Trabalho Final - R para Iniciantes

Maurílio Bonora Júnior

15/05/2019

## Introdução

Sempre que vamos escrever um projeto ou relatório, somos bombardeados por um gigantesca base de dados com artigos de relevância variável sobre o determinado assunto. Organizar e escolher os melhores artigos dentre essa imensidão de referências é uma dura tarefa, que muitas vezes pode demandar um tempo que muitos pesquisadores não têm. Contudo, com o advento do R e seus diversos pacotes, entre eles o “Bibliometrix”, essa tarefa de comparar artigos para escolher os melhores se tornou muito mais fácil. Para exemplificar isso, foram escolhidos dois termos para se começar a busca de arquivos em base de dados (Scopus). Tais termos foram: “TCR” e “Signalling”. A escolha dos termos foi feito baseado na experiência do autor desse trabalho, vide que a área de pesquisa desse se dá em tal área (Imunologia celular e molecular). Abaixo segue todo o procedimento com os códigos e gráficos possíveis de se fazer com o pacote “Bibliometrix”.

## Requerimentos

Para a realização do trabalho, foram necessários alguns programas e pacotes, dentre eles:

* O próprio R: garante toda a linguagem R para se trabalhar;
* RStudio: um software com uma interface mais elegante e amigável ao usuário para ele conseguir trabalho;
* Git: um sistema de controle de versão distribuído. Utilizado juntamente do Rmarkdown para se fazer o upload direto do trabalho para um repositório em nuvem (Github);
* MikTex: programa necessário para a conversão do Script do Rmarkdown para PDf;
* Pacote “Rmarkdown”: converte scripts do R (mais especificamente do Rmarkdown) em uma variedade de formatos incluindo HTML, MS Word, PDF e Beamer. Além disso, o Rmarkdown consegue compilar os scripts em especíes de “livros” onde é possível colocar comentários, códigos fontes e a saída (resultado) do código do script;
* Pacote “Bibliometrix”: garante um conjunto de ferramentas muito útil para análises na área de cientometria e bibliometria;
* Um arquivo Scopus.bib: arquivo baixado da base de dados Scopus com todas as informações de artigos escolhidos a partir das palavras-chaves selecionado.

## Desenvolvimento

#### Carregamento e Conversão dos Dados

Após a instalação do R, RStudio, Miktex e Git (lembrando que é necessário a criação de um diretório com o nome do trabalho no Git), é necessário fazer o download dos pacotes “Rmarkdown” e “Bibliometrix”.

Em seguida é necessário carregar o pacote “Bibliometrix”:

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()

Já com o arquivo scopus.bib dentro da sua pasta de trabalho está na hora de começar a preparar os dados para as análises. A primeira função que será utilizada é “read.Files”, que converte todos os arquivos de texto de scopus.bib em um grande vetor de caracteres que chamaremos de D.

D<-readFiles("scopus.bib")

Esse grande objeto D pode então ser convertido em um DataFrame a partir da função “convert2df”.

M <- convert2df(D, dbsource = "scopus", format = "bibtex")

##   
## Converting your scopus collection into a bibliographic dataframe  
##   
## Articles extracted 100   
## Articles extracted 200   
## Articles extracted 300   
## Articles extracted 400   
## Articles extracted 500   
## Articles extracted 511   
## Done!  
##   
##   
## Generating affiliation field tag AU\_UN from C1: Done!

#### Análise Bibliométrica

O primeiro passo para fazer uma análise descritiva desse Data Frame é utilizara função “biblioAnalysis”, que irá calcular as principais medidas bibliométricas de toda nossa base de dados.

results<-biblioAnalysis(M, sep = ";")

Tal função retorna um objeto da classe bibliometrix (introduzida com o pacote homônimo), que contem os seguintes componentes como o número total de artigos, primeiro autor de cada manuscrito, número de vezes que cada manustrico foi citado, entre outros. Para mais informações use a função:

help("biblioAnalysis")

