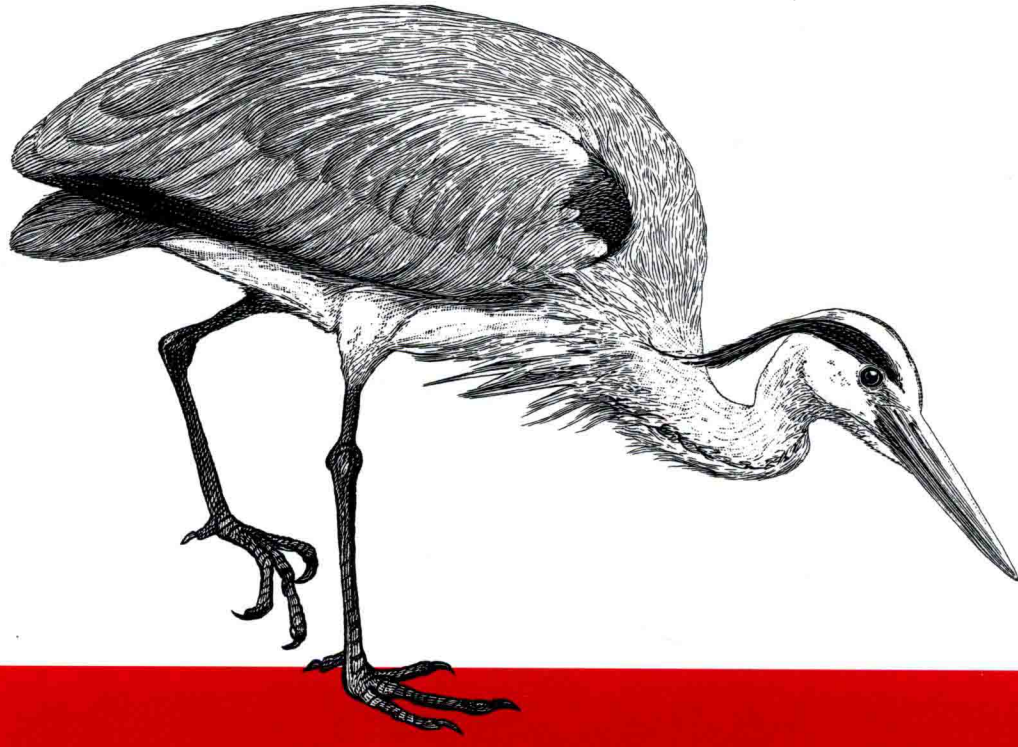
O'REILLY®



高效R语言编程

Efficient R Programming

Colin Gillespie Robin Lovelace 著
张燕妮 译

中国电力出版社

高效R语言编程

Colin Gillespie 和 Robin Lovelace 著

张燕妮 译

topol • Tokyo O'REILLY®

Media, Inc. 授权中国电力出版社出版

中国电力出版社

图书在版编目（CIP）数据

高效R语言编程 / (美) /科林·吉尔斯比（Colin Gillespie），(美) 罗宾·洛夫莱斯（Robin Lovelece）著；张燕妮译. — 北京：中国电力出版社，2018.8

书名原文：Efficient R Programming

ISBN 978-7-5198-2085-5

I. ①高… II. ①科… ②罗… ③张… III. ①程序语言－程序设计 IV. ①TP312

中国版本图书馆CIP数据核字(2018)第108264号

北京市版权局著作权合同登记 图字：01-2018-1599号

出版发行：中国电力出版社

地　址：北京市东城区北京站西街19号（邮政编码100005）

网　址：http://www.cepp.sgcc.com.cn

责任编辑：刘 炽（liuchi1030@163.com）

责任校对：王小鹏

装帧设计：Randy Comer，张 健

责任印制：杨晓东

印　刷：三河市航远印刷有限公司

版　次：2018年8月第一版

印　次：2018年8月北京第一次印刷

开　本：750毫米×980毫米　16开本

印　张：14

字　数：262千字

印　数：0001—3000册

定　价：48.00元

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

R语言是属于GNU系统的一个自由、免费、开放源代码的软件，它是一个用于统计计算和统计制图的优秀工具。2011年前，R在国内的发展一直不温不火。2011年，作为数据处理方面的优秀工具R在国内迅速流行了起来。

