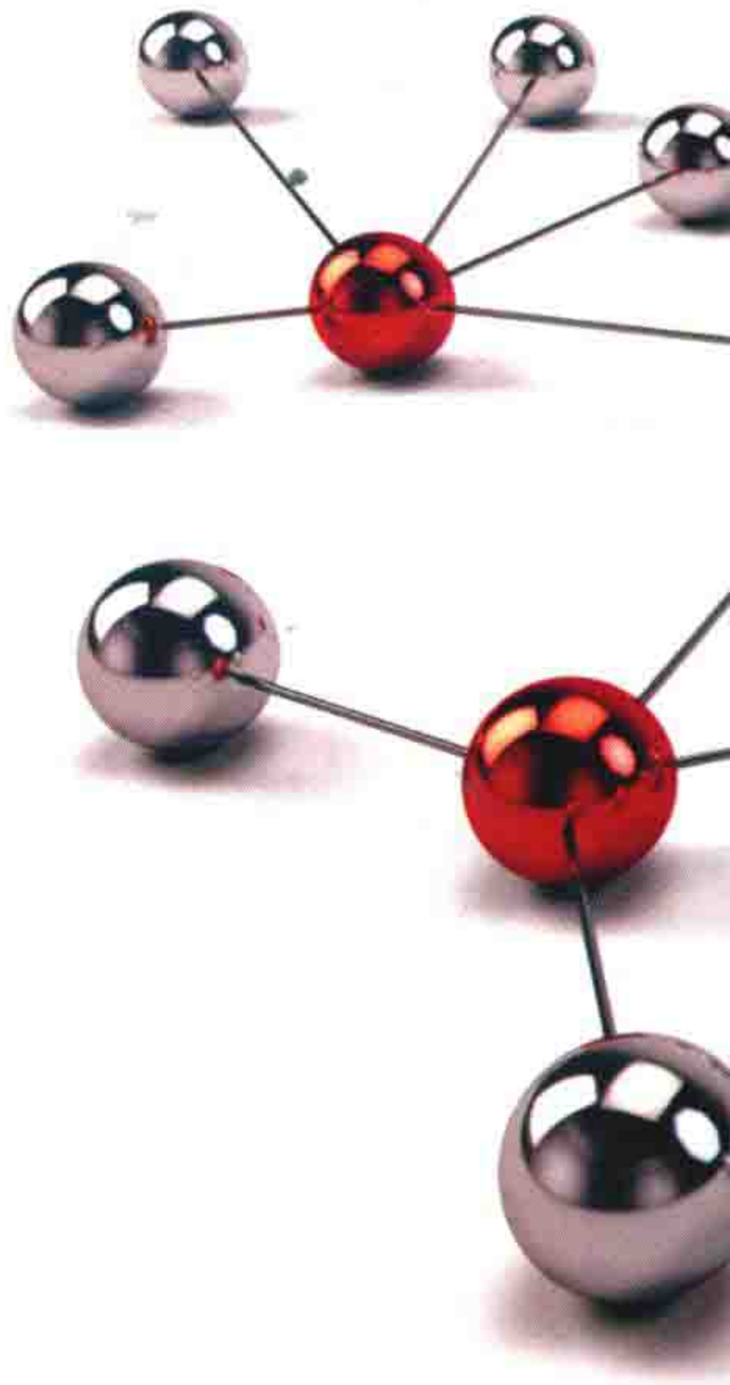
TURING
图灵程序
设计丛书

书中数据和R代码可下载

R数据挖掘入门



[日] 山本义郎 藤野友和 久保田贵文 / 著
朱建春 / 译

R × Data Mining

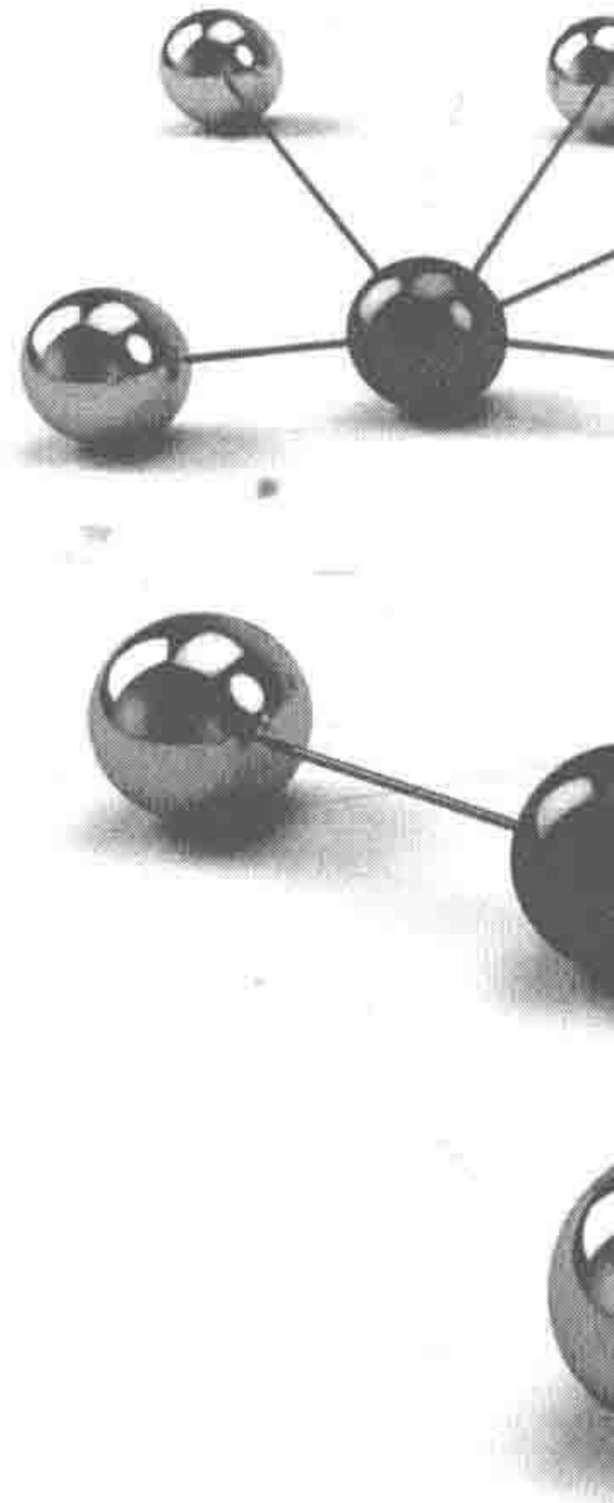
从理论基础到实例应用，边学习边实践！

网罗数据挖掘中10种经典、常用的方法

回归分析/Logistic回归分析/决策树分析/支持向量机/记忆基础推理/聚类分析
/神经网络/主成分分析/对应分析/关联规则分析

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING
图灵程序
设计丛书



R数据挖掘入门

R × Data Mining

[日] 山本义郎 藤野友和 久保田贵文 / 著
朱建春 / 译

人民邮电出版社
北 京

图书在版编目(CIP)数据

R数据挖掘入门/(日)山本义郎,(日)藤野友和,(日)久保田贵文著;朱建春译. -- 北京:人民邮电出版社,2018.3

(图灵程序设计丛书)

ISBN 978-7-115-47878-8

Ⅰ.①R… Ⅱ.①山…②藤…③久…④朱… Ⅲ.①数据采集—研究 Ⅳ.①TP274

中国版本图书馆CIP数据核字(2018)第024767号

内 容 提 要

本书使用R,结合大量实例,详细介绍了数据挖掘的理论和分析方法。全书分为3部分:第1部分简单介绍了使用R进行数据挖掘的流程和数据挖掘的概要;第2部分介绍了数据挖掘的10种常用方法,并在此基础上使用R实际进行数据挖掘;第3部分结合实际的数据挖掘事例介绍了如何使用这些方法。本书适合数据挖掘的初学者,以及正在从事数据分析相关工作,想了解更多分析方法的读者阅读。

