HZ BOOKS
华章IT

WILEY

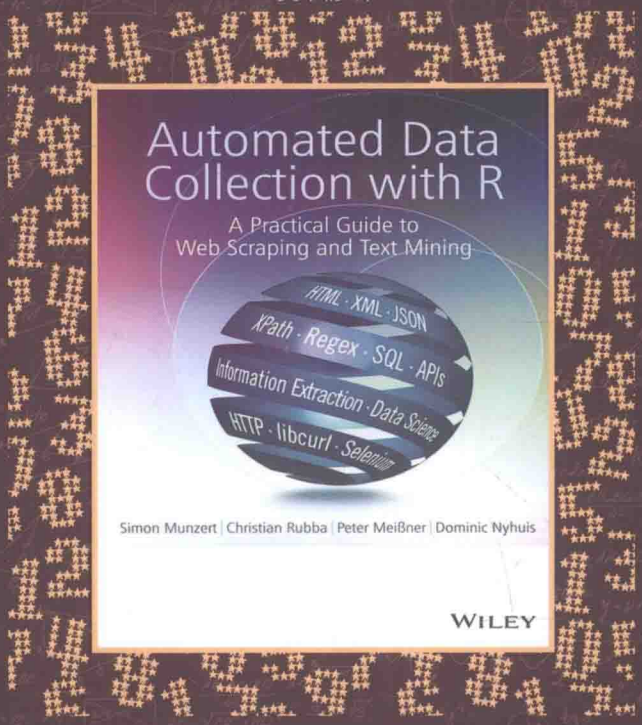
数据科学与工程技术丛书

基于R语言的自动数据收集

网络抓取和文本挖掘实用指南

西蒙·蒙策尔特（Simon Munzert）
克里斯蒂安·鲁巴（Christian Rubba）
彼得·迈博纳（Peter Meißner）
多米尼克·尼胡斯（Dominic Nyhuis）
[德] 著

吴今朝 译



AUTOMATED DATA COLLECTION WITH R

A Practical Guide to Web Scraping and Text Mining

机械工业出版社
China Machine Press

数据科学与工程技术丛书

AUTOMATED DATA COLLECTION WITH R

A Practical Guide to Web Scraping and Text Mining

基于R语言的自动数据收集

网络抓取和文本挖掘实用指南

[德] 西蒙·蒙策尔特（Simon Munzert）
克里斯蒂安·鲁巴（Christian Rubba）
彼得·迈博纳（Peter Meißner）
多米尼克·尼胡斯（Dominic Nyhuis） 著

吴今朝 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

基于R语言的自动数据收集：网络抓取和文本挖掘实用指南 /（德）蒙策尔特（Munzert, S.）等著；吴今朝译 . —北京：机械工业出版社，2016.2
（数据科学与工程技术丛书）
书名原文：Automated Data Collection with R：A Practical Guide to Web Scraping and Text Mining

ISBN 978-7-111-52750-3

I. 基… II. ①蒙… ②吴… III. ①程序语言－程序设计 ②数据采集 IV. ①TP312 ②TP274

中国版本图书馆CIP数据核字（2016）第014662号

本书版权登记号：图字：01-2015-4153

本书从社会科学研究者角度系统且深入阐释利用R语言进行自动化数据抓取和分析的工具、方法、原则和最佳实践。深入剖析自动化数据抓取和分析各个层面的问题，从网络和数据技术到网络抓取和文本挖掘的实用工具箱，重点阐释利用R语言进行自动化数据抓取和分析，能为社会科学研究者与开发人员设计、开发、维护和优化自动化数据抓取和分析提供有效指导。

出版发行：机械工业出版社（北京市西城区百万庄大街22号　邮政编码：100037）
责任编辑：秦　健　　责任校对：董纪丽
印　　刷：北京瑞德印刷有限公司　　版　　次：2016年3月第1版第1次印刷
开　　本：185mm×260mm　1/16　　印　　张：24
书　　号：ISBN 978-7-111-52750-3　　定　　价：99.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 88378991　88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