R语言的语法通俗易懂，初学者很容易上手。对没有编程基础，又想处理数据的其他领域专家来说R语言是一个很好的选择。国内关于R语言方面的书籍有很多，尤其是关于可视化、数据科学和包开发方面的R语言资源更是琳琅满目，还有很多R语言社区如火如荼。虽然R语言容易，资源又易得，但很多R语言初学者或程序员会掉入“低效的陷阱”，写出的代码虽然能运行，但效率却很低，幸运的是现在我们有了这本《高效R语言编程》。

本书的两位作者多年来为各种层次的学员讲授R语言，并在实际项目中使用R语言，积累了丰富的经验。书中从多个方面介绍了如何提高R语言的效率，从基础的配置到C++接口，每一章中作者都列出了本章内容所需的软件配置（众所周知，R语言包的依赖性很强），几乎每一部分内容都跟有实例演示及配套练习，以便读者切实的掌握这部分内容。

本书在翻译过程中充分发挥了个人专长，是团队协作的结果。刘芳擅长硬件配置，第8章由刘芳翻译；丁维才做过多个大型项目，对团队协作方面深有感悟，所以第9章由丁维才翻译；其余章节由张燕妮翻译。感谢丁维才使用SVN为全书翻译做了版本管理。

张燕妮

目录

前言

本书可使你的R编程工作事半功倍，它是关于计算效率和编程效率的。现在有着大量优秀R语言资源，例如可视化（如Chang 2012）、数据科学（如Grolemund与Wichham 2016）以及包开发（如Wickham 2015）。另外还有大量的关于R语言在特定领域的使用资源，包括贝叶斯统计、机器学习和地理信息系统。然而，关于如何轻松高效地应用R语言的资料非常稀缺。有关技巧、注意事项和数十年的该主题知识沉淀在大量网页、Email讨论和论坛中广泛流传着，但这种情况使R用户更难理解如何编写高效R代码。

在教学过程中，我们发现该状况同时存在于初学者和老手中。不管是理解避免循环应用R的向量对象和如何设置你的*.Rprofile*和*.Renviron*文件的问题，还是驾驭R的优秀接口提升性能，效率的概念是关键的。本书的目标是从技巧、警告、编程诀窍中为R程序员提取出长久有效的单一、内聚的精华。

本书内容来自于我们的学生的问题，各种培训中不同水平的、各行各业的学生数年来不断咨询如何使他们的R更快。怎样将计算机科学的通用原则（例如不要编写同一代码，即DRY）应用到R语言的具体代码中。怎样使R代码融入高效流程中，包括项目启动、协作开发及简评阶段？如何才能快速学习使用新的包或者新函数？

本书包含的10章内容解答了不止上述问题。每章从基本原理开始介绍，而后逐步推进到高级内容，这样可适合不同水平的读者。然而更高级的话题（如并行编程及C++）可能暂时对R新手无任何意义，本书是采用慢开头进而打下扎实基础的讲述方式帮助读者走完R语言的糟糕的陡峭学习曲线。这样即

使资深 R 用户也能从中找到先前隐藏的高明建议。在讲述这类资料时，我们经常听到“为何以前没有人告诉我可以这样做呢？”

高效编程不应看成可有可无的附属品，随着项目和数据集的增大，高效编程的重要性也在增加。事实上，本书是在我教基于 R 的大数据课程过程中构思出来的，而该课程是你要处理大量数据集以及保证代码高效的必修课。即使面对小量数据情况，快速编写运行速度快的高效代码是成功代码的必备特征。我们发现，高效编程的概念在所有 R 社区的分区中都是非常重要的。无论你偶尔使用 R（例如，因为它的无与伦比的统计包），还是为了开发包或者正在开发大型的合作项目（高效是至关重要的项目），代码高效均严重影响到你的工作效率。

