科学计量与知识图谱系列丛书

R 科学计量数据可视化

USING R FOR SCIENTOMETRICS DATA VISUALIZATION

李 杰 编著

科学计量数据可视化是我们认识所关注领域过去、现在和未来的一种有益的方法！

首都经济贸易大学出版社
Capital University of Economics and Business Press

科学计量与知识图谱系列丛书

R 科学计量数据可视化

USING R FOR SCIENTOMETRICS DATA VISUALIZATION

李　杰　编著

首都经济贸易大学出版社
Capital University of Economics and Business Press
·北 京·

图书在版编目（CIP）数据

R科学计量数据可视化 / 李杰编著. --北京：首都经济贸易大学出版社，2018.6

ISBN 978-7-5638-2805-0

Ⅰ. ①R… Ⅱ. ①李… Ⅲ. ①科学计量学—可视化软件 Ⅳ. ①G301

中国版本图书馆CIP数据核字（2018）第110938号

R科学计量数据可视化

李杰 编著

责任编辑 薛晓红

封面设计 风得信·阿东 FondesyDesign

出版发行 首都经济贸易大学出版社

地 址 北京市朝阳区红庙（邮编 100026）

电 话 （010）65976483 65065761 65071505（传真）

网 址 http://www.sjmcb.com

E - mail publish@cueb.edu.cn

经 销 全国新华书店

照 排 北京砚祥志远激光照排技术有限公司

印 刷 北京玺诚印务有限公司

开 本 710毫米×1000毫米 1/16

字 数 238千字

印 张 13.5

版 次 2018年6月第1版 2018年6月第1次印刷

书 号 ISBN 978-7-5638-2805-0 / G · 417

定 价 48.00元

Preface

We heard about bibliometrics 10 years ago for the first time. In 2008 Corrado was writing a monograph on fast growing firms, a niche theme, which he approached for the first time. Scientific literature was fairly limited. Scholars came from different disciplines with a variety of approaches and methods that made it difficult to cumulate the findings.

We talked about this research problem once during a football match among scholars. Our discussion continued for several days on the various techniques of systematic analysis of literature. We enjoyed the exchange and concluded that bibliometrics was an interesting method and that it would have been fun to explore it together.

Our goal became to examine the intellectual structure of fast growing firms research. We analyzed all the scientific production published in academic English-written journals. The analysis was complex because it required several steps and diverse analysis and mapping software tools, which were often available only under commercial licenses. All the process was unwieldy, from data-collection to data-visualization. Massimo greatly contributed with his statistical and coding skills. Our collaboration continued in moments of fun, such as our frequent football matches. While analyzing data, we discovered that we enjoyed working together. In short, our friendship soon turned into a scientific collaboration that still lasts.

Within our departments and academic communities, the reaction to our work was positive. At that time, few people talked about bibliometrics in Italy, even from the point of view of research evaluation. Years later we presented a bibliometric analysis paper on performance management at the Annual Conference of the Academy of Management, the largest international management meeting. Also on that occasion, we got positive feedbacks that pushed us to persist. In the same years, young Italian colleagues asked us

for suggestions for their literature reviews and for their research. Massimo opened some statistical analysis laboratories in R and together we presented the bibliometric analysis at some workshops.

We are telling this story because without these feedbacks and stimuli we would not have published the bibliometrix release 0.1 in 2016. A year later we are at version 1.7, thanks to our growing passion for bibliometrics and to the suggestions that today come from scholars from all around the world.

R-bibliometrix is currently a free tool for quantitative research in scientometrics and bibliometrics that includes all the main bibliometric methods of analysis, easy to use even for those who have no coding skills.

Bibliometrix is a unique tool, developed in the statistical computing and graphic R language, according to a logical bibliometric workflow. R is highly extensible because it is an object-oriented and functional programming language, and therefore is pretty easy to automate analyses and create new functions. As it has an open-software nature, it is also easy to get help from the users community, mainly composed by prominent statisticians. Therefore, bibliometrix is flexible and can be rapidly upgraded and can be integrated with other statistical R-packages. That why, it is useful in a constantly changing science such as bibliometrics.

Today bibliometrix is more than just a statistical tool. It is becoming a community of international developers and users who exchange questions, impressions, opinions, and examples within an open source project. For this reason, we are very honored that Dr Jie Li of the Research center for Safety and security SCITECH trends at the Department of Safety Science and Engineering, Shanghai Maritime University gave us the opportunity to tell you this story and to write an English preface for his book "Using R for Scientometrics data Visualization" that mainly introduces the BIBLIOMETRIX package to scholars and students.

