Real Citation Count

Kong

Saturday, January 10, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

**What is real citation?**

Typically, knowledge is spreading by means of **citations**. The more citations an article earns, the lagrer scientific impact it exerts. But knowledge is not equally expanded by all its citing papers. Some of them may be highly cited, while others are poor. So when we are trying to measure the scientific impact of the articles, we cannot treat the citing papers equally. Then I came up with the idea: Real Citation Index.

Real citation of an article is defined as follows:

Suppose that we have an article which has *n* citing papers. For every paper that citing , they have a citing number respectively. Now I have two ways to calculate the real citation number .

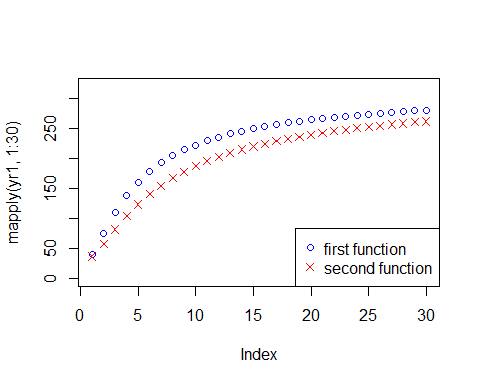
=

=

Now I will write an example to compare the two functions.

#We use a vector x to express the citations of every citing paper  
x <- c(1,4,10,7,35,50,0,21,19,25,28,21,17,1,16,50,26,55,19,12,14,15,68,25,8,16,53,33,202,315,11,147,7)

#cr1 to express the first function  
cr1 <- function(x,theta){  
 x <- as.numeric(x)  
 y <- 0   
 y <- ((sum(x^(theta)))/length(x))^(1/(theta))   
 y   
}   
  
#cr2 to express the second funtion  
cr2 <- function(x,theta){  
 x <- as.numeric(x)  
 y <- 0   
 y <- (sum(x^theta))^(1/(theta+1))   
 y   
}  
  
#Then we write two funcitons to plot the trends of cr1 & cr2.  
  
yr1 <- function(theta){cr1(x,theta)}  
yr2 <- function(theta){cr2(x,theta)}



**Here I think the first function is more logical,so I will focus on the discussion on the first fucntion.**

# Main features

As we can see, when gets bigger, the will gets closer to the maximum value of the .

研究思路： 1.首先需要下载数据，从Web of Science数据库，可以先从某一个研究者入手，然后下载全部的文献，并且下载引文的数据，这样就可以获得 **引文的引文**的数据，然后构建数据库。

2.我们讨论 的取值，可能在某一个值的情况下，刚好可以把文献分为两部分，一部分比较高，一部分则比较低。应该说，我们可以根据 的取值，然后比较和的大小，将文献分为两类。

3.可以根据真实引用值去预测以后的被引频率，可能对于每一篇文献，我们都可以找到一个较好的 值来预测，从而得到最好的拟合结果

4.利用真实引用来描述*h*指数，这种基于真实引用的*h*指数，是否更能反映一个研究者某个时期的影响力？

5.如何预测单篇文献的被引趋势？

- 在文献发表的初期，被引的可能性主要取决于文献发表的期刊以及作者的影响力，而作者的影响力则主要取决于他发表的那些高质量论文。  
- 在文献发表的一段时间后，这主要取决于文章已经获得的引用数量。  
- 同时，一个领域内研究者的活跃度对论文的被引产生很大影响。

6.对于不同的研究者来说，合适的 可能取值也不一样。

7.有可能对于诺奖获得者来说，可能在某一个阶段，他的真实被引达到了一个拐点，很有可能存在这样的特征。

8.我们可以做一个对比，论文的被引频次和真实被引的比较。

9.有些引文可能发表日期比较晚，没有什么引用，这部分问题值得讨论（方法的缺陷）。如果近几年的引文比较多，则很明显会影响到结果，是否考虑选取3或5年之前发表的引文。另外，引文的引文会出现重复，如何去掉重复？

10.精炼只选article。或者把所有的都选上。被引次数相同的文献，可能也会差异很大

## Data collecting and database build

We should collect a series of papers ,their citations should vary from 10,20,30,50,80,100,200,500,700,1000,1500,2000,3000. Then we should choose a series of to compare and .

library(xlsx) d2389 <- read.xlsx2("2389.xlsx",1,startRow = 1, endRow = 2390,colClasses = c(rep("character",6),"Date","numeric",rep("character",11),rep("numeric",118)))

d2389 <- read.xlsx2("2389.xlsx",1,startRow = 1, endRow = 2390)