译 者 序

本书是以非计算机专业人士（尤其是社会科学领域的研究者）为目标读者的，但它对于广大开发者也有很好的参考价值。本书介绍的思路和方法并不仅仅局限于 R 语言的应用，在很多其他开发平台上也不难实现。

对我个人来说，这本书让我对社会科学和信息技术都有了全新的认识，开阔了眼界。大数据技术的应用领域除了搜索、电商、社交网络、垂直应用，还可以和很多专业领域结合起来，挖掘出非常有价值的信息。开源社区已经有了支持各种技术需求的现成组件，大大简化了所需的编程工作。可以说，这本书介绍的技术让大数据、网页抓取、机器学习这些貌似高大上和高深莫测的概念变得具体实际了。

在翻译本书过程中我最大的收获与其说是技术上的，不如说是理念上的：学科之间的交叉能够产生如此奇妙的反应，让很多我们以前想得到却做不到甚至根本不敢想的事情能够轻松地实现。尤其是在大数据时代，自动化数据抓取和文本挖掘技术为各专业领域的研究者提供了前所未有的强大工具，让社会科学家也能像自然科学家一样通过建模、采集数据、分析统计的过程产生量化的结果，以此来支持他们的分析和结论。

本书的核心内容是自动化数据抓取和分析的方法，R 语言及其一些组件在其中承担了基础架构的作用。比如，书中介绍了通过定期抓取 Twitter 相关推文对奥斯卡奖得主进行预测的案例，我们同样可以利用新浪微博提供的开放接口做到类似的事情（请参阅 http://open.weibo.com/wiki/2/search/topics）。利用 R 语言及其众多组件提供的支持，我们可以避开大量技术细节，专注于研究主题，真正需要编写的代码其实是相当简单的。

管中窥豹，可见一斑。我们还可以看到一个趋势：编程将不再是计算机专业人士的专利，而是一种越来越方便、越来越简单的工具。随着各种编程语言（如本书用到的 R 语言）及其配套工具的完善，几乎每个人都有机会具备基本的编程能力，就像现在大部分人都能学会开车和使用电脑上网一样。

本书就反映了上述趋势。在书中，作者给出了简洁的代码、详细的讲解以及真实的例子，让我们切切实实地看到大数据在社会科学领域运用的效果。作者尽可能回避晦涩的术语和高深的理论，为我们提供了非常实用的组件，并探讨了一些很有趣的实际问题。这样的讲解方式非常有利于我们快速上手、循序渐进学习，并且马上就能把学到的技术运用到实际研究项目中。

在翻译的过程中，我尽了最大努力让译文通顺易懂且忠于原文，但是，由于本人水平有限，难免会有不足和遗漏之处，望读者不吝指正，我在此先行感谢。另外，我在 GitHub 开辟了一个讨论区：https://github.com/coderLMN/AutomatedDataCollectionWithR/issues，欢迎读者在这里提出自己的疑问和观点并参与讨论。

最后，我要感谢我的父母、妻子和儿子在本书翻译过程中给予我的支持和理解。翻译是个精益求精的工作，是他们帮我分担了很多事情，才让我能全力以赴地投入。

前　言

过去20年，互联网的快速发展改变了我们分享、收集和发布数据的方式。企业、政府机构和个人用户都提供了各种类型的信息，新的沟通渠道也带来了有关人类行为的大量数据。社会科学领域曾经的根本性问题——观测数据稀缺和难以获取的情况——正在快速扭转为数据取之不尽用之不竭的局面。这种翻天覆地的形势也并非尽善尽美。例如，传统的数据采集和分析技术可能不足以应对复杂的大量数据。对这种大数据的需求进行分析的结果之一是所谓“数据科学家”的诞生，他们能对数据进行筛选，在研究者和企业那里都很受欢迎。

