co-citation network

dongwan kim

2020-08

## library packages

library(bibliometrix)

## To cite bibliometrix in publications, please use:  
##   
## Aria, M. & Cuccurullo, C. (2017) bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, 11(4), pp 959-975, Elsevier.  
##   
##   
## http:\\www.bibliometrix.org  
##   
##   
## To start with the shiny web-interface, please digit:  
## biblioshiny()

library(igraph)

##   
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

## The following object is masked from 'package:base':  
##   
## union

library(stringr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:igraph':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

options(encoding = "CP949")

## loading data & cleansing

# Information of 96 articles   
# WOS에서 가져온 TXT 파일을 불러들인다  
  
M <- convert2df("C://Users/dongwan/Desktop/dongwan/간호학과/final.ahrq.txt", dbsource="isi",format="plaintext")

##   
## Converting your isi collection into a bibliographic dataframe  
##   
## Done!  
##   
##   
## Generating affiliation field tag AU\_UN from C1: Done!

# Matrix article x reference  
# 논문 x 참조문헌 으로 matrix를 만든다   
  
cr<-cocMatrix(M,Field = "CR",sep =";")   
  
# 96 articles , 1939 references  
dim(cr)

## [1] 96 1939

# check the matrix  
cr[1:4,1:4]

## 4 x 4 sparse Matrix of class "dgCMatrix"  
## ACKERMAN DB 2010 J ARTHROPLASTY   
## GALBRAITH JG, 2011, 1  
## GUARASCIO-HOWARD L, 2011, .  
## VAN GAAL BGI, 2011, .  
## ANG E, 2011, .  
## ALEXANDER BH 1992 AM J PUBLIC HEALTH   
## GALBRAITH JG, 2011, 1  
## GUARASCIO-HOWARD L, 2011, .  
## VAN GAAL BGI, 2011, .  
## ANG E, 2011, .  
## AZHAR A 2008 IR MED J   
## GALBRAITH JG, 2011, 1  
## GUARASCIO-HOWARD L, 2011, .  
## VAN GAAL BGI, 2011, .  
## ANG E, 2011, .  
## BARRETT JA 2004 J AM GERIATR SOC   
## GALBRAITH JG, 2011, 1  
## GUARASCIO-HOWARD L, 2011, .  
## VAN GAAL BGI, 2011, .  
## ANG E, 2011, .

# convert reference information into "first name, last name, year, journal"   
# 이름, 성, 연도, 저널이름 순으로 전처리를 한다  
# 전처리 방법은 사람마다 다르다  
  
A<-as.matrix(cr)  
name<-dimnames(A)[[2]]  
name[1]

## [1] "ACKERMAN DB 2010 J ARTHROPLASTY "

for(i in 1:length(name)){  
   
 tmp<-strsplit(name[i],' ')[[1]]  
 tmp[1]<-tolower(tmp[1])  
 ttmp<-strsplit(tmp[1],'')[[1]]  
 ttmp[1]<-toupper(ttmp[1])  
 tmp[1]<-paste(ttmp,collapse = "")  
   
 ind<-grep("\\d",tmp)  
   
 tmp[ind]<-paste0(tmp[ind],",")  
 tmp[ind-1]<-paste0(tmp[ind-1],",")  
 name[i]<-paste(tmp,collapse=" ")  
}  
name[1]

## [1] "Ackerman DB, 2010, J ARTHROPLASTY"

dimnames(A)[[2]]<-name

## bibliographic co-citation

# We talk about co-citation of two articles when both are cited in a third article. Thus, co-citation can be seen as the counterpart of bibliographic coupling. # A co-citation network can be obtained using the general formulation: C = A' x A , where A is a bipartite network.  
  