We said that Bibliometrix includes all the main bibliometric methods of analysis, but we use it especially for science mapping and not for measuring

science, scientists, or scientific productivity. Synthesizing past research findings is one of the most important tasks in advancing a line of research. Various methods exist to summarize the amount of scientific activity in a domain, but bibliometrics has the potential to introduce a systematic, transparent and reproducible review process. This is very relevant in an age when the number of academic publications is rising at a very fast pace and it is increasingly unfeasible to keep track of everything that is being published; and when the emphasis on empirical contributions is resulting in voluminous and fragmented research streams, and a contested field. Literature reviews are increasingly playing a crucial role in synthesizing past research findings to effectively use the existing knowledge base, advance a line of research, and give evidence-based insights into the practice of exercising and sustaining professional judgment and expertise. The overwhelming volume of new information, conceptual developments and data are the milieu in which bibliometrics becomes useful, by providing a structured analysis to a large body of information, to infer trends over time, themes researched, identify shifts in the boundaries of the disciplines, to detect most the prolific scholars and institutions, and to show the "big picture" of extant research.

Naples, Italy July 2017

Massimo Aria and Corrado Cuccurullo

前　言

当前，我们处于科学文献大数据时代。面对海量的文献我们如何快速地了解一个研究领域、研究方向或者主题的整体格局以及未来的趋势？在此背景下，与该问题直接相关，科学计量理论、方法和技术适时发展，成为解决上述科研问题的一种有效的途径，掌握科学计量相关的技术和方法也成为科研工作者在新时代进行科学研究活动的基本技能。在过去十余年里，科学计量数据可视化的理论与方法已经大量地渗透到其他学科的研究实践中。在国内，这种以科学文本数据为研究对象，通过可视化技术来揭示学科结构、演进和互动的研究领域被统称为“科学知识图谱”。

科学计量数据可视化背后涉及了大量的科学计量学（还包含文献计量学、网络计量学以及信息计量学）方面的基础理论，比如论文的作者生产率分布、论文的共被引、耦合、主题共现以及作者合作等。还包含了统计学和网络科学等方面的技术和方法，比如多维尺度分析、聚类分析、复杂网络分析、自然语言处理和文本挖掘等分析方法。上面的理论和方法构成了进行科学计量数据可视化分析的知识基础，是进行知识图谱分析的前提。在理论和方法的支持下，当前国内外的相关学者已经开发了数十种进行科技文本挖掘方面的软件或者工具包，这些知名的工具包含了 HistCite、BibExcel、CiteSpace、SCI2 以及 VOSviewer 等。这些工具为有意借助领域文献分析以获取学科研究格局和动态的学者提供了可能。

笔者在过去 5 年从事科学计量和知识图谱的实践研究中，相继撰写了关于 CiteSpace、VOSviewer 以及 BibExcel 等方面的书籍，主要目的在于帮助非科学计量学领域的学者快速应用该方法辅助科学研究；从 2016 年开始已经相继组织了 4 次与科学计量和知识图谱相关的活动，与来自国内的数百名知识图谱爱好者有过交流。在交流中，最为常见和令笔者反思的一个问题是：“我得到的图谱结果应该怎样解释呢？”笔者认为科学计量及知识图谱的方法仅仅给我们提供了一种认识知识世界的新方式，但这种认识方式更需要知识图谱实践者结合自身的专业背景和知识图谱的理论与方法去思考。在进行科学计量和知识图谱分析的时候，读者一定要明确自己要解决的问题是什么？以及为什么知识图谱能够解决提出的问题，它与其他方法相比优势在哪里？等等。即在进行科学计量和知识图谱分析之前，一定要确定自己所要研究的问题，然后选择要使用何种知识图谱呈现方式来

解决问题。

本书是《CiteSpace：科技文本挖掘及可视化》《科学计量与知识网络分析——基于 BibExcel 等软件的实践》《科学知识图谱原理及应用——VOSviewer 与 CiteNetExplorer 初学者指南》的姊妹篇。与前面这些知识图谱工具不同的是，本书详细介绍了意大利那不勒斯菲里德里克第二大学（University of Naples Federico II）经济与统计系 Massimo Aria 和 Corrado Cuccurullo 基于 R 语言开发的 BIBLIOMETRIX 工具包（Version 1.6 和 1.7）[1]。该 R 工具包基本上涵盖了进行科学计量和知识可视化的功能（图 0.1），可以满足爱好 R 软件，并试图使用 R 软件进行科学计量和知识图谱分析的读者。在此基础上，对与科学计量与知识图谱相关的一些 R 工具包，如 rAltmetric、wordcloud2、gender 以及 tidytext 等进行了介绍。本书对使用 R 软件进行英文全文本挖掘进行了很少的介绍，对中文全文本挖掘尚未涉及。在今后的更新中将对使用 R 软件进行全文本挖掘进行适当的完善。

图 0.1　bibliometrix 功能概览

为了便于读者熟悉 bibliometrix 工具包，大多数的案例运行采用了工具包自

[1] 读者在应用时，建议通过提供的链接来检查是否为最新版的 BIBLIOMETRIX，在实际的研究中尽可能地使用最新版来对数据进行分析。BIBLIOMETRIX-R Package for Bibliometric and Co-Citation Analysis http://www.bibliometrix.org/

带的数据，一些案例专门下载了 Web of Science 和 Scopus 数据集进行了分析。案例中呈现了所分析的结果，但并未就结果进行描述性或者带有特定研究目的的解读。通过对这些结果的学习，读者可以自己去思考可以做些什么？或者至少可以通过这种方法了解自己所关注领域的基本情况。

本书在撰写中有如下约定：

```
> 后为代码
# 为代码的说明
## 为代码运行的结果
```

感谢 Massimo Aria 和 Corrado Cuccurullo，他们在本书写作过程中给予了大力帮助，并为本书撰写了英文序言。感谢首都经济贸易大学出版社杨玲社长在科学计量与知识图谱系列丛书出版中的极大支持，感谢中国科学院李彬彬博士在提取子矩阵问题上的帮助，感谢滑铁卢大学博士后于淼对文稿提出的修改建议，感谢本书的责任编辑薛晓红以及研究生李平对本书的编辑和详细校对。

回首自己在科学计量和知识图谱研究与实践上的经历，感受五味杂陈。衷心地期望本书及相关系列丛书能进一步促进科学计量与知识图谱实践研究在国内的发展和普及，并使每一位读者受益。

李 杰

2018 年 5 月于北京

目　录

第 1 讲

R基础

1.1 R 下载

问题

如何下载 R 软件[1]？

方法

第 1 步：登录 CRAN 网站[2]，在主界面中点击 CRAN 链接（图 1.1），进入 R 的下载镜像链接。例如这里选择在中国的镜像链接来下载 R（图 1.2）。

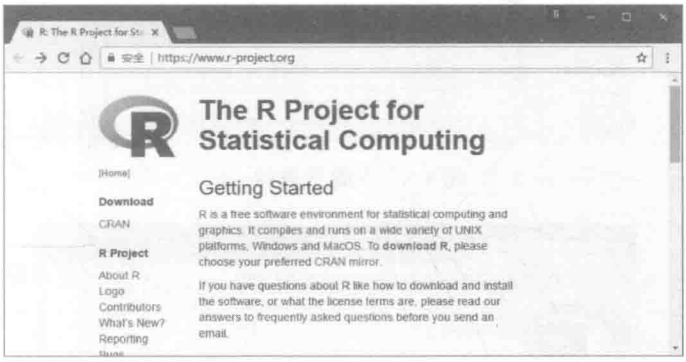

图 1.1 R 软件的下载主页

[1] R 语言是基于 S 语言开发的，S 语言则由贝尔实验室 John Chambers 和其同事于 1976 年共同开发。在 20 世纪 90 年代初，Ross Ihaka 和 Robert Gentleman 开发了 R 语言。

[2] R软件下载 https://www.r-project.org/; CRAN 是 Comprehensive R Archive Network 的简写，是拥有同一资料，包括 R 的发布版本、包、文档和源代码的网络集合。

图 1.2　R 下载的镜像文件

第 2 步：在 Download and Install R 中提供了 Mac 和 Windows 版本的 R 软件。此处我们使用的电脑系统为 Windows，因此这里选择 Download R for Windows（图 1.3）。在接下来的界面 R for Windows 中点击 install R for the first time（图 1.4）。