随着互联网的高歌猛进，我们还见证了第二个趋势，那就是像R这样的开源软件越来越流行，越来越有影响力。对计量社会科学家来说，R是最重要的分析软件之一。它得益于有一个不断发布新组件的活跃社区，而且该社区一直在快速成长。到现在，R已经不仅仅是一个免费统计软件包，它还包含了许多其他编程语言和软件包的接口，这样就大大简化了对各种来源的数据进行处理的工作。

从个人角度来说，我们对社会科学数据所做的工作的特点可以总结如下：

- 资金比较稀缺。
- 既没有时间也没有意愿进行数据的手工采集。
- 感兴趣的是利用最新、高质量和海量的数据来源。
- 需要记录从开始（数据采集）到结束（发布结果）的整个研究过程，这样它就可以被重现。

在过去，我们经常受困于对各种来源的数据进行手工整理，还要寄希望于手工整理不可避免带来的编码和复制－粘贴错误只是非系统性的。最终，我们越来越厌倦那种不可重现的研究数据采集方式，这种方式易于出错、缓慢复杂，而且提高了因烦躁而死的风险。因此，我们不断地把数据采集和发布流程纳入在统计分析过程中已熟悉的软件环境——R。这个程序提供了一套很好的基础架构，可以把日常工作流程扩展为实际数据分析前后的一系列步骤。

虽然目前R本身还不是用来采集数据或进行实验的，但我们还是认为本书讲述的技术不仅仅是对于成本高昂的调查、实验和学生助理编程者的“穷人的替代品”。我们相信它们是现代数据分析工具组合的有力补充。我们推崇对在线资源的数据进行采集，不仅认为它是比传统数据采集方法性价比更高的解决方案，更将其视为从新的和不断开发中的数据源中整合数据集的特有方法。此外，我们重视基于电脑程序的解决方案，因为它们能确保可靠性、重现能力、时间效率以及对高质量数据集的整合。除了工作效率，你还会发现自己乐于通过写代

码和设计算法方案替代乏味的手工劳动。简而言之，我们相信，如果你愿意花时间学习和采用本书中推荐的技术，在数据分析的简便性和质量上得到的持续提升一定会让你受益匪浅。

假定你已经确定在线数据是你的项目所适用的资源，那么是否真的有必要采用网络抓取或统计性文本处理技术，以及随之而来的自动化或半自动化数据采集流程？虽然我们不能指望拿出一锤定音的准则，但下面是一些有用的判断条件。如果你发现自己符合其中的多个条件，那么自动化的方法很可能就是正确选择：

- 你是否计划经常重复这项任务？比如，需要通过它来保持数据库的更新。
- 你是否需要让其他人能重复你的数据采集过程？
- 你是否经常处理在线的数据源？
- 这项任务在规模和复杂度上是否非同小可？
- 如果这项任务也可以手工完成……你是否缺乏必要的资源来调动其他人参与？
- 你是否愿意通过编程的手段实现自动化流程？

理想情况下，本书讲述的技术让你能够以相当合理的成本创建强大的数据集，这些数据集来自现有的、非结构化或未排序的数据，之前也没有人分析过它们。在很多情况下，根据你的研究主题的特点，你需要对本书讲述的技术重新思考、提炼和组合，才能有所成效。在任何情况下，我们都希望你能发现本书的主题对你有所启发，能开阔你的眼界：网络的街道是用数据铺成的，这些数据正迫不及待地等着被采集。

你从本书中不会学到的内容

当你浏览目录的时候，你会对阅读本书之后有望学到的东西有个初步的印象。虽然我们很难确定哪些部分是你希望看到却不在本书讨论范围内的，但是我们还是会指出你在本书中找不到的某几个方面的内容。

你在本书中不会看到对 R 环境的介绍。这方面已经有很多出色的介绍材料——不管是印刷版的还是在线的——本书不再赘述。如果你之前没有用过 R 语言，也大可不必失望地将此书束之高阁。我们还会推荐一些写得很好的 R 入门教材。