最后，高效是一个事半功倍的话题。拿汽车作类比，你是想一满箱油（或者充满电）开车驾驶 1000km，还是每 50km 就更换一辆老破旧的汽车呢？或者你愿意选择一辆高效的汽车还是一辆自行车呢？同理，高效的 R 代码在各方面均超越低效 R 代码：读、写、执行、分享及维护都更简单轻松。本书不可能给出如何写出这样代码的所有答案，但它肯定提供了一些思想、实例代码和技巧来帮助你的编程之旅有一个良好开端。

排版约定

本书使用下述排版约定。

斜体（*Italic*）

表示新术语、URL、电子邮件地址、文件名和文件扩展名。

黑体（Bold）

表示 R 包名。

等宽字体（`Constant width`）

表示程序清单，以及在段落中应用的代码片段，例如变量、函数名、数据库、数据类型、环境变量、语句和关键字。

等宽粗体（**`Constant width bold`**）

表示命令或需用户输入的文本内容。

等宽斜体（*Constant width italic*）

表示需要用户提供的值或由上下文决定的值来替代的文本内容。

这个图标表示提示或建议。

这个图标表示一般性说明。

这个图标表示警告或提醒。

使用示例代码

补充材料（代码示例、图片等）可从以下网址下载：*https://github.com/csgillespie/efficient*。

本书的目的是帮助你完成工作。通常，你可以在你的程序和文档中使用本书的代码。除非你要复制大量代码，否则你无需为许可联系我们。例如，使用本书中的多个代码片段编写程序就无需获得许可。但以 CD-ROM 形式分发或销售 O'Reilly 书中的代码就需要我们的许可。回答问题时引用本书以及代码例程无需许可。在你自己的项目文档使用了大量的示例代码则需要获得许可。

我们不强制要求署名，但如果这样做，我们深表感谢。署名格式一般包括书名、作者、出版社和 ISBN 号，例如："Efficient R Programming by Colin Gillespie and Robin Lovelace （O'Reilly）. Copyright 2017 Colin Gillespie, Robin Lovelace, 978-1-491-95078-4"。

如果你觉得你的代码示例超出正常使用范围或者上述许可范围，请通过 *permissions@oreilly.com* 联系我们。

O'Reilly Safari

Safari（以前的 Safari Books Online）是一个针对企业、政府、教育工作者以及个人的会员制培训和参考平台。

订阅者可以访问来自超过 250 个出版社的大量书籍、培训视频、学习路径、交互式教程，以及精选的播放列表，这些出版社包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、AdobePress、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett 等。

更多信息请访问 *http://orieilly.com/safari*。

如何联系我们

有关本书的建议和疑问请联系出版社：

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

这本书有专属网页，你可以在那里找到本书的勘误、示例和其他信息，请访问：*http://bit.ly/efficient-r-programming*。

对本书做出评论或者咨询技术问题，发送Email至*bookquestions@oreilly.com*。

需要了解更多关于我们书籍、课程、大会和新闻信息，请访问我们的官方网站：*http://www.oreilly.com*。

在 Facebook 上查看我们：*http://facebook.com/oreilly*

在 Twitter 上关注我们：*http://twitter.com/oreillymedia*

在 YouTube 上查看我们：*http://www.youtube.com/oreillymedia*

致谢

我在开放平台上编写了这本书，很多人帮我修正了多个小问题。特别感谢 O'Reilly 的本书工作人员，以及每个通过 GitHub 的帮助者：@Delvis、@richelbilderbeek、@adamryczkowski、@CSJCampbell、@tktan、@nachti，Conor Lawless、@timcdlucas，Dirk Eddelbuettel、@wolfgangleder、@HenrikBengtsson、@giocomai 和 @daattali。

感谢提出详细意见的技术审核人员：Richard Conton 和 Garrett Grolemund。

Colin

感谢 Esther、Nathan 和 Niamh，谢谢你们的耐心。

Robin

感谢我在 Cornerstone Housing Cooperative 的室友容忍我写书时的邋遢，感谢 Leeds 大学中支持我从事非常规的期刊文章和学术会议的学术工作的每个人。感谢由开源代码的开发者、用户、传播者构成的社区的每个人，他们使这一切成为可能。

第 1 章

概述

本章从 R 语言和编程经验的角度，介绍了本书所适用的广大群体，以及如何更好地利用本书。任何想提高效率的读者首先应准确地理解效率这个术语，本章在“什么是效率？”一节讨论了算法和程序设计者的效率，并在“什么是有效 R 编程？”一节专门讨论了 R 语言的效率问题。本章“为何需要效率？”一节谈论了为什么现在功能强大的计算机很便宜，却还要困扰于代码的效率。

幸运的是本书不止只适用于 R。本章“通用效率技巧”一节概述了高效的 R 编程所需要的非 R 编程技巧。不同于典型的编程书，该节讲述了盲打和一致性这些编程以外的、通用的改善效率的技巧。然而，本书首先是一本编程书，因此每一章都配以代码示例。尽管开篇有比较多的概念以及零零碎碎的知识，本章“基准测试与性能测试”一节介绍了高效 R 程序员工具箱中的两个非常重要的工具，并通过两个演示例程讲解如何使用。本章的“本书资源”是本书所用到的相关包及源代码。

软件要求

阅读本书时，边读边运行和体验代码是非常有必要的。在后续章节的开头都有一节介绍当前章所需要的软件，请确保你安装每一章所需的软件包。本章所需的软件是：

- 在你的电脑上安装 R 软件（见“安装与升级 RStudio”）。
- 安装和加载 **microbenchmark**、**profvis** 与 **ggplot2** 包（“安装 R 包”，第 2 章介绍了如何安装和更新包）。可通过下面命令检查你是否安装了这些

包：

```
library("microbenchmark")
library("profvis")
library("ggplot2")
```

运行整本书所需的软件配置，在本章“本书资源”一节中列出。

读者对象和如何使用本书

任何想快速编写 R 代码、使 R 代码运行更快且更具扩展性的人都应该阅读本书。通常在学习了 R 语言中关于数据分析的基础知识之后会有这些需求。我们假设你熟悉 R 语言或精通其他语言编程，同时本书也适用于初学者。根据技术水平可将本书广大的读者群分为如下三类：

刚接触 R 语言的编程人员

本书将帮助你遍历 R 语言的特性以便高效工作，如果你将 R 如同其他语言一样使用的话，则难以编写高效的代码。

少许编程经验的 R 用户

本书为你提供了很多概念与编程技巧（某些属于计算机科学），它们可提高你的效率。

少许编程经验的 R 初学者

本书将引领你有一个良好开端。毕竟坏习惯容易养成且很难改掉。在你的编程生涯开始时阅读本书将为你节约大量的网上搜索时间。

了解自己的情况有助于你从书中最大限度地获益。当你有个正在做的项目时推荐你读一下本书，无论你的项目是工作上的相互协作项目，还是在家中就能完成的简单的个人项目。为何这么说呢？因为本书所涉及的内容比大多数编程教程更广泛（例如第 4 章谈及了项目管理），做项目使你能够将概念、建议、代码付诸实际工作中。从书本直接到工作的方式确保了实战中能强化所学内容。

如果你是上述分类的最后一组：R 的初学者，我们推荐你所用的练手 R 项目不是需要交付给客户的重要项目，而是别的一些 R 资源项目。尽管本书属于通用类型，但你的 R 应用可能具有相应专业特点。基于此，我们建议你阅读