9. 단어의 연관성보기

arules패키지를 이용한 단어의 연관성 분석

I. 연관성 규칙 - Apriori algorithm

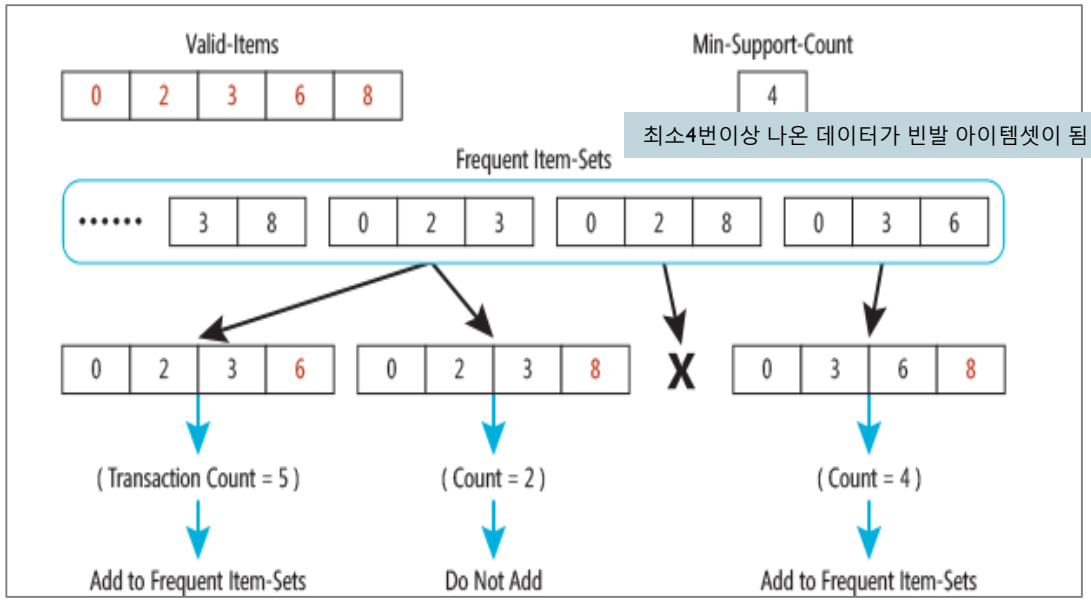
- 연관성분석 – 함께 발생하는 필드의 서로 다른 값들을 설명하는 규칙을 제공
- apriori 알고리즘 – 모든 필드의 값들을 카운트하지 않고 빈번한 아이템 셋을 계산하여 연관규칙을 얻어냄.

구매목록1	구매목록2	구매목록3
라면	생수	
담배	맥주	초콜릿
라면	맥주	파
상추	돼지고기	맥주
와인	치즈	
담배	껌	
맥주	초콜릿	
:	:	:

가장 빈번하게 구매하는 상품은 맥주,라면,초콜릿, 담배 등등이 되겠네요? 데이터가 많은 상태에서 빈번한 항목들을 알아낸다면 더 마트의 매출을 높일 수 있게 활용할 수 있어요.

빈발 아이템 셋(frequent item-set)을 알아내면, 만일 어떤 고객이 "담배와 초콜릿을 구매했다면 그 고객은 맥주를 함께 구매할 확률이 높다 " 는 규칙을 얻어 낼 수 있어요.
즉, 결과<=조건(포함도,정확도) ➔ 맥주 <= 담배 & 초콜릿(3000:20%,0.75)
"전체 고객의 20%인 3000명이 담배와 초콜릿을 구매했으며, 이 중 75%가 맥주도 함께 구매하였다."