Then I tried that when there are a lot of citing articles with very low citations will result in a very low . So I have to remove part of the data.First I should draw a picture to decide from which year(publication age) the data should be removed.

library(rCharts)  
# d2389 <- d2389[,c(1,2,6,8,20:137)]  
# colnames(d2389) <- c("ti","au","so","py","tc","avc",1900:2015)  
# table(d2389$py)  
# d2389 <- d2389[c(1:6,115:122)]  
# d2008 <- subset(d2389, py == 2008)# publication year == 2008  
# d2008 <- d2008[7:14, ]  
#   
# dd2008 <- data.frame(year = NA, d2008) #plus year as avaraible  
# d2008$year <-as.numeric(rownames(d2008))  
# m1 <- mPlot(x = 2008:2015, y = d2008[,1:5],data = d2008)

## 疑问

1.很明显，对于不同的文献来说，他们达到最大引用以及后续的趋势不一样。 2.同样，为什么那些超高被引的文献，他们的值一定会变小，这个问题该如何解释？

3.如果一篇文献被引越多，由于为0的引用越多，反而使其真实引用值变小，显然不合理。 4.其次在引文的引文中，也有很多被引为0的文献，如果按照原来的思路，这部分也要做考虑的。

options(java.parameters = "Xmx4096m")  
#set java.parameters

先把其他的都算出来比较一下再说

我们可以选择一个theta值，让他接近于100

rsdata <- function(x){  
 x <- x[,c(1,2,6,8,20:137)]  
 colnames(x) <- c("ti","au","so","py","tc","avc",1900:2015)  
 return(x)  
}  
  
rdxlsx <- function(file,start,end){  
 x <- read.xlsx2(file,1,startRow = start, endRow =   
 end,colClasses = c(rep("character",6  
 ),"Date","numeric",rep("character",11),rep  
 ("numeric",118)))  
}

we can now do the nls fit,see if the real citation have an better effect. the selected article 鏍囬: (The Hedgehog and Wnt signaling pathways in cancer)

x <- 1:14  
y <- c(8,50,57,76,56,84,63,64,56,63,40,62,40,44)  
z <- vector()  
for (i in 1:14) {  
 z[i] <- sum(y[1:i])  
}  
fit <- nls(z ~ 40 \* (exp(a1 \* (pnorm((log(x)-a2)/a3))-1)),start = list(a1=3,a2=3,a3=2))  
  
mcmodel <- function(x,y,m = 35,data,lambda = 2,mu = 2,sigma = 2){  
 fit <- nls(y ~ m \* (exp(lambda \* (pnorm((log(x)-mu)/sigma))-1)),start = list(lambda=lambda,mu=mu,sigma=sigma))  
 fit  
}

写一个函数，如何计算出每一年的真是引用值。

#calculate the real citation value of the selected year  
  
cry <- function(data,year,theta = 1,m = 35){  
 data <- data[!is.na(data[,"py"]),] #delete the data that have no pulication years  
 year <- as.numeric(year)  
 if(year < min(as.numeric(data[,"py"]),na.rm = TRUE)) {  
 stop("Invalid year:too small!")  
   
 }  
 ydata <- subset(data, py <= year) # find the necessary articles  
 n1 <- dim(ydata)[1] #find number of articles  
 n2 <- dim(ydata)[2] # number of columns  
 h <- vector() #storage the citation of each paper  
 for (i in 1:n1){  
 puby <- as.numeric(ydata[i,"py"])  
 h[i] <- sum(ydata[i,(122-(2015-puby)):(122-(2015-year))],na.rm = TRUE) # count number of citaitons till this year   
 }  
   
 cdata <- NULL  
 cdata$rc <- round(sum(log((3\*h/m)^theta+1)),3)  
 cdata$ct <- n1 #ordinary citation  
 as.data.frame(cdata)  
}  
# m represents the average citations in a research area  
  
#But often review has much more citations than ordinary #research articles  
  
twocit <- function(data,theta = 1, m = 35){  
 data <- data[!is.na(data[,"py"]),] #delete the data that have no pulication years  
 miny <- min(as.numeric(data[,"py"]),na.rm = TRUE)  
 xa <- NULL  
 for (year in miny:2014){  
 xa <- rbind(xa,cry(data,year,theta,m))   
 }  
   
 xa$year <- 1:(2014-min(as.numeric(data[,"py"]))+1)  
 xa  
}

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

You can also embed plots, for example:



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