Trabalho Final - R para Iniciantes

Maurílio Bonora Júnior

26/04/2019

## 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”. 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

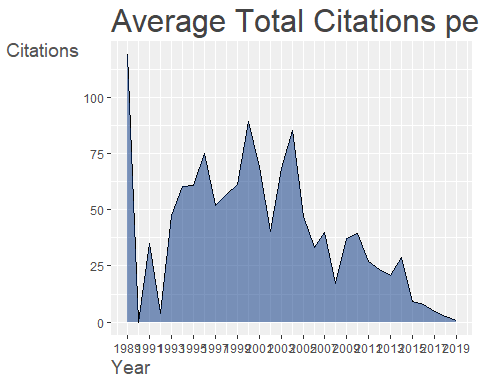
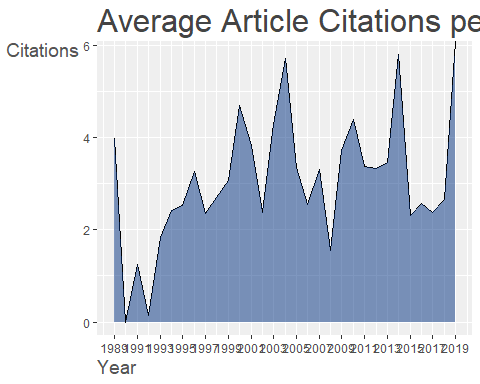
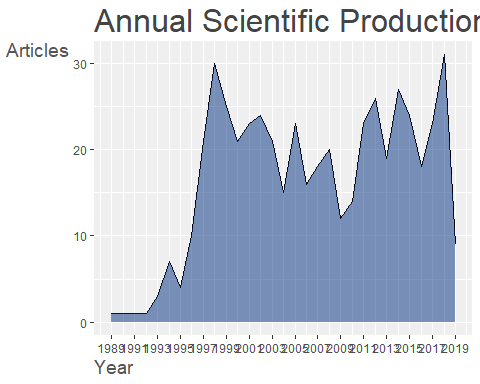
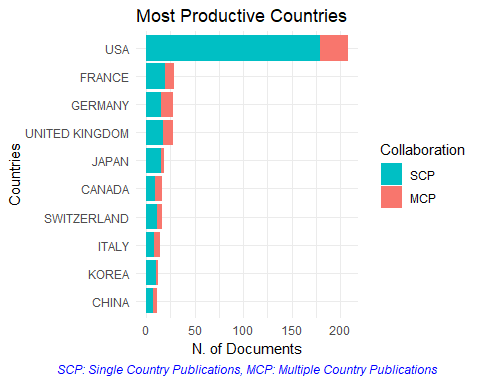
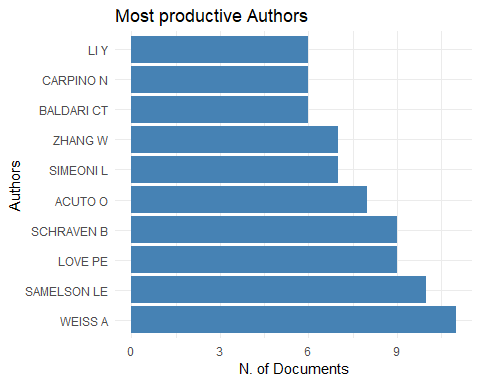
Contudo, o objeto “results” que obtemos ainda é muito grande e complicado de se ler, sendo muito laborioso obter informações gerais a partir dali. Para isso, usa-se a função “summary” para se obter os principais resultados da nossa análise bibliométrica. Essa função aceita ainda dois argumentos adicionais: “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



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