你也不要指望本书针对网络抓取或文本挖掘进行全面讲解。首先，我们专门使用了一套软件环境，而它并不是为实现这些目的量身定制的。在一些应用需求下，R 对于你要完成的任务并非理想解决方案，其他软件包可能更合适。我们也不会用如 PHP、Python、Ruby 或 Perl 等替代环境来干扰你。要想知道本书是否对你有帮助，你应该扪心自问，你是否已经或计划把 R 用在日常工作中。如果对这两个问题的答案都是否定的，很可能你就应该考虑替代方案了。但是，如果你已经在用或倾向于使用 R 了，你就可以省下学习另一个开发语言的精力，留在熟悉的开发环境里。

本书也不会严谨地介绍数据科学。在这个主题上也有一些出色的教材，例如 O'Neil and Schutt（2013）、Torgo（2010）、Zhao（2012），以及 Zumel and Mount（2014）。在这些书中偶尔缺失的部分是如何在真实环境中获取数据科学中用到的数据。在这方面，本

书可以作为数据分析的准备阶段的参考书，它还给出了关于如何管理可用信息并让它们保持及时更新的指导原则。

最后，你最不可能从本书看到的是针对你的具体问题的完美解答。在数据采集过程中，获取数据的领域从来都不会完全相似，而且其形式有时也会快速变化，这都是固有的问题。我们的目标是让你能改写例子和案例分析中提供的代码，并创建新的代码，以此帮助你在采集自己所需数据的工作中获得成功。

为什么使用 R

对于本书中涵盖的问题，R 是一个很好的解决方案，我们这么考虑是有很多原因的。对我们来说，最重要的几点原因如下：

- R 可以自由和简便地获得。你可以按自己的需要随时随地下载、安装和使用它。不去钻研那些昂贵的专有软件对你是大有裨益的，因为你不需要依赖于雇主支付软件版权费的意愿。
- 作为一个主要专注于统计学的软件环境，R 拥有一个巨大而且持续繁荣的社区。R 被用于各种专业领域，如社会科学、医学科学、心理学、生物学、地理学、语言学以及商业等。这样大的专业范围让你能与很多开发者共享代码，并从文档完善的多领域应用中获益。
- R 是开源的。这意味着你能够轻松地分析函数的工作原理并毫不费力地修改它们。这也意味着对程序的修改不会被一个维护产品的独家程序员团队所控制。即使你无意为 R 的开发贡献代码，你仍然可以从种类繁多的可选扩展项（组件）中获益。组件的数量与日俱增，很多已有的组件也会经常更新。你能在这里找到相当棒的关于 R 应用的流行主题的概述：http://cran.r-project.org/web/views/。
- 对常规任务而言，R 是相当快的。如果你用过类似于 SPSS 或 Stata 的其他统计软件，并养成了在计算复杂模型的时候顺便度个假的习惯，你应该会赞同这个印象，更不用提那种由“同一个会话，同一个数据框”的逻辑带来的切肤之痛了。甚至还有一些扩展可以用来加速 R，例如，在 R 的内部通过 Rcpp 组件调用 C 语言代码。
- R 在构建数据可视化效果方面也很强大。虽然这对于数据采集并非显著的增值，在日常工作中你还是不应该错过 R 的图形特色。我们后面会讲解到，对被采集数据的视觉检查能够且必须作为数据校验的第一步，以及图形如何给海量数据提供直观的总结方式。
- 使用 R 进行的工作主要是基于命令行的。这在 R“菜鸟”听起来也许像是一个不足，但相比那些要用鼠标点击的程序来说，这是唯一能支持产生可重现结果的方式。
- R 对于操作系统是不挑剔的。它通常可以在 Windows、Mac OS 和 Linux 下运行。
- 最后，R 是能自始至终支持研究过程的完整软件包。如果你在读这本书，你应该不是专职程序员，而是对于你要从事的某个主题或特定数据源有相当大的兴趣。

在这种情况下，学习另一门语言不会有成效，反而会让你无法开展研究工作。普通研究流程的一个例子如图 1 所示。它的特点是永远在各种程序之间切换。如果你需要对数据采集过程进行修正，你就不得不顺着整个梯子爬回去。而使用 R 的研究过程，正如在本书中所讲述的，只在单一的软件环境中进行（见图 2）。对于网络抓取和文本处理而言，这意味着你不必为这项任务去学习另一门编程语言。你需要学习的只是标记语言 HTML、XML、正则表达式逻辑和 XPath 的一些基础知识，但所需的操作都是在 R 内部执行的。

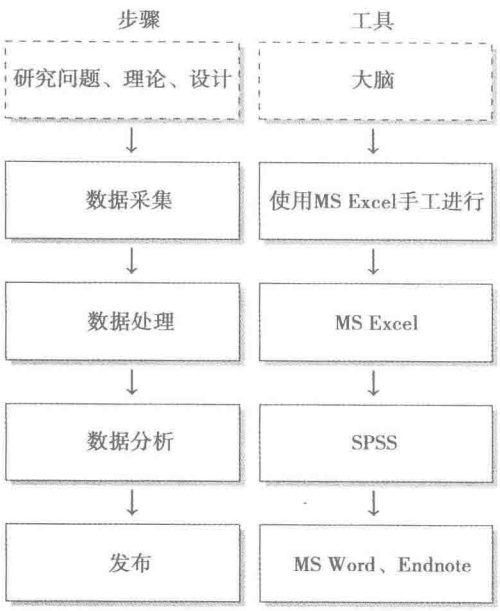


图 1　不使用 R 的研究过程——形象化的例子

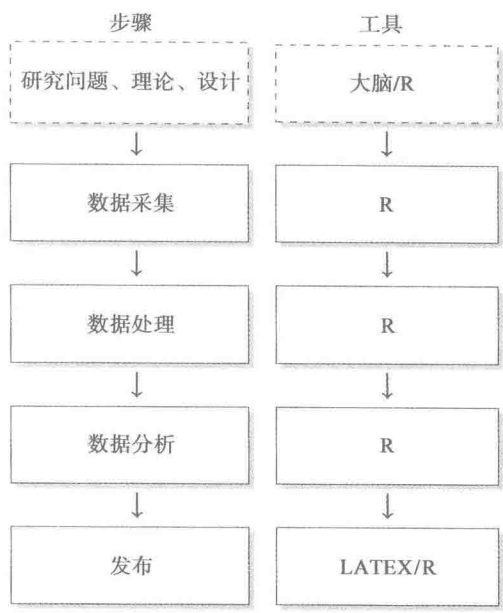


图 2　使用 R 的研究过程——形象化的例子

R 起步阶段的推荐读物

市面上有很多写得很好的介绍 R 的书。在它们当中，我们发现以下几本尤其有帮助：

Crawley, Michael J. 2012. *The R Book*, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Adler, Joseph. 2009. *R in a Nutshell. A Desktop Quick Reference*. Sebastopol, CA: O'Reilly.

Teetor, Paul. 2011. *R Cookbook*. Sebastopol, CA: O'Reilly.

除了这些商业化的资源，网上还有很多免费的信息。对绝对的新手来说，Code School 上有个真正超棒的在线教程，可以在 http://tryr.codeschool.com 看到。另外，Quick-R（http://www.statmethods.net/）里有很多基本命令的索引。最后，你也可以在 http://www.ats.ucla.edu/stat/r/ 找到很多免费资源和例子。

R 是一个不断成长中的软件，为了跟上它的进展，你或许需要定期访问 Planet R（http://planetr.stderr.org/），该网站提供了已有组件的发布历史，偶尔还会介绍一些有意思

的应用。R-Bloggers（http://www.r-bloggers.com/）是个博客大杂烩，它专门收集有关R的各种领域的博客。它提供了由数以百计的R应用构成的广阔视角，这些应用涉及的领域从经济学到生物学再到地理学，大部分都附有重现博文内容所必需的代码。R-Bloggers甚至推介了一些探讨自动化数据采集的例子。

当你遇到问题的时候，R的帮助文件有时候不是特别有用。去Stack Overflow（http://stackoverflow.com）这样的在线论坛或Stack Exchange网络旗下的其他站点寻求帮助往往会更有启发性。对于复杂问题，可以考虑去GitHub（http://github.com）上找一些R的专家。另外请注意，还有很多特别兴趣小组（SIG）的邮件列表（http://www.r-project.org/mail.html），里面划分了多种多样的主题，甚至还覆盖全世界的同城R用户小组（http://blog.revolutionanalytics.com/local-r-groups.html）。最后，有人建了个CRAN任务视图，较好地概括了近期Web技术的进展和R框架里的服务：http://cran.r-project.org/web/views/WebTechnologies.html。

配套资源

本书的配套网站见http://www.r-datacollection.com。

该网站提供了书中例子和案例分析的相关代码，以及其他一些内容。这意味着你无须手工从书中复制代码，直接访问和修改相应的R文件即可。你也可以在该网站找到某些练习题的解答，以及本书的勘误表。如果你在书中发现了任何错误，也请不吝告知。

免责声明

这不是一本关于网络蜘蛛（spider）抓取数据的书。网络蜘蛛是在互联网上抓取信息的程序，它能很快地从一个网页跳转到另一个网页，往往会抓取整个网站的内容。如果你想追随的是Google的Googlebot的足迹，那你很可能拿错书了。本书介绍的技术是用于更明确、更温柔的工作，也就是从特定的网站抓取特定的信息。最后，你要为自己学会这些技术之后的所作所为负责。本书示例的代码和得罪网站管理员的程序之间往往没有太大的鸿沟。所以，下面是关于如何做好一个网络数据采集从业者的一些重要忠告：

- 牢记你的数据是从何而来的，在可能的时候，感谢那些最初采集并发布它的人们。
- 如果你打算二次发布那些在网络上找到的数据，切勿违反版权协议。如果这些信息不是你自己采集的，有时你需要所有权人的许可才能对其进行加工。
- 不要做任何违法的事情！要了解你在数据采集过程中能做和不能做的事情，可以去Justia BlawgSearch（http://blawgsearch.justia.com/，这是一个搜索法律相关博客的网站）核对一下。在里边搜索被标记为“web scraping”（网络抓取）的结果有助于你了解相关法律的进展和近期的判决。另外，电子前沿基金会（http://www.eff.org/）早在1990年就成立了，它致力于保护消费者和大众的数字权利。

不过，我们希望你永远不需要仰仗他们的帮助。

关于从网络抓取内容的行为，本书 9.3.3 节提出了一些更详细的建议。

致谢

本书的问世有赖于很多人的帮助。我们想利用这个机会表达我们对他们的感激之情。首先我们想对 Peter Selb 说声谢谢，是他向我们提出了创建一门关于数据采集的新课程的构想。正是出自他的这一灵光闪现，我们才开始把之前相对零散的一些经验整理成一本全面的教材。我们也要向对本书部分章节提出宝贵意见的几位朋友表示感谢。我们重点感谢 Christian Breunig、Holger Döring、Daniel Eckert、Johannes Kleibl、Philip Leifeld 和 Nils Weidmann，他们的意见极大地提高了本书的水平。我们还要感谢 Kathryn Uhrig 对手稿的校对工作。

本书的早期版本用于 Konstanz 大学 2012 ～ 2013 学年夏季学期的“新的数据采集方法”和“从互联网采集数据”两门课程。我们要感谢各位同学提出的意见，以及他们在学习本书、R 以及正则表达式的时候表现出来的耐心。我们还要感谢参加了 2012 年 12 月在曼海姆召开的“助力实证性政治改革研究：R 自动化数据采集”和 2013 年 4 月在苏黎世召开的“R 自动化在线数据采集”两个研讨会的朋友。我们要特别感谢 Bruno Wüest 对苏黎世研讨会成功举办的协助，以及 Fabrizio Gilardi 的热心支持。