◆ 著　[日]山本义郎　藤野友和　久保田贵文
译　朱建春
责任编辑　杜晓静
执行编辑　刘香娣
责任印制　周昇亮

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷

◆ 开本:880×1230　1/32
印张:6.5
字数:194千字　2018年3月第1版
印数:1-3 000册　2018年3月北京第1次印刷

著作权合同登记号　图字:01-2017-4817号

定价:45.00元

读者服务热线:(010)51095186转600　印装质量热线:(010)81055316
反盗版热线:(010)81055315
广告经营许可证:京东工商广登字20170147号

译者序

2016年美国总统大选前，虽然希拉里的民调支持度始终高于特朗普，但印度一家公司推出的名为MogIA的人工智能系统却预测特朗普将赢得大选，令所有专家都跌破眼镜。MogIA是通过对Google、Facebook、YouTube等网站上收集的超过2000万条资讯进行分析而得出的结论，事实证明其预测完全正确。该公司称，至今MogIA已经预测过多次国际选举，最终结果都符合预期。

世界已经进入了“云、物、大、智”的时代，即云计算、物联网、大数据、人工智能。在这个时代，你能做点什么呢？

前几天看到腾讯新闻，说河北柿子滞销，山区2亿斤柿子卖不掉，只能当饲料。如果果农有充分的市场信息，或者相关机构能帮助他们做一下市场数据分析，情况是否能好得多？我想是的。虽然仍然存在竞争，但是不会输得这么惨。

现在，技术的进步已经使得大数据的处理成为可能。提起大数据，人们很快就联想到Hadoop、Spark和Storm。这些框架把若干廉价的计算设备连接起来，成为具有超级计算和存储能力的分布式系统，令人惊叹。但是它们只是提供了大数据存储和分析的平台，并不提供分析方法。真正的利器则是数据分析和数据挖掘。

本书的主题是数据挖掘。数据挖掘领域有很多著名的算法，如关联规则发现的Apriori算法等。不过，对于这些算法的细节，本书不做介绍。本书要做的是，把那些想了解和使用数据挖掘技术的人，带入这一领域。通过使用已经开发好的工具，让读者切身体会从数据中找宝的过程，去了解数据挖掘的各种方法的差异。

译者从事IT行业已经接近30年了。其中的大部分时间都在做普通

的企业应用的软件开发，和“云、物、大、智”的关系不大。但是，“普通”的工作做久了，自然就想去做一些“高级”的，但后来译者惊讶地发现，大数据和人工智能也并不高级，也只是一些IT的工具而已。数据挖掘也是这样。关键的是，你要迈出第一步，才能利用这些工具为你的组织创造价值。而迈出这一步其实并不是很难。

本书使用R，结合大量的实例，带领读者走进数据挖掘的世界。R是一种简洁、高效且易扩展的语言。它的好处在于，有无数的专业人士在不停地为其添砖加瓦，开发扩展程序包。而数据挖掘的技术也是日新月异的，比如，对于本书没有涉及的图挖掘、语音挖掘，译者翻译本书时，尚没有合适的R程序包。但是，当你购买本书时，或许这些新方法的程序包就已经面世，你可以免费使用。因此，R不仅是一门语言，而且是一个体系，一个绝不应该忽略的体系。因为它无偿为我们提供了很多优质的功能。

本书中所运行的R代码和数据，全部可以从网上下载，你只需按照书中说明进行操作，逐步执行，就可以看到多彩的曲线和图形，这会带给你愉悦，让你发现数据挖掘不再是抽象的理论，而是一种鲜活的技术。当你看到寥寥几行R脚本，就能使你从庞大的数据中找到端倪时，你就会惊讶于数据挖掘的魅力。