빈발 아이템 셋 – 최소 지지도 이상을 갖는 아이템 셋을 의미



연관규칙의 효용성을 나타내는 지표

- 지지도(support) – 빈발 아이템셋을 판별할 때 사용하는것으로 조건이 발생할 확률을 의미
- 신뢰도(confidence) – 아이템 셋 간의 연관성 강도를 측정시 사용하는것으로, 조건절이 주어졌을때 결과가 발생할 확률을 의미
- 향상도(lift) – 생성된 연관규칙의 효용성을 판단하는 수치

2. apriori 알고리즘을 이용한 문서요약작업

how to 1) R에서 필요한 패키지 설치하기

```
install.packages(c("KoNLP","arules"))
library(KoNLP)
library(arules)
```

apriori알고리즘을 사용할 수 있는 패키지

2) bigdata.txt 파일을 읽어와서 명사를 추출하는 작업

```
> big_source=readLines("D:\\mining-r\\bigdata.txt",encoding="utf-8")
> find_word<-Map(extractNoun,big_source)
> head(find_word)
$`빅데이터를 활용한 의사결정`
[1] "빅데이터를" "활용"       "한"         "의사"       "결"

$`인구조사를 위해 인구 센서스 대신 빅데이터를 활용하는 최초의 정부가 등장한$
[1] "인구조사"   "인구"       "센서스"     "대"         "빅데이터를"
[6] "활용"       "최초"       "정부"       "등장"

$`커뮤니티에 대한 데이터는 그 어느때보다도 넘쳐나고 있다. 이러한 데이터를 이$
 [1] "커뮤니티"     "데이터"       "어느때보다도" "데이터"
 [5] "이해"         "관리"         "능력"         "향상"
 [9] "정부"         "기존"         "데이터"       "수집"
[13] "방식"

$`효율적이지 않다는 점을 인식하기 시작했고, 현재 프로그램을 자동화해 국민에$
 [1] "효율"     "적"       "점"       "인식하"   "시작"     "프로그램"
 [7] "자동화해" "국민"     "서비스"   "제공"     "수"       "혁신"
[13] "적"       "방법"     "찾기위해" "빅데이터" "기술"     "관심"
[19] "수"
```

3) 중복되는 단어를 제거하고 필터링 할 단어의 음절을 2~4글자 사이로 설정

```
> ex_word<-unique(find_word)
> ex_word2<-sapply(ex_word,unique)
> filter_con<-function(x) {nchar(x)>=2 && nchar(x)<=4}
> filter_apply<-function(x) {Filter(filter_con,x)}
> ex_word2<-sapply(ex_word2,filter_apply)
> head(ex_word2)
[[1]]
[1] "활용" "의사"

[[2]]
[1] "인구조사" "인구"     "센서스"   "활용"     "최초"     "정부"
[7] "등장"

[[3]]
 [1] "커뮤니티" "데이터"   "이해"     "관리"     "능력"     "향상"
 [7] "정부"     "기존"     "수집"     "방식"

[[4]]
 [1] "효율"     "인식하"   "시작"     "프로그램" "자동화해" "국민"
 [7] "서비스"   "제공"     "혁신"     "방법"     "찾기위해" "빅데이터"
[13] "기술"     "관심"

[[5]]
 [1] "활용"     "산업"     "분야"     "신속"     "의사결정" "가능"
 [7] "자동화"   "시민"     "번거로움" "기업"     "정부"     "고객"

[[6]]
[1] "세금신고" "납부"     "모든것에" "실시간"   "서비스"   "지원"
```

4) apriori 알고리즘은 트랜잭션에서 시작해요. 그림처럼 트랜잭션을 생성하고 행렬 구조로 단어사이의 연관성을 살펴보아요.

```
> word_trans=as(ex_word2,"transactions")
> word_trans
transactions in sparse format with
 7 transactions (rows) and
 76 items (columns)
> word_table=crossTable(word_trans)
> word_table
```