## starting httpd help server ... done

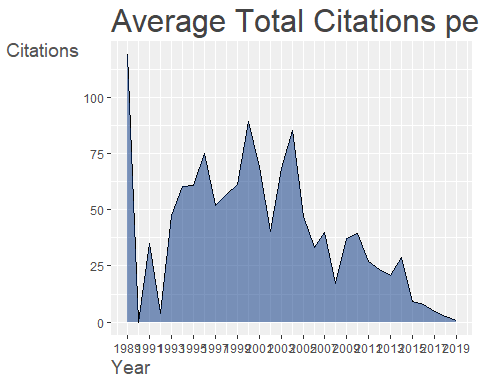
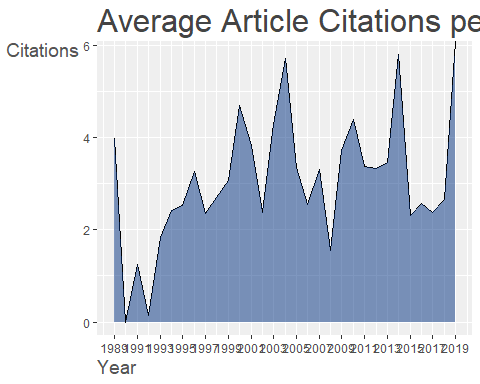
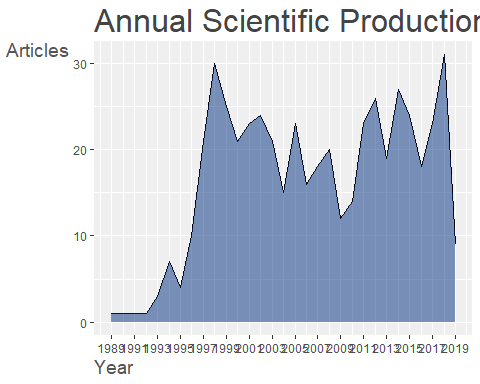
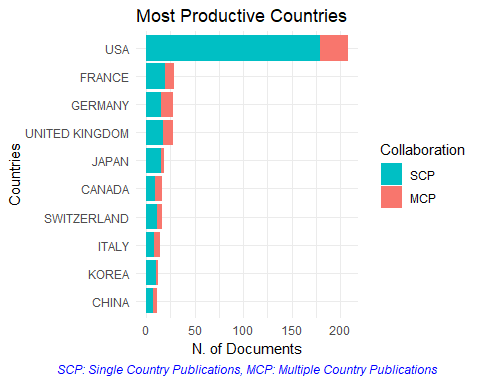
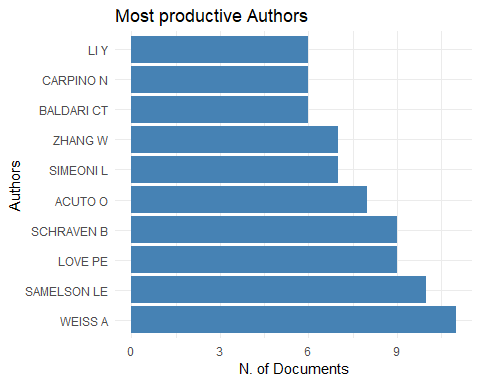
Contudo, o objeto “results” que obtemos ainda é muito grande e complicado de se ler, sendo muito laborioso obter informações gerais a partir dali. Para isso, usa-se a função “summary” para se obter os principais resultados da nossa análise bibliométrica. Essa função aceita ainda dois argumentos adicionais: “k” é um valor de formatação que indica quantas linhas de cada tabela serão mostradas, enquanto “pause” é um valor lógico usado para permitir ou não a pausa na rolagem da tela.

summary(results, k =10, pause = FALSE)

##   
##   
## Main Information about data  
##   
## Documents 511   
## Sources (Journals, Books, etc.) 127   
## Keywords Plus (ID) 3213   
## Author's Keywords (DE) 609   
## Period 1989 - 2019   
## Average citations per documents 38.77   
##   
## Authors 2560   
## Author Appearances 3047   
## Authors of single-authored documents 16   
## Authors of multi-authored documents 2544   
## Single-authored documents 17   
##   
## Documents per Author 0.2   
## Authors per Document 5.01   
## Co-Authors per Documents 5.96   
## Collaboration Index 5.15   
##   
## Document types   
## ARTICLE 445   
## BOOK CHAPTER 6   
## CONFERENCE PAPER 5   
## EDITORIAL 2   
## ERRATUM 4   
## NOTE 5   
## REVIEW 40   
## SHORT SURVEY 4   
##   
##   
## Annual Scientific Production  
##   
## Year Articles  
## 1989 1  
## 1990 1  
## 1991 1  
## 1992 1  
## 1993 3  
## 1994 7  
## 1995 4  
## 1996 10  
## 1997 21  
## 1998 30  
## 1999 25  
## 2000 21  
## 2001 23  
## 2002 24  
## 2003 21  
## 2004 15  
## 2005 23  
## 2006 16  
## 2007 18  
## 2008 20  
## 2009 12  
## 2010 14  
## 2011 23  
## 2012 26  
## 2013 19  
## 2014 27  
## 2015 24  
## 2016 18  
## 2017 23  
## 2018 31  
## 2019 9  
##   
## Annual Percentage Growth Rate 7.598962   
##   
##   
## Most Productive Authors  
##   
## Authors Articles Authors Articles Fractionalized  
## 1 WEISS A 11 SAMELSON LE 2.52  
## 2 SAMELSON LE 10 WEISS A 2.42  
## 3 LOVE PE 9 ACUTO O 2.21  
## 4 SCHRAVEN B 9 NA NA 2.00  
## 5 ACUTO O 8 HARDER T 1.81  
## 6 SIMEONI L 7 ABRAHAM RT 1.70  
## 7 ZHANG W 7 WANGE RL 1.64  
## 8 BALDARI CT 6 LOVE PE 1.61  
## 9 CARPINO N 6 MICHEL F 1.60  
## 10 LI Y 6 SCHRAVEN B 1.58  
##   
##   
## Top manuscripts per citations  
##   
## Paper TC TCperYear  
## 1 MERESSE B, 2004, IMMUNITY 545 36.33  
## 2 WANGE RL, 1996, IMMUNITY 443 19.26  
## 3 ACUTO O, 2003, NAT REV IMMUNOL 413 25.81  
## 4 MOSSMAN KD, 2005, SCIENCE 361 25.79  
## 5 DOWER NA, 2000, NAT IMMUNOL 334 17.58  
## 6 JANES PW, 2000, SEMIN IMMUNOL 322 16.95  
## 7 KERSH GJ, 1998, IMMUNITY 248 11.81  
## 8 KRAUSE M, 2000, J CELL BIOL 240 12.63  
## 9 ZHANG W, 1999, INT IMMUNOL 218 10.90  
## 10 CHAN AC, 1994, J IMMUNOL 218 8.72  
##   
##   
## Corresponding Author's Countries  
##   
## Country Articles Freq SCP MCP MCP\_Ratio  
## 1 USA 208 0.4771 179 29 0.139  
## 2 FRANCE 29 0.0665 19 10 0.345  
## 3 GERMANY 28 0.0642 15 13 0.464  
## 4 UNITED KINGDOM 28 0.0642 17 11 0.393  
## 5 JAPAN 18 0.0413 15 3 0.167  
## 6 CANADA 16 0.0367 9 7 0.438  
## 7 SWITZERLAND 16 0.0367 11 5 0.312  
## 8 ITALY 14 0.0321 8 6 0.429  
## 9 KOREA 12 0.0275 10 2 0.167  
## 10 CHINA 11 0.0252 7 4 0.364  
##   
##   
## SCP: Single Country Publications  
##   
## MCP: Multiple Country Publications  
##   
##   
## Total Citations per Country  
##   
## Country Total Citations Average Article Citations  
## 1 USA 10614 51.03  
## 2 FRANCE 1619 55.83  
## 3 UNITED KINGDOM 1461 52.18  
## 4 GERMANY 1046 37.36  
## 5 CANADA 1032 64.50  
## 6 JAPAN 660 36.67  
## 7 SWITZERLAND 597 37.31  
## 8 ITALY 529 37.79  
## 9 NETHERLANDS 397 56.71  
## 10 AUSTRALIA 278 34.75  
##   
##   
## Most Relevant Sources  
##   
## Sources Articles  
## 1 JOURNAL OF IMMUNOLOGY 158  
## 2 EUROPEAN JOURNAL OF IMMUNOLOGY 32  
## 3 NATURE IMMUNOLOGY 26  
## 4 BLOOD 16  
## 5 JOURNAL OF EXPERIMENTAL MEDICINE 16  
## 6 FRONTIERS IN IMMUNOLOGY 14  
## 7 IMMUNITY 14  
## 8 INTERNATIONAL IMMUNOLOGY 14  
## 9 MOLECULAR IMMUNOLOGY 14  
## 10 PLOS ONE 9  
##   
##   
## Most Relevant Keywords  
##   
## Author Keywords (DE) Articles Keywords-Plus (ID) Articles  
## 1 SIGNAL TRANSDUCTION 28 SIGNAL TRANSDUCTION 788  
## 2 TCR 27 MICE 477  
## 3 TCR SIGNALING 18 RECEPTORS 420  
## 4 T CELL RECEPTOR 17 ARTICLE 418  
## 5 T LYMPHOCYTES 16 T LYMPHOCYTE RECEPTOR 402  
## 6 T CELL 15 ANTIGEN 400  
## 7 T CELL ACTIVATION 15 T CELL 397  
## 8 T CELLS 14 PRIORITY JOURNAL 380  
## 9 SIGNALING 11 HUMAN 315  
## 10 T CELL SIGNALING 10 MOUSE 314