本书原版由日本著名的出版公司欧姆社出版。和欧美图书的精准详实相比，日文图书的特点是娓娓道来，深浅结合。这次翻译本书对我来说也是一次有益的尝试。由于原书最后一章的文本挖掘是针对日文的，不适合中国读者，因此在征得原作者同意的基础上，译者进行了重写。对于原书的数据，有些包含日文文字，译者也把它译成了中文，以便国内读者使用。

由于时间仓促和译者水平有限，翻译过程中难免有疏漏，请读者不吝指正。

朱建春

2017年12月

序言

近年来，随着大数据逐渐受到关注，数据科学的重要性被认识，实际承担数据分析任务的数据挖掘再次引起了人们的注意。在十多年前的数据挖掘大潮中，人们对数据挖掘充满了“有数据就能挖到金矿”的期待，然而通过数据挖掘“挖金矿”并非易事，所谓对数据使用某种工具进行处理后就能自动找到宝贵的信息，不过是一种幻想而已，当人们明白这一点后，大潮就渐渐退去了。近年来人们对数据科学家的关注中所透露出的对数据科学的期待，和数据挖掘大潮时期开始有所不同，纵然数据分析需要一定程度的分析技能和统计学的思考方式，但人们开始努力从数据中找出一些有用的信息，使其在实际业务中发挥作用。而本书就是为希望了解数据挖掘的相关概念，并运用到实际的数据上，以及正在从事数据分析相关工作，想了解更多分析方法的读者而写的。

数据挖掘不是老旧过时的方法，数据科学之所以受到关注，就是因为数据挖掘解决了各种业务问题。数据挖掘的很多方法和之前没有太大变化。

本书使用真实的数据，希望通过实际的操作，帮助读者理解数据挖掘，为此，本书分成三部分。

第I部分为“使用R进行数据挖掘的准备”，第1章“基于R的数据分析入门”简单介绍使用R进行数据挖掘的流程，第2章“数据挖掘概述”介绍数据挖掘的概要。

第II部分为“数据挖掘方法”，对10种数据挖掘方法分别介绍其概要，并使用R实际进行数据分析。首先是关于预测的方法，第3章介绍回归分析，第4章介绍Logistic回归分析，第5章介绍决策树分析，第6章介绍支持向量机，第7章介绍记忆基础推理。然后是关于分组和分

割的方法，第 8 章介绍聚类，第 9 章介绍自组织映射。接下来是关于在数据挖掘中通过低维度可视化来把握大量数据的维规约的方法，第 10 章介绍主成分分析，第 11 章介绍对应分析。最后，作为购物篮分析，第 12 章介绍数据挖掘中著名的关联规则分析的方法。

第 III 部分为“数据挖掘实战”，第 13 章通过具体的例子对多种预测方法进行评估比较。此外，作为两个实际的大规模数据分析的例子，第 14 章介绍从股价生成综合指数，第 15 章介绍社交网络分析。

为了让大家实际操作 R 以加深理解，本书除了第 2 章以外都提供了 R 代码。

通过 RStudio 实际操作演练，是理解本书内容的捷径。

本书中使用的数据和 R 脚本，可在以下网址下载。

http://www.ituring.com.cn/book/1914

R 脚本的编码方式（Encoding）为 UTF-8。在 RStudio 中打开 R 脚本时如果出现乱码，请在“File”菜单的“Reopen with Encoding”中指定“UTF-8”，然后重新打开。