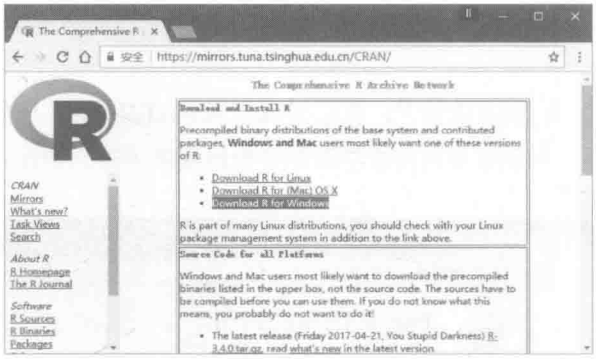

图 1.3　下载 R 软件 –1

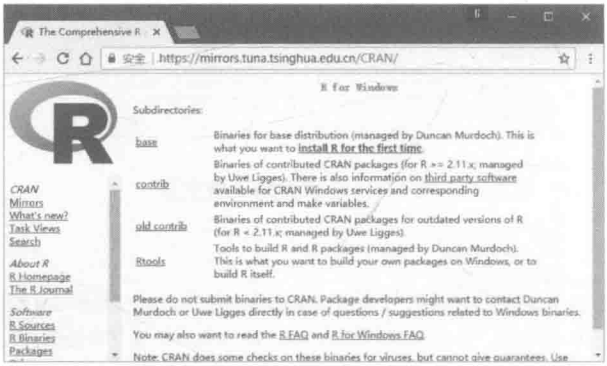

图 1.4　下载 R 软件 –2

第 3 步：本次下载 R 时的最新版本为 R-3.4.0 for Windows（图 1.5），在页面上点击 Download R 3.4.0 for Windows (76 megabytes, 32/64 bit) 即可下载 R 软件。

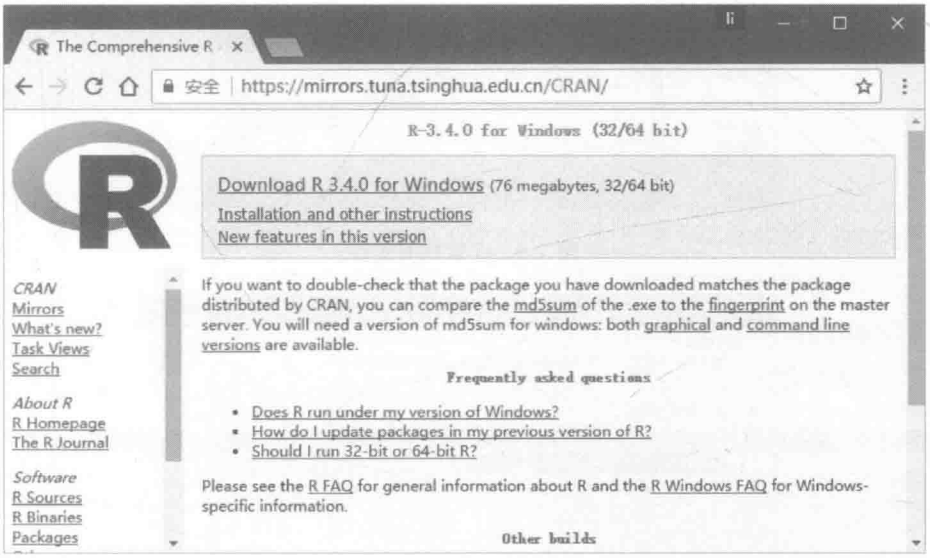

图 1.5　下载 R 软件 -3

1.2　R 安装

问题

如何安装 R 软件?

方法

双击所下载的 R-3.4.0-win.exe 安装包，并点击确定（图 1.6），进入 R 的安装向导。在“欢迎使用 R for Windows 3.4.0 安装向导”中点击下一步（图 1.7）。对初学者，若对软件没有特别的关注和要求，可以连续点击“下一步”直至软件完成安装（图 1.8）。