A partir daí obtemos uma série de resultados interessantes, desde as principais informações dos nossos dados - como nº de documentos, média de citações por artigos, nº de autores, nº de artigos com autores únicos - até informações mais curiosas como a produção de artios por ano desde o primeiro que foi publicado, autores mais produtivos, países com maiores nº de citações, principais fontes de artigos, entre diversas outras. Por fim, a partir dessas mesmos informações básicas obtidas, é possível montar gráficos que nos dão uma ideia bem mais visual da relação entre nossos dados, usando a função “plot”.

plot(x=results, k=10, pause=FALSE)



#### Análise de Referências Citadas

É possível produzir uma tabela mostrando as referências mais citadas dos autores mais citados, utilizando-se a função “citations”. Para se obter os artigos mais citadas, usa-se:

CR <- citations(M, field = "article", sep = ";")  
cbind(CR$Cited[1:10])

## [,1]  
## ZHANG, W., SLOAN-LANCASTER, J., KITCHEN, J., TRIBLE, R.P., SAMELSON, L.E., LAT: THE ZAP-70 TYROSINE KINASE SUBSTRATE THAT LINKS T CELL RECEPTOR TO CELLULAR ACTIVATION (1998) CELL, 92, PP. 83-92 24  
## SMITH-GARVIN, J.E., KORETZKY, G.A., JORDAN, M.S., T CELL ACTIVATION (2009) ANNU. REV. IMMUNOL., 27, PP. 591-619 21  
## WEISS, A., LITTMAN, D.R., SIGNAL TRANSDUCTION BY LYMPHOCYTE ANTIGEN RECEPTORS (1994) CELL, 76, PP. 263-274 20  
## WEISS, A., LITTMAN, D.R., SIGNAL TRANSDUCTION BY LYMPHOCYTE ANTIGEN RECEPTORS (1994) CELL, 76, P. 263 16  
## XAVIER, R., BRENNAN, T., LI, Q., MCCORMACK, C., SEED, B., MEMBRANE COMPARTMENTATION IS REQUIRED FOR EFFICIENT T CELL ACTIVATION (1998) IMMUNITY, 8, PP. 723-732 13  
## VARMA, R., CAMPI, G., YOKOSUKA, T., SAITO, T., DUSTIN, M.L., T CELL RECEPTOR-PROXIMAL SIGNALS ARE SUSTAINED IN PERIPHERAL MICROCLUSTERS AND TERMINATED IN THE CENTRAL SUPRAMOLECULAR ACTIVATION CLUSTER (2006) IMMUNITY, 25, PP. 117-127 12  
## VIOLA, A., SCHROEDER, S., SAKAKIBARA, Y., LANZAVECCHIA, A., T LYMPHOCYTE COSTIMULATION MEDIATED BY REORGANIZATION OF MEMBRANE MICRODOMAINS (1999) SCIENCE, 283, PP. 680-682 12  
## WANGE, R.L., SAMELSON, L.E., COMPLEX COMPLEXES: SIGNALING AT THE TCR (1996) IMMUNITY, 5, P. 197 11  
## ZHANG, W., TRIBLE, R.P., SAMELSON, L.E., LAT PALMITOYLATION: ITS ESSENTIAL ROLE IN MEMBRANE MICRODOMAIN TARGETING AND TYROSINE PHOSPHORYLATION DURING T CELL ACTIVATION (1998) IMMUNITY, 9, PP. 239-246 11  
## CANTRELL, D., T CELL ANTIGEN RECEPTOR SIGNAL TRANSDUCTION PATHWAYS (1996) ANNU. REV. IMMUNOL., 14, P. 259 10

Enquanto que para se obter os autores mais citados, usa-se:

CR<- citations(M, field = "author", sep = ";")  
cbind(CR$Cited[1:10])

## [,1]  
## WEISS A 669  
## SAMELSON L E 403  
## KORETZKY G A 268  
## DAVIS M M 244  
## VON BOEHMER H 240  
## CHAN A C 219  
## ZHANG W 216  
## GERMAIN R N 192  
## ALLEN P M 185  
## DUSTIN M L 177

Por fim, é possível montar ainda uma tabela onde mostra-se os autores mais citados “localmente”, isto é, dentre a base de dados que temos, quais foram os autores mais citados entre os outros autores dessa base. Para isso, utilizamos a função “LocalCitations”. Utilizaremos as mesmas funções auxiliares usadas anteriormente para ver os 10 principais autores e principais artigos citados local e respectivamente:

CR <- localCitations(M, sep = ";")

## Articles analysed 100   
## Articles analysed 200   
## Articles analysed 300   
## Articles analysed 400   
## Articles analysed 462

CR$Authors[1:10,]

## Author LocalCitations  
## 11 ACUTO O 30  
## 1776 SAMELSON LE 28  
## 1389 MICHEL F 21  
## 2213 WANGE RL 21  
## 1255 LOVE PE 20  
## 2121 TUOSTO L 16  
## 88 ATTAL BONNEFOY G 14  
## 1302 MANGINO G 14  
## 2227 WEISS A 14  
## 1896 SHORES EW 13

CR$Papers[1:10,]

## Paper DOI Year LCS GCS  
## 24 WANGE RL, 1996, IMMUNITY 10.1016/S1074-7613(00)80315-5 1996 19 443  
## 58 TUOSTO L, 1998, EUR J IMMUNOL 10.1002/(SICI)1521-4141(199807)28:07<2131::AID-IMMU2131>3.0.CO;2-Q 1998 9 85  
## 124 AZZAM HS, 2001, J IMMUNOL 10.4049/JIMMUNOL.166.9.5464 2001 9 180  
## 191 CIOFANI M, 2004, J IMMUNOL 10.4049/JIMMUNOL.172.9.5230 2004 9 183  
## 99 MICHEL F, 2000, J IMMUNOL 10.4049/JIMMUNOL.165.7.3820 2000 7 73  
## 132 MICHEL F, 2001, IMMUNITY 10.1016/S1074-7613(01)00244-8 2001 7 128  
## 133 RAAB M, 2001, IMMUNITY 10.1016/S1074-7613(01)00248-5 2001 7 84  
## 362 GUY CS, 2013, NAT IMMUNOL 10.1038/NI.2538 2013 7 97  
## 139 KOSUGI A, 2001, IMMUNITY 10.1016/S1074-7613(01)00146-7 2001 6 85  
## 195 CARPINO N, 2004, IMMUNITY 10.1016/S1074-7613(03)00351-0 2004 6 100

#### Ranking de Dominância de Autores

O fator de dominância é uma taxa que indica a fração de artigos com múltiplos autores em que um pesquisador aparece como primeiro autor. A função “dominance” calcula o ranking da dominância entre os autores.

DF <- dominance(results, k = 10)  
DF

## Author Dominance Factor Tot Articles Single-Authored Multi-Authored First-Authored Rank by Articles Rank by DF  
## 1 HARDER T 0.6666667 4 1 3 2 1 1  
## 2 MICHEL F 0.5000000 6 0 6 3 5 2  
## 3 CARPINO N 0.3333333 6 0 6 2 5 3  
## 4 HOUTMAN JCD 0.2500000 4 0 4 1 1 4  
## 5 LEITENBERG D 0.2500000 4 0 4 1 1 4  
## 6 VAN OERS NSC 0.2000000 5 0 5 1 4 6  
## 7 LI Y 0.1666667 6 0 6 1 5 7  
## 8 ZHANG W 0.1428571 7 0 7 1 8 8  
## 9 ACUTO O 0.1250000 8 0 8 1 9 9  
## 10 LOVE PE 0.1111111 9 0 9 1 10 10

#### O índice-h dos autores

O índice-h é uma métrica que mede ambas produtividade e impacto de citação das publicações de cientistas. Esse índice é baseado no conjunto de artigos mais citados do autor em específico e no número de citações que esses artigos receberam em outras publicações. A função “Hindex” calcula o índice-h de autores ou de fontes e suas variantes (índice-g e índice-m) em uma conjunto de dados bibliográficos. Para calcular o índice-h dos 10 autores mais produtivos dessa coleção em específico, utilizaremos os seguintes códigos:

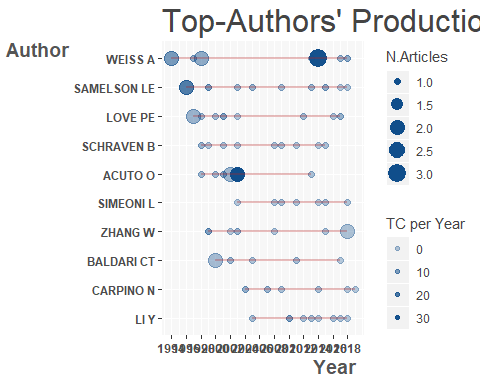
authors=gsub(","," ",names(results$Authors)[1:10])  
indices <- Hindex(M, field = "author", elements=authors, sep = ";", years = 50)  
indices$H

## Author h\_index g\_index m\_index TC NP PY\_start  
## 1 WEISS A 10 11 0.3846154 819 11 1994  
## 2 SAMELSON LE 8 10 0.3333333 907 10 1996  
## 3 LOVE PE 7 9 0.3043478 471 9 1997  
## 4 SCHRAVEN B 7 9 0.3181818 222 9 1998  
## 5 ACUTO O 8 8 0.3636364 905 8 1998  
## 6 SIMEONI L 4 7 0.2352941 69 7 2003  
## 7 ZHANG W 4 7 0.1904762 319 7 1999  
## 8 BALDARI CT 5 6 0.2500000 174 6 2000  
## 9 CARPINO N 4 6 0.2500000 197 6 2004  
## 10 LI Y 5 8 0.3333333 177 8 2005

#### Produtividade dos principais autores no decorrer do tempo

A função “AuthorProdOverTime” facilmente consegue nos dar um gráfico mostrando a produção dos 10 primeiros autores (em termos de número de publicações e número de citações totais por ano) no decorrer do tempo.

topAU <- authorProdOverTime(M, k = 10, graph = TRUE)



#### Estimativa do coeficiente da regra de Lotka

A regra de Lotka descreve a frequência de publicação dos autores de um determinado campo como uma lei do quadrado inversa, onde o número de autores publicando um determinado número de artigos é uma taxa fixa para o número de autores publicando um artigo único. Essa suposição implica que o coeficiente beta teórico da regra de Lotka é igual a 2. A função “lotka” estima os coeficientes dessa regra de Lotka. Usando tal função é possível estimar por exemplo o coeficiente beta, dito anteriormente, de nossa coleção bibliográfica e acessar - a partir de um teste estatístico - a similaridade da distribuição empírica com uma teórica.

* Distribuição empírica da produtividade dos autores:

L <- lotka(results)  
L$AuthorProd

## N.Articles N.Authors Freq  
## 1 1 2248 0.878125000  
## 2 2 224 0.087500000  
## 3 3 48 0.018750000  
## 4 4 23 0.008984375  
## 5 5 6 0.002343750  
## 6 6 4 0.001562500  
## 7 7 2 0.000781250  
## 8 8 1 0.000390625  
## 9 9 2 0.000781250  
## 10 10 1 0.000390625  
## 11 11 1 0.000390625

* Coeficiente beta estimado:

L$Beta

## [1] 3.370245

* Constante:

L$C

## [1] 0.7867831

* Qualidade de ajuste:

L$R2

## [1] 0.980799

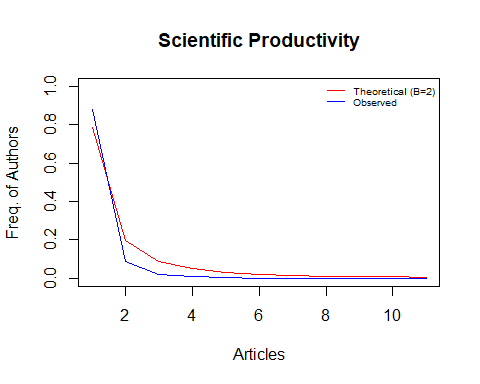
* Valor p dos dois testes K-S da amostra

L$p.value

## [1] 0.02325117

A tabela L$AuthorProd mostra a distribuição da produção científica observada em nosso exemplo. Nosso coeficiente beta é de aproximadamente 3.3 com uma qualiidade de ajuste de 0.98. Os dois testes (estatísticos) Kolmogorov-Smirnoff nos provê um valor p de 0.02, o que significa que basicamente não há diferença estatística entre nossos distribuições Lotka esperada e observada. Para uma visão gráfica disso, é possível usar a função plot:

Observed=L$AuthorProd[,3]  
Theoretical=10^(log10(L$C)-2\*log10(L$AuthorProd[,1]))  
plot(L$AuthorProd[,1],Theoretical,type="l",col="red",ylim=c(0, 1), xlab="Articles",ylab="Freq. of Authors",main="Scientific Productivity")  
lines(L$AuthorProd[,1],Observed,col="blue")  
legend(x="topright",c("Theoretical (B=2)","Observed"),col=c("red","blue"),lty = c(1,1,1),cex=0.6,bty="n")



#### Matrizes de Redes Bibliográficas

As características de um artigo científico estão todas conectadas entre si a partir do próprio artigo, via os próprios autores ao revista ou periódico publicado, as palavras chaves com a data de publicação, etc. As conexões entre diferentes características geram redes bipartidas que podem ser representadas na forma de matrizes retangulares (artigo x qualquer que seja a característica). Além disso, publicação científicas normalmente possuem referências a outros artigos. Isso acaba por gerar uma nova rede, chamada rede de acoplamento ou de co-citação. Essas redes são utilizadas para analizar e capturar o real significado de certas propriedades em sistemas de pesquisas e, em particular, determinar a influência de algumas unidades bibliométricas como revistas e periódicos.

#### Redes bipartidas

A função “cocMatrix” é usada para computar uma rede bipartida selecionando um dos atributos dos dados, por exemplo, uma rede Artigo x Fonte da Publicação:

A <- cocMatrix(M, Field = "SO", sep = ";")  
sort(Matrix::colSums(A), decreasing = TRUE)[1:5]

## JOURNAL OF IMMUNOLOGY EUROPEAN JOURNAL OF IMMUNOLOGY NATURE IMMUNOLOGY   
## 158 32 26   
## JOURNAL OF EXPERIMENTAL MEDICINE BLOOD   
## 16 16

Seguindo a mesma lógica, é possível fazer outros tipos de redes bipartidárias:

* Rede de citações

A <- cocMatrix(M, Field = "CR", sep = ". ")  
sort(Matrix::colSums(A), decreasing = TRUE)[1:5]

## WEISS A. LITTMAN D.R. SIGNAL TRANSDUCTION BY LYMPHOCYTE ANTIGEN RECEPTORS (1994)   
## 16   
## SMITH-GARVIN J.E. KORETZKY G.A. JORDAN M.S. T CELL ACTIVATION (2009)   
## 6   
## VON BOEHMER H. FEHLING H.J. STRUCTURE AND FUNCTION OF THE PRE-T CELL RECEPTOR (1997)   
## 6   
## ROBEY E. FOWLKES B.J. SELECTIVE EVENTS IN T CELL DEVELOPMENT (1994)   
## 5   
## STARR T.K. JAMESON S.C. HOGQUIST K.A. POSITIVE AND NEGATIVE SELECTION OF T CELLS (2003)   
## 4

* Rede de Autores:

A <- cocMatrix(M, Field = "AU", sep = ";")  
sort(Matrix::colSums(A), decreasing = TRUE)[1:5]

## WEISS A SAMELSON LE LOVE PE SCHRAVEN B ACUTO O   
## 11 10 9 9 8

* Rede de palavras-chaves (dos autores):

A <- cocMatrix(M, Field = "DE", sep = ";")  
sort(Matrix::colSums(A), decreasing = TRUE)[1:5]

## SIGNAL TRANSDUCTION TCR TCR SIGNALING T CELL T CELL RECEPTOR   
## 28 26 18 15 14

#### Redes de acoplamento

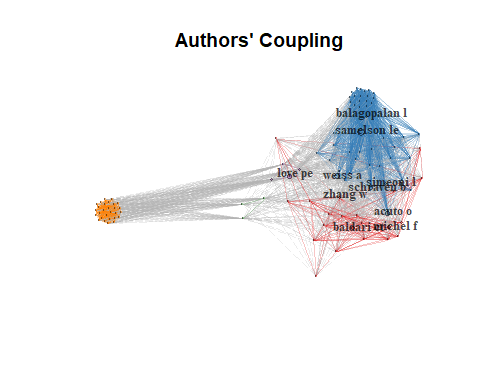
De acordo com Kessler (1993), podemos dizer que dois artigos estão acoplados bibliográficamente se pelo menos uma das fontes citadas aparece nas referências bibliográficas de ambos os artigos. A função “biblioNetwork” calcula, a partir de um data frame bibliográfico, as redes de acoplamento mais comumente usadas: de autores, de fontes e países.

Artigos com poucas referências, consequentemente, tendem a ficar menos bibliograficamente acoplados (mais distantes um do outro em um gráfico), se a “força” de acoplamento é medida simplesmente pelo número de referências que aquele artigo tem em comum com outros. Isso nos sugere que pode ser melhor utilizar uma medida relativa (e não absoluta) para o acoplamento bibliográfico.

Através da função “normalizeSimilarity”, é possível se calcular a força de associação, inclusão e a similaridade de Jaccard ou Salton entre os vértices de uma rede.

Como exemplo, segue uma rede de acomplamento para os autores de nossos dados:

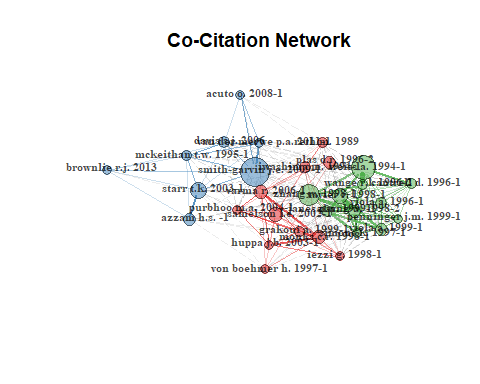
NetMatrix <- biblioNetwork(M, analysis = "coupling", network = "authors", sep = ";")  
net=networkPlot(NetMatrix, normalize = "salton", weighted=NULL, n = 100, Title = "Authors' Coupling", type = "fruchterman", size=5,size.cex=T,remove.multiple=TRUE,labelsize=0.8,label.n=10,label.cex=F)



#### Co-citação e colaboração bibliográfica

Dois artigos são co-citados quando ambos são citadas por um mesmo artigo. Como qualquer outra rede, é possível fazer uma para co-citação utilizando-se a função “biblioNetwork”.

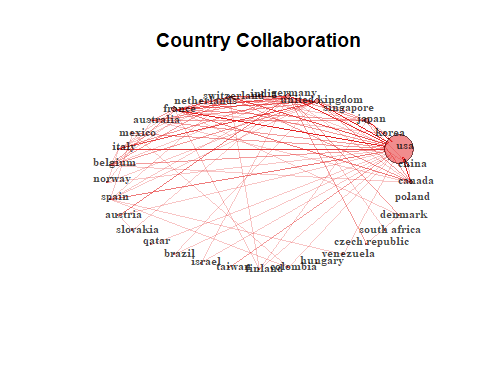
NetMatrix <- biblioNetwork(M, analysis = "co-citation", network = "references", sep = ";")  
net=networkPlot(NetMatrix, n = 30, Title = "Co-Citation Network", type = "fruchterman", size=T, remove.multiple=FALSE, labelsize=0.7,edgesize = 5)



Em uma rede de colaboração científica os nós são os autores enquanto que os links são as co-autorias de um artigo, pois já se sabe que o segundo é uma das formas mais bem documentadas de colaboração científica entre dois pesquisadores (de acordo com Glanzel, 2004).

Novamente, usando a função “biblioNetwork”, é possível calcular uma rede de colaboração científica por países, por exemplo:

M <- metaTagExtraction(M, Field = "AU\_CO", sep = ";")  
NetMatrix <- biblioNetwork(M, analysis = "collaboration", network = "countries", sep = ";")  
net=networkPlot(NetMatrix, n = dim(NetMatrix)[1], Title = "Country Collaboration", type = "circle", size=TRUE, remove.multiple=FALSE,labelsize=0.7,cluster="none")



#### Análise descritiva das características gráficas de redes

A função “networkStat” calcula várias estatísticas das nossas redes de forma resumida. Partindo de uma matriz bibliográfica, dois grupos de medidas descritivas são computadas:

* as estatísticas resumidas de uma rede

NetMatrix <- biblioNetwork(M, analysis = "co-occurrences", network = "keywords", sep = ";")  
netstat <- networkStat(NetMatrix)  
names(netstat$network)

## [1] "networkSize" "networkDensity" "networkTransitivity" "networkDiameter"   
## [5] "networkDegreeDist" "networkCentrDegree" "networkCentrCloseness" "networkCentrEigen"   
## [9] "networkCentrbetweenness" "NetworkAverPathLeng"

* e os principais indíces referentes a centralidade e prestígio dos vértices

NetMatrix <- biblioNetwork(M, analysis = "co-occurrences", network = "keywords", sep = ";")  
netstat <- networkStat(NetMatrix)  
names(netstat$vertex)

## NULL

Para resumir os principais resultados da função “networkStat”, é simples: usa-se a função “summary”, que nos mostrará as principais informações sobre a rede e a descrição dos vértices através de várias tabelas.

summary(netstat, k=10)

##   
##   
## Main statistics about the network  
##   
## Size 3275   
## Density 0.042   
## Transitivity 0.198   
## Diameter 3   
## Degree Centralization 0.907   
## Average path length 1.958   
##

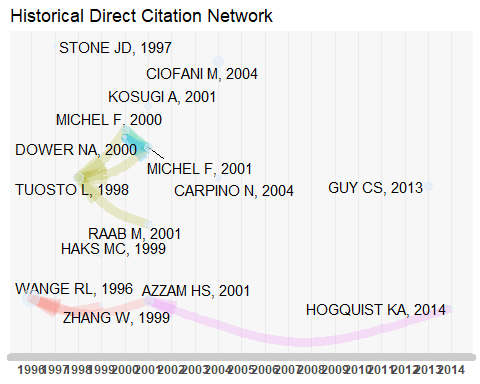
#### Rede histórica de citações diretas

Proposto por E. Garfield em 2004, o mapa historiográfico é um gráfico que representa uma rede cronológica mostrando as citações diretas mais importantes a partir da nossa coleção bibliográfica. Essa mapa é gerado a partir da função “histPlot” após se aplicar a função “histNetwork”, capaz de montar essa rede cronológica de citações diretas.

options(width=130)  
histResults <- histNetwork(M, min.citations = 10, sep = ";")

## Articles analysed 100   
## Articles analysed 200   
## Articles analysed 300   
## Articles analysed 340

net <- histPlot(histResults, n=15, size = 20, labelsize=10, size.cex=TRUE, arrowsize = 0.5, color = TRUE)