# co-citation matrix with weighted edge  
# co-citation matrix 형태로 만든다 matrix A : 논문 x 참조문헌   
  
co.rf<-t(A) %\*% A  
  
# check the matrix  
co.rf[1:4,1:4]

## Ackerman DB, 2010, J ARTHROPLASTY  
## Ackerman DB, 2010, J ARTHROPLASTY 1  
## Alexander BH, 1992, AM J PUBLIC HEALTH 1  
## Azhar A, 2008, IR MED J 1  
## Barrett JA, 2004, J AM GERIATR SOC 1  
## Alexander BH, 1992, AM J PUBLIC HEALTH  
## Ackerman DB, 2010, J ARTHROPLASTY 1  
## Alexander BH, 1992, AM J PUBLIC HEALTH 2  
## Azhar A, 2008, IR MED J 1  
## Barrett JA, 2004, J AM GERIATR SOC 1  
## Azhar A, 2008, IR MED J  
## Ackerman DB, 2010, J ARTHROPLASTY 1  
## Alexander BH, 1992, AM J PUBLIC HEALTH 1  
## Azhar A, 2008, IR MED J 1  
## Barrett JA, 2004, J AM GERIATR SOC 1  
## Barrett JA, 2004, J AM GERIATR SOC  
## Ackerman DB, 2010, J ARTHROPLASTY 1  
## Alexander BH, 1992, AM J PUBLIC HEALTH 1  
## Azhar A, 2008, IR MED J 1  
## Barrett JA, 2004, J AM GERIATR SOC 4

#convert co-citation matrix into igraph  
#아이그래프 에서 쓸수있는 데이터로 만든다  
  
co.rf.g<-graph.adjacency(co.rf,mode = "undirected",weighted = T)

## network statistics

# eigenvector centrality  
eig\_cen<-eigen\_centrality(co.rf.g)[1]  
eig\_cen<-data.frame(eig\_cen)  
eig\_cen$name<-name  
  
# betweenness centrality  
bet\_cen<-betweenness(co.rf.g)  
bet\_cen<-data.frame(bet\_cen)  
  
# closeness centrality   
clo\_cen<-closeness(co.rf.g)

## Warning in closeness(co.rf.g): At centrality.c:2617 :closeness centrality is not  
## well-defined for disconnected graphs

clo\_cen<-data.frame(clo\_cen)  
  
# degree & degree centrality (degree/(n-1))  
deg<-degree(co.rf.g)  
deg<-data.frame(deg)  
deg\_cen<-deg/(dim(deg)[1]-1)

## network detection(clustering)

# leading.eigenvector.community by MEJ Newman, 2006  
eig<-leading.eigenvector.community(co.rf.g)  
  
# cluster\_louvain by Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, 2008  
lou<-cluster\_louvain(co.rf.g)  
  
# groups & modularity  
eig

## IGRAPH clustering leading eigenvector, groups: 24, mod: 0.64  
## + groups:  
## $`1`  
## [1] "Ackerman DB, 2010, J ARTHROPLASTY"   
## [2] "Azhar A, 2008, IR MED J"   
## [3] "Barrett JA, 2004, J AM GERIATR SOC"   
## [4] "Bass E, 2008, J AM MED DIR ASSOC"   
## [5] "Chang JT, 2004, BRIT MED J"   
## [6] "Chapuy MC, 1992, NEW ENGL J MED"   
## [7] "Cotter PE, 2006, IRISH J MED SCI"   
## [8] "Coussement J, 2008, J AM GERIATR SOC"   
## [9] "Fonda D, 2006, MED J AUSTRALIA"   
## + ... omitted several groups/vertices

lou

## IGRAPH clustering multi level, groups: 24, mod: 0.68  
## + groups:  
## $`1`  
## [1] "Ackerman DB, 2010, J ARTHROPLASTY"   
## [2] "Alexander BH, 1992, AM J PUBLIC HEALTH"   
## [3] "Azhar A, 2008, IR MED J"   
## [4] "Barrett JA, 2004, J AM GERIATR SOC"   
## [5] "Bass E, 2008, J AM MED DIR ASSOC"   
## [6] "Bates DW, 1995, AM J MED"   
## [7] "Chang JT, 2004, BRIT MED J"   
## [8] "Chapuy MC, 1992, NEW ENGL J MED"   
## [9] "Cotter PE, 2006, IRISH J MED SCI"   
## + ... omitted several groups/vertices

## bind the information

ahrq.info<-cbind.data.frame(eig\_cen[,2],eig\_cen[,1],bet\_cen,clo\_cen,deg\_cen,deg,eig$membership,lou$membership)  
names(ahrq.info)<-c("name","eigen\_centrality","betweenness\_centrality","closeness","degree\_centrality","degree","leading.eigen.cluster","louvain.cluster")  
  
#sort by eigenvector centrality  
#아이젠벡터 중심 기준으로 데이터를 정렬한다  
ahrq.info<-arrange(ahrq.info,desc(ahrq.info$eigen\_centrality))  
ahrq.info[1:10,]