	가능	개인정보	결과	고객	과정	관리	관심	구축	국민	기반	기술	기업
가능	1	0	0	1	0	0	0	0	0	0	0	1
개인정보	0	1	1	0	1	0	0	1	0	1	0	0
결과	0	1	1	0	1	0	0	1	0	1	0	0
고객	1	0	0	1	0	0	0	0	0	0	0	1
과정	0	1	1	0	1	0	0	1	0	1	0	0
관리	0	0	0	0	0	1	0	0	0	0	0	0
관심	0	0	0	0	0	0	1	0	1	0	1	0
구축	0	1	1	0	1	0	0	1	0	1	0	0
국민	0	0	0	0	0	0	1	0	1	0	1	0
기반	0	1	1	0	1	0	0	1	0	1	0	0
기술	0	0	0	0	0	0	1	0	1	0	1	0
기업	1	0	0	1	0	0	0	0	0	0	0	1
기존	0	0	0	0	0	1	0	0	0	0	0	0
기회	0	1	1	0	1	0	0	1	0	1	0	0
납부	0	0	0	0	0	0	0	0	0	0	0	0
능력	0	0	0	0	0	1	0	0	0	0	0	0
대체	0	1	1	0	1	0	0	1	0	1	0	0
데이터	0	1	1	0	1	1	0	1	0	1	0	0
등장	0	0	0	0	0	0	0	0	0	0	0	0
명확	0	1	1	0	1	0	0	1	0	1	0	0
모든것에	0	0	0	0	0	0	0	0	0	0	0	0
발생	0	1	1	0	1	0	0	1	0	1	0	0
방법	0	0	0	0	0	0	1	0	1	0	1	0
방식	0	1	1	0	1	1	0	1	0	1	0	0
번거로움	1	0	0	1	0	0	0	0	0	0	0	1
분야	1	0	0	1	0	0	0	0	0	0	0	1

4) apriori(트랜잭션,지지도,신뢰도)를 입력하여 연관성규칙을 생성하게 함.

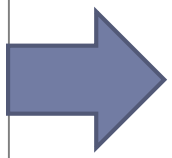
```
> ar_assoc<-apriori(word_trans,parameter=list(supp=0.25,conf=0.08))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
       0.08    0.1    1 none FALSE            TRUE       5    0.25      1
 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: 1

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[76 item(s), 7 transaction(s)] done [0.00s].
sorting and recoding items ... [8 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [21 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```



```
> inspect(ar_assoc)
     lhs                rhs          support   confidence lift     count
[1]  {}              => {서비스}     0.2857143 0.2857143  1.000000 2
[2]  {}              => {정부}       0.4285714 0.4285714  1.000000 3
[3]  {}              => {데이터}     0.2857143 0.2857143  1.000000 2
[4]  {}              => {방식}       0.2857143 0.2857143  1.000000 2
[5]  {}              => {수집}       0.2857143 0.2857143  1.000000 2
[6]  {}              => {의사결정}   0.2857143 0.2857143  1.000000 2
[7]  {}              => {빅데이터}   0.2857143 0.2857143  1.000000 2
[8]  {}              => {활용}       0.5714286 0.5714286  1.000000 4
[9]  {정부}          => {활용}       0.2857143 0.6666667  1.166667 2
[10] {활용}          => {정부}       0.2857143 0.5000000  1.166667 2
[11] {데이터}        => {방식}       0.2857143 1.0000000  3.500000 2
[12] {방식}          => {데이터}     0.2857143 1.0000000  3.500000 2
[13] {데이터}        => {수집}       0.2857143 1.0000000  3.500000 2
[14] {수집}          => {데이터}     0.2857143 1.0000000  3.500000 2
[15] {방식}          => {수집}       0.2857143 1.0000000  3.500000 2
[16] {수집}          => {방식}       0.2857143 1.0000000  3.500000 2
[17] {의사결정}      => {활용}       0.285714
[18] {활용}          => {의사결정}   0.285714
[19] {데이터,방식}   => {수집}       0.2857143 1.0000000  3.500000 2
[20] {데이터,수집}   => {방식}       0.2857143 1.0000000  3.500000 2
[21] {방식,수집}     => {데이터}     0.2857143 1.0000000  3.500000 2
>
```

단어간의 연관 효용성이 높음을 의미