##   
## Legend  
##   
## Paper DOI Year LCS GCS  
## 1996 - 21 WANGE RL, 1996, IMMUNITY 10.1016/S1074-7613(00)80315-5 1996 19 443  
## 1997 - 33 STONE JD, 1997, J IMMUNOL <NA> 1997 5 85  
## 1998 - 54 TUOSTO L, 1998, EUR J IMMUNOL 10.1002/(SICI)1521-4141(199807)28:07<2131::AID-IMMU2131>3.0.CO;2-Q 1998 9 85  
## 1999 - 81 ZHANG W, 1999, INT IMMUNOL 10.1093/INTIMM/11.6.943 1999 5 218  
## 1999 - 87 HAKS MC, 1999, IMMUNITY 10.1016/S1074-7613(00)80084-9 1999 5 104  
## 2000 - 91 MICHEL F, 2000, J IMMUNOL 10.4049/JIMMUNOL.165.7.3820 2000 7 73  
## 2000 - 103 DOWER NA, 2000, NAT IMMUNOL 10.1038/79766 2000 5 334  
## 2001 - 115 AZZAM HS, 2001, J IMMUNOL 10.4049/JIMMUNOL.166.9.5464 2001 9 180  
## 2001 - 121 MICHEL F, 2001, IMMUNITY 10.1016/S1074-7613(01)00244-8 2001 7 128  
## 2001 - 122 RAAB M, 2001, IMMUNITY 10.1016/S1074-7613(01)00248-5 2001 7 84  
## 2001 - 127 KOSUGI A, 2001, IMMUNITY 10.1016/S1074-7613(01)00146-7 2001 6 85  
## 2004 - 173 CIOFANI M, 2004, J IMMUNOL 10.4049/JIMMUNOL.172.9.5230 2004 9 183  
## 2004 - 176 CARPINO N, 2004, IMMUNITY 10.1016/S1074-7613(03)00351-0 2004 6 100  
## 2013 - 303 GUY CS, 2013, NAT IMMUNOL 10.1038/NI.2538 2013 7 97  
## 2014 - 317 HOGQUIST KA, 2014, NAT IMMUNOL 10.1038/NI.2938 2014 5 91

## Respostas encontradas

Segue agora uma lista com todas os gráficos e tabelas montados durante o trabalho:

* Tabela contendo as principais informações de toda a coleção bibliográfica: número de documentos, fontes, palavras-chaves, tempo desde a publicação do primeiro ao último artigo, média de citações por artigo, informações referentes ao número de autores per se e por artigos, número de artigos por tipo de publicação, número de artigos publicados por ano, autores e países mais produtivos, artigos e países com mais citações, fontes e palavras chaves mais relevantes;
* Gráficos de autores e países mais produtivos, produção anual, média de citações por artigo e total por ano;
* Ranking de dominancia de autores;
* O índice-h dos principais autores;
* Gráfico e tabela mostrando a produtividade com o passar dos anos dos principais autores;
* Gráfico comparando a produtividade cientifica observada e teórica, dentro do tema escolhido;
* Vários gráficos demonstrando diferentes rede bibliográficas: de autores, de citação de colaboração entre países;
* Por fim, um gráfico demonstrativo de uma rede histórica das citações históricas mais importantes dentro de nossos dados

## Dificuldades encontradas

As principais dificuldades foram relativas à algumas características referentes ao próprio Rmarkdown. Enquanto o guia do pacote “Bibliometrix” era muito bom e intuivo, o Rmarkdown teve alguns problemas. O principal deles foi referente ao salvamente (knit) para outros formatos. Mesmo com o software Miktex tendo sido instalado, não houve a mínima possibilidade de se salvar o script em pdf. Além disso, no formato word todos os gráficos e tabelas acabavam ficando desconfigurados em relação ao seu tamanho original e, por causa disso, acabou por se escolher como formato final de visualização o html. Além disso, outra dificuldade foi a em encontrar funções ou formas de se fazer o negrito. itálico e sublinhado no próprio rmarkdown, vide que a página de help dele dentro do próprio RStudio ou mesmo no Cran não detinham tal informação (ou se detinham estavam muito bem escondidas pois o autor desse trabalho e outros cinco alunos foram incapazes de encontrá-la). Outro problema com o próprio Rmarkdown foi a incapacidade de se inserir logo no começo do desenvolvimento as funções install.packages para o pacote do Rmarkdown per se e do Bibliometrics, pois repetidas vezes erros eram dados e o salvamento (knit) era impossibilitado de ser feito. Dito tudo isso, o trabalho foi relativamente fácil de se fazer, garantindo uma confiança muito maior ao autor deste que se acostumou a procurar a ajuda de diversas funções na tentativa de entende-las melhor e aplicar em seus próprios projetos algumas ideas obtidas no decorrer da execução deste.

## Bibliografia

Aria, M. & Cuccurullo, C. (2019). A brief introduction to bibliometrix (<https://cran.r-project.org/web/packages/bibliometrix/vignettes/bibliometrix-vignette.html>).

Arquivo de Referência do Pacote “rmarkdown” (<https://cran.r-project.org/web/packages/rmarkdown/rmarkdown.pdf>).

Arquivo de Referência do Pacote “bibliometrix” (<https://cran.r-project.org/web/packages/bibliometrix/bibliometrix.pdf>).

Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, 11(4), pp 959-975, Elsevier, DOI: 10.1016/j.joi.2017.08.007 (<https://doi.org/10.1016/j.joi.2017.08.007>).

Cuccurullo, C., Aria, M., & Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains, Scientometrics, DOI: 10.1007/s11192-016-1948-8 (<https://doi.org/10.1007/s11192-016-1948-8>).

Cuccurullo, C., Aria, M., & Sarto, F. (2015). Twenty years of research on performance management in business and public administration domains. Presentation at the Correspondence Analysis and Related Methods conference (CARME 2015) in September 2015 (<http://www.bibliometrix.org/documents/2015Carme_cuccurulloetal.pdf>).

Sarto, F., Cuccurullo, C., & Aria, M. (2014). Exploring healthcare governance literature: systematic review and paths for future research. Mecosan (<http://www.francoangeli.it/Riviste/Scheda_Rivista.aspx?IDarticolo=52780&lingua=en>).

Cuccurullo, C., Aria, M., & Sarto, F. (2013). Twenty years of research on performance management in business and public administration domains. In Academy of Management Proceedings (Vol. 2013, No. 1, p. 14270). Academy of Management (<https://doi.org/10.5465/AMBPP.2013.14270abstract>).