图 1.6　R 软件的安装

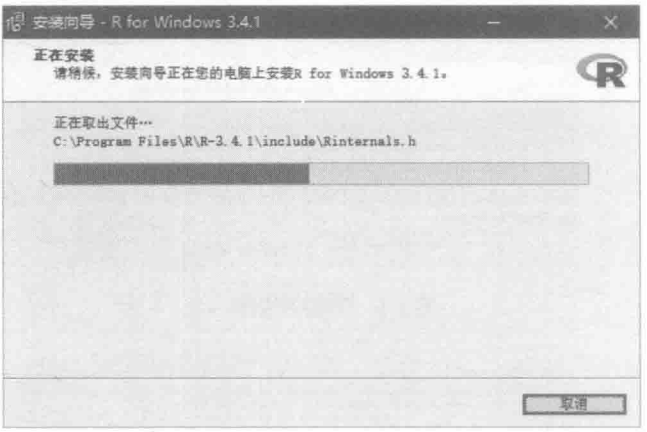

图 1.7　R 软件的安装进程

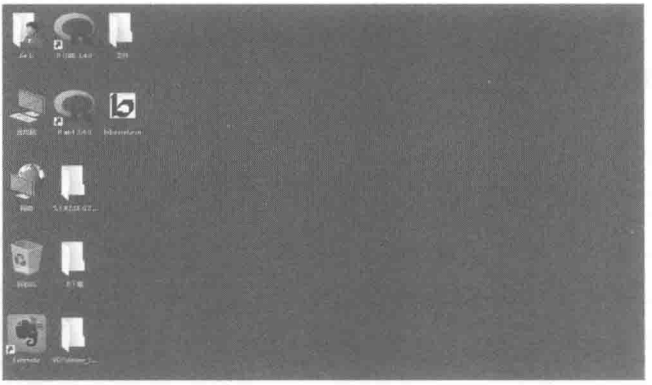

图 1.8　R 软件安装完成

1.3 Rstudio 安装

问题

如何安装 Rstudio？

方法

为了更好地使用 R，建议在安装 R 之后下载 Rstudio[1] 进行安装。安装 Rstudio 后，在以后的 R 使用中，相关操作可以在 Rstudio 中进行。

在 Rstudio 中基本包含了 4 个界面（图 1.9），左上角为代码界面，在这里写好代码后，点击 Run 即可运行所选择的代码；左下角为代码运行的过程窗口；右上角为代码环境和代码运行的历史界面，可以显示代码运行的基础结果和数据情况（例如数据的记录数和变量数）；右下角包含了绘图结果、软件包及其帮助界面。

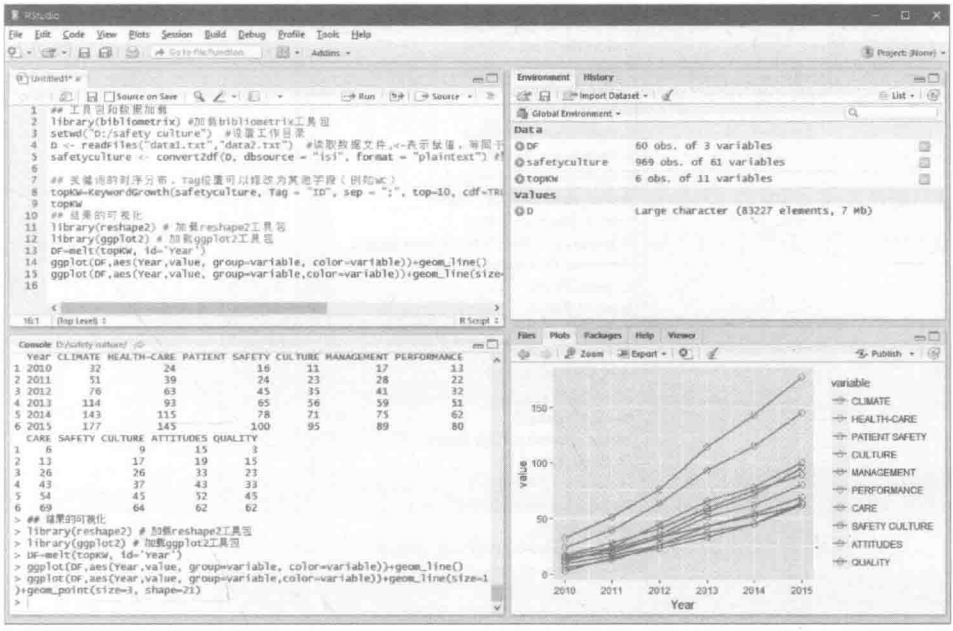

图 1.9 Rstudio 界面

[1] R Rstudio 主页：http://www.rstudio.com.
RStudio 的前世今生——RStudio 创始人专访．https://cosx.org/2016/11/interview-j-j-allaire/.