary(rules)

## set of 191 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 2 3 4 5   
## 11 100 66 14   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 3.000 3.000 3.435 4.000 5.000   
##   
## summary of quality measures:  
## support confidence lift   
## Min. :0.002009 Min. :0.7500 Min. : 40.89   
## 1st Qu.:0.002131 1st Qu.:0.8113 1st Qu.: 86.44   
## Median :0.002278 Median :0.8468 Median :131.36   
## Mean :0.002593 Mean :0.8569 Mean :129.68   
## 3rd Qu.:0.002695 3rd Qu.:0.9065 3rd Qu.:158.77   
## Max. :0.005830 Max. :0.9882 Max. :321.89   
##   
## mining info:  
## data ntransactions support confidence  
## bookbaskets\_use 40822 0.002 0.75

규칙을 평하하는 측도: coverage, fishersExactTest

measures <- interestMeasure(rules,   
 measure=c("coverage", "fishersExactTest"),   
 transactions=bookbaskets\_use)   
summary(measures)

## coverage fishersExactTest   
## Min. :0.002082 Min. : 0.000e+00   
## 1st Qu.:0.002511 1st Qu.: 0.000e+00   
## Median :0.002719 Median : 0.000e+00   
## Mean :0.003039 Mean :5.080e-138   
## 3rd Qu.:0.003160 3rd Qu.: 0.000e+00   
## Max. :0.006982 Max. :9.702e-136

# coverage는 0.002~0.007 (100~250명)  
# Fisher 검정의 p값은 작으므로 규칙들은 고객의 행동패턴을 실제 반영하고 있음  
  
# 신뢰도가 높은 5개의 규칙 살펴보기  
inspect(head((sort(rules, by="confidence")), n=5))

## lhs rhs support confidence lift  
## 1 {Four to Score,   
## High Five,   
## Seven Up,   
## Two for the Dough} => {Three To Get Deadly : A Stephanie Plum Novel} 0.002057714 0.9882353 165.33500  
## 2 {Harry Potter and the Order of the Phoenix,   
## Harry Potter and the Prisoner of Azkaban,   
## Harry Potter and the Sorcerer's Stone} => {Harry Potter and the Chamber of Secrets} 0.002866102 0.9669421 72.82751  
## 3 {Four to Score,   
## High Five,   
## One for the Money,   
## Two for the Dough} => {Three To Get Deadly : A Stephanie Plum Novel} 0.002082211 0.9659091 161.59976  
## 4 {Four to Score,   
## Seven Up,   
## Three To Get Deadly : A Stephanie Plum Novel,   
## Two for the Dough} => {High Five} 0.002057714 0.9655172 180.79975  
## 5 {High Five,   
## Seven Up,   
## Three To Get Deadly : A Stephanie Plum Novel,   
## Two for the Dough} => {Four to Score} 0.002057714 0.9655172 167.72062

조건에 따라 규칙 찾기

# The Lovely Bones => 인 규칙 (…rhs이면.. 결과값이지않나? =>..이아니라?)  
brules <- apriori(bookbaskets\_use,  
 parameter =list(support = 0.001,   
 confidence=0.6),  
 appearance=list(rhs=c("The Lovely Bones: A Novel"),   
 default="lhs"))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport support minlen maxlen  
## 0.6 0.1 1 none FALSE TRUE 0.001 1 10  
## target ext  
## rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 40   
##   
## set item appearances ...[1 item(s)] done [0.00s].  
## set transactions ...[216031 item(s), 40822 transaction(s)] done [1.25s].  
## sorting and recoding items ... [3172 item(s)] done [0.06s].  
## creating transaction tree ... done [0.04s].  
## checking subsets of size 1 2 3 4 5 6 7 8 done [0.48s].  
## writing ... [46 rule(s)] done [0.08s].  
## creating S4 object ... done [0.05s].

summary(brules)

## set of 46 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 3 4   
## 44 2   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.000 3.000 3.000 3.043 3.000 4.000   
##   
## summary of quality measures:  
## support confidence lift   
## Min. :0.001004 Min. :0.6000 Min. :21.81   
## 1st Qu.:0.001029 1st Qu.:0.6118 1st Qu.:22.24   
## Median :0.001102 Median :0.6258 Median :22.75   
## Mean :0.001132 Mean :0.6365 Mean :23.14   
## 3rd Qu.:0.001219 3rd Qu.:0.6457 3rd Qu.:23.47   
## Max. :0.001396 Max. :0.7455 Max. :27.10   
##   
## mining info:  
## data ntransactions support confidence  
## bookbaskets\_use 40822 0.001 0.6

# 신뢰도가 높은 5개 규칙  
brulesConf <- sort(brules, by="confidence")   
inspect(head(lhs(brulesConf), n=5))

## items   
## 1 {Divine Secrets of the Ya-Ya Sisterhood: A Novel,   
## Lucky : A Memoir}   
## 2 {Lucky : A Memoir,   
## The Notebook}   
## 3 {Lucky : A Memoir,   
## Wild Animus}   
## 4 {Midwives: A Novel,   
## Wicked: The Life and Times of the Wicked Witch of the West}  
## 5 {Lucky : A Memoir,   
## Summer Sisters}

# Lucky : A Memoir를 포함하는 규칙 … (포함하지 않는 아닌가>..?)  
brulesSub <- subset(brules, subset=!(lhs %in% "Lucky : A Memoir"))   
brulesConf <- sort(brulesSub, by="confidence")

## 연습문제

Hobbit, Harry Potter와 관련된 규칙들을 찾아보시오.