现在看来，编写一本有关自动化数据采集的著作是相当耗费时间的。我们在攻读博士学位期间都在从事这个项目，花费了很多本应用在自己课题上的时间来研究复杂的网络抓取技术。我们要感谢我们的导师 Peter Selb、Daniel Bochsler、Ulrich Sieberer 和 Thomas Gschwend 在我们走弯路时候的耐心和支持。Christian Rubba 还要感谢瑞士国家科学基金会的慷慨资助（经费授予号：137805）。

我们还要衷心感谢在本书中运用的众多组件的创建者和维护者。是他们的持续投入开启了新的学术研究的大门，并为个人研究者开辟了通往广阔数据源的道路。虽然我们没办法在这一段中提到所有的组件开发者，我们还是要对 Duncan Temple Lang 和 Hadley Wickham 的杰出工作表示感谢。我们还要感谢 Yihui Xie 开发的组件，它对本书的排版尤为关键。

我们要感谢来自出版社同仁的帮助，特别是 Heather Kay、Debbie Jupe、Jo Taylor、Richard Davies、Baljinder Kaur 和其他负责校对与排版的朋友，以及在写作过程中各个阶段提供帮助的朋友。

最后，我们要很高兴地向朋友和家庭成员的支持表示感谢。我们特别要衷心感谢 Karima Bousbah、Johanna Flock、Hans-Holger Friedrich、Dirk Heinecke、Stefanie Klingler、Kristin Lindemann、Verena Mack 以及 Alice Mohr。

Simon Munzert
Christian Rubba
Peter Meißner
Dominic Nyhuis

目　录

第 1 章
概　　述

你是否准备好第一次尝试网络抓取？让我们从一个你能直接在你的电脑上重建的小例子开始，假设你已经安装好 R。此案例会让你对本书的核心主题有个初步的印象。

1.1　案例研究：濒危世界遗产地

联合国教科文组织（UNESCO）是联合国的一个机构，其职责包括对全世界自然和文化遗产的保护。迄今为止（截至 2013 年 11 月），已经有 981 个地点被列入世界遗产，其中大部分是像胡夫金字塔这样的人造建筑，但也包括像大堡礁这样的自然景观。遗憾的是，一些被划定为世界遗产的地点正在遭受人类活动的威胁。有哪些地点正在遭受威胁？它们分别在哪些地方？世界上是不是有一些地区的遗产比其他地区的遗产处于更加濒危的状态？遗产地面临濒危风险的原因有哪些？这些都是我们在第一个案例研究中想要检验的问题。

当科学家要掌握某个主题的基本情况时，他们一般首先会做哪件事呢？他们会去维基百科（Wikipedia）查阅这个主题！打开世界遗产地的页面（http://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger），我们会看到一份清单，其中列出了现在和以前濒危的遗产地。这份列表包含了名称、位置（所在城市、国家及其地理坐标）、遗产地面临威胁的种类、该地点被列入世界遗产的年份以及该地点被列入濒危世界遗产的年份。我们首先调查一下这些地点在全世界的分布情况。

虽然这张列表包含了有关遗产地的信息，但它们所在的位置或区域性聚集的情况并不是特别直观。相比用肉眼扫描列表，更有效的方法是在地图上标出每个遗产地的位置。因为人类善于处理视觉信息，所以我们在本书中会尽可能地让分析结果可视化。不过，如何把来自列表的信息对应到地图上呢？这听起来像是个有难度的任务，但实际上并非如此，在后面的内容中，我们会充分讨论一些相关的技术。现在，我们先让你有一个如何用 R 处理这类任务的初步印象。本书在后面的章节里会更系统化地详细讲解下面的代码片段中的命令。

一开始，我们需要加载一批组件。虽然 R 只有一组基本函数——主要是与数学和统计学相关的，但是通过用户编写的组件可以轻松地对它进行扩展。对本例而言，我们要用 library() 函数[⊖]加载下列组件：

⊖ 此处假定所有相关组件已经安装好。否则，请在你的控制台输入下面的命令：install.packages(c("stringr", "XML", "maps"))。