本书中的部分代码依赖启动时的随机数，导致运行结果不尽相同。因此，实际运行源代码时，可能会出现和本书不同的结果。

由于像这样运行时间不同导致结果不同的情况很多，所以不可进行一次分析后就做出判断，而应进行多次分析，弄清楚所得到的结果是稳定的结果还是偶然的结果。

从本书的策划到出版，欧姆社的各位负责人给予了很大的支持和帮助。在本书出版之际，谨对他们的努力和热情，深表谢意。

山本义郎 藤野友和 久保田贵文

2015 年 10 月

目　录

第 I 部分

使用R进行数据挖掘的准备

第 1 章

基于 R 的数据分析入门

统计分析软件 R 的知名度不断上升，应用也日渐广泛，基本上一提到统计分析软件人们就会想到 R。R 获得广泛支持的原因可以列举如下几点：

- 作为开源软件，可以免费使用
- 很多统计专家对开发提供了支持
- 由于是脚本语言，因此可以很容易地追踪统计分析的过程，并重复运行
- 有众多的程序包可以使用[①]

R 中提供的程序包不仅支持最新的统计分析方法，还支持对外部资源的访问和网站开发，一般编程语言所能实现的功能，R 也都能做到。也就是说，只使用 R 也可以实现数据相关的环境的搭建。

本章将使用 R 搭建数据挖掘的执行环境，并简单介绍从 R 的安装、R 语言的基础到数据可视化的相关内容。

---

① 根据执笔时（2015 年 9 月 3 日）的统计，官网上共有 7096 个上传的程序包。

1.1 R及RStudio的安装

(1) R的安装

R的官方网站为http://www.r-project.org/，R和R程序包从名为CRAN（The Comprehensive R Archive Network）的仓库[①]进行下载。在国内也有几个镜像网站[②]，大家使用离自己最近的即可。访问CRAN后，页面的上方有针对各种操作系统的下载链接，请下载并安装适合自己的系统的文件（图1.1）。另外，在本书执笔时，R的最新版本为3.2.2[③]，使用Windows系统的情况下，依次点击“Download R For Windows”→“base”→“Download R 3.2.2 for Windows”，即可下载R的安装程序。

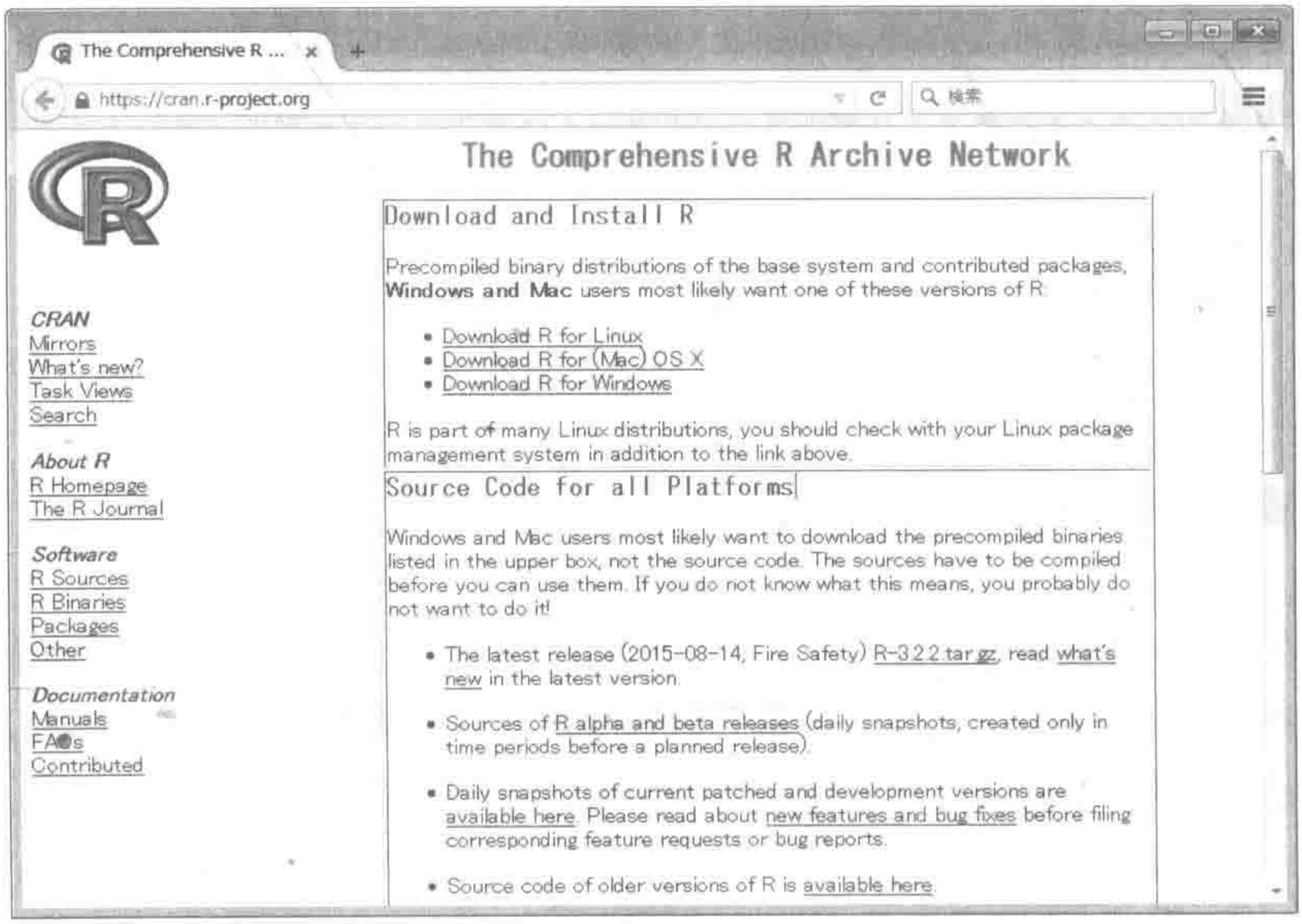


图1.1 CRAN的网页

① http://cran.r-project.org/

② 比如清华大学的http://mirrors.tuna.tsinghua.edu.cn/CRAN/，或中科大的http://mirrors.ustc.edu.cn/CRAN/等。——译者注

③ 读者下载最新版本即可。——编者注

双击下载的安装文件，启动安装器（图 1.2）。如果按默认设置进行安装，可一直点击“下一步”，直到完成安装。然后，双击桌面上生成的快捷方式，启动 R。但是，由于 R 自身的用户界面只提供最基本的功能（控制台和图像输出），因此不便于编辑和执行 R 脚本。这里，我们导入 R 的集成开发环境 RStudio，以提高工作效率。

图 1.2 R 的安装画面

（2）RStudio 的安装

RStudio 是对基于 R 的数据分析和开发提供支持的集成开发环境，它提供了如下功能。

- 脚本编辑器

语法高亮显示 / 代码补全 / 自动缩进 / 运行选择的代码

- 图像管理

各种格式的输出 / 图像的历史记录及重复显示

- DEBUG支持
- 项目、工作区及目录的管理

RStudio作为开源软件，可以免费使用。下面将从RStudio的官方网站http://www.rstudio.com/进行下载并安装。RStudio不仅有桌面版，还有服务器版、商用版（Commercial License）等版本，请选择自己需要的版本。这里安装的是桌面版的Open Source License。

顺着RStudio主页的引导进入下载页，或者直接进入下载页[①]，下载适合自己的操作系统的安装文件并执行（图1.3）。点击两次“下一步”，再点击“安装”，其后的安装过程将自动进行。安装完成后，“开始”菜单中将有对应的图标，可以从这里启动RStudio。

图1.3 RStudio的安装画面

1.2 RStudio的基本操作

（1）脚本编辑器

RStudio启动后，系统会显示R的控制台（Console）、根文件夹

① http://www.rstudio.com/products/rstudio/download