## name eigen\_centrality betweenness\_centrality  
## 1 Oliver D, 1997, BMJ-BRIT MED J 1.0000000 114467.69  
## 2 Oliver D, 2000, J AM GERIATR SOC 0.8437843 82293.77  
## 3 Oliver D, 2004, AGE AGEING 0.8188914 123891.35  
## 4 Bates DW, 1995, AM J MED 0.7090256 62733.97  
## 5 Haines TP, 2004, BMJ-BRIT MED J 0.5781602 43603.67  
## 6 Halfon P, 2001, J CLIN EPIDEMIOL 0.5421982 51279.49  
## 7 Perell KL, 2001, J GERONTOL A-BIOL 0.4945042 31861.53  
## 8 Lundebjerg N, 2001, J AM GERIATR SOC 0.4785977 133765.14  
## 9 Hendrich A, 1995, APPL NURS RES 0.4670978 36743.01  
## 10 Tideiksaar R, 1993, MT SINAI J MED 0.4371443 56007.15  
## closeness degree\_centrality degree leading.eigen.cluster louvain.cluster  
## 1 5.719515e-05 0.3059856 593 1 1  
## 2 5.690224e-05 0.2683179 520 4 1  
## 3 5.768011e-05 0.2786378 540 1 1  
## 4 5.670542e-05 0.2218782 430 4 1  
## 5 5.677624e-05 0.1738906 337 1 1  
## 6 5.691520e-05 0.1873065 363 1 1  
## 7 5.656749e-05 0.1749226 339 1 1  
## 8 5.760701e-05 0.2079463 403 4 14  
## 9 5.634755e-05 0.1656347 321 4 14  
## 10 5.688606e-05 0.1744066 338 4 1

## network plot

# E(graph) is about edge information , V(graph) is about vertex information  
# V(graph)$size is eigenvertor centrality   
# E(grpah)$widh is weighted edge  
  
E(co.rf.g)$width = E(co.rf.g)$weight  
V(co.rf.g)$size<-unlist(eig\_cen[1])  
  
# top 30 references by egienvector centrality  
# top 30 의 문헌만 가져온다  
  
ind<-V(co.rf.g)$name %in% ahrq.info[1:30,]$name  
sum(ind)

## [1] 30

# subgraph with top 30 references by egienvector centrality  
# induced.subgraph 함수를 쓰면 본인이 설정하는 대로 sub network 를 뽑을 수 있다  
  
induce1<-induced.subgraph(co.rf.g,V(co.rf.g)$name[which(ind==T)])  
  
# labeling top 20 references  
ind<-V(induce1)$name %in% ahrq.info[1:20,]$name  
sum(ind)

## [1] 20

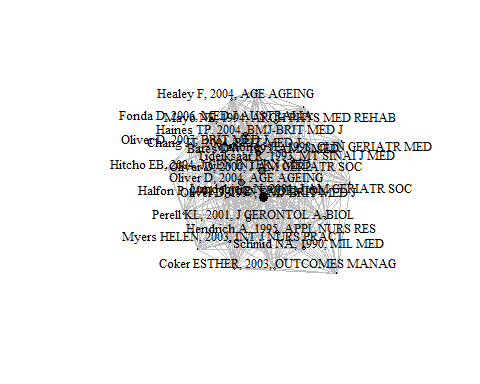
V(induce1)$initial[which(ind==T)]<-V(induce1)$name[which(ind==T)]

## Warning in vattrs[[name]][index] <- value: number of items to replace is not a  
## multiple of replacement length

V(induce1)$initial[which(ind==F)]<-""  
sum(V(induce1)$initial=="")

## [1] 10

# remove loops & multiple  
  
g2<-simplify(induce1,remove.loops=TRUE,remove.multiple = T)  
  
# color 1 black, 2-5, dim gray, the others gray  
V(g2)$color<-ifelse(V(g2)$size>=0.99,"black",  
 ifelse(V(g2)$size>= 0.5781601,"Dim gray","Gray"))  
  
# network plot  
  
set.seed(12)  
plot(g2,layout=layout.fruchterman.reingold, vertex.color=V(g2)$color, vertex.size=unlist(V(g2)$size)\*10, vertex.label.color="black",  
 vertex.label.cex=0.8, vertex.label.dist=1, edge.curved=0,edge.width=E(g2)$weight/3,edge.color="gray",vertex.label=V(g2)$initial)



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.