HZ BOOKS
华章科技

资深R语言用户多年实战经验的结晶，介绍了各种性能奇特的R语言包，提升R语言性能的方法，以及R语言在实际使用时与Java、MySQL、MongoDB、Hive、HBase、Hadoop等技术的综合运用的解决方案。

数据分析
技术丛书

# R的极客理想

## 工具篇

张丹/著


机械工业出版社
China Machine Press

数据分析
技术丛书

# R的极客理想

## 工具篇

张丹/著


机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 的极客理想——工具篇 / 张丹著 . —北京：机械工业出版社，2014.8
（数据分析技术丛书）

ISBN 978-7-111-47507-1

I. R… II. 张… III. 程序语言－程序设计 IV. TP312

中国版本图书馆 CIP 数据核字（2014）第 170133 号

本书首先介绍了 R 的工具包、时间序列包和性能监控包；然后阐述 R 语言与其他编程语言的通信，以及 R 语言作为服务器的应用；最后阐释 R 语言与各种数据库的通信以及 R 语言如何与 Hadoop 集成。附录介绍了 Java、各种数据库以及 Hadoop 的安装方式。书中内容涉及计算机、互联网、数据库、大数据、统计、金融等领域，详细总结了 R 语言在实际使用时与 Java、MySQL、Redis、MongoDB、Cassandra、Hadoop、Hive、HBase 等技术综合运用的解决方案，具有实战性及可操作性强等特点。

本书适合所有 R 语言工作者，包括软件工程师、DBA、数据科学家、科研工作者以及相关专业的学生。读者可以选择任何感兴趣的章节进行阅读，每节之间没有特别的顺序要求。

# R 的极客理想——工具篇

张丹 著

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：明永玲 责任校对：董纪丽
印 刷：蓟城市京瑞印刷有限公司 版 次：2014 年 8 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：19.5
书 号：ISBN 978-7-111-47507-1 定 价：59.00 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



# Foreword 序　一

生活中我是三分熟的牛排，张丹兄是五分熟的牛排，一天他给我发邮件邀请我写序，我好些天都没答复，为什么呢？因为我们互相都不熟啊！嘿嘿，好冷。从没有人找我写过序，张丹有胆找一个浑身负能量爆棚的新手及话痨写，还说自由发挥，风格不限，作者自己都不怕，我怕啥，大不了把脖子往你们面前一横！当然那是张丹的脖子……其实我自己看书一般不太看推荐序（除非里面有重大八卦），因为推荐序里通常也就是友情帮点一百二十八个赞，你要是像鲁迅先生那样半夜爬起来翻看推荐序，当然你肯定不会看见“吃人”两个字，而是看见一个特殊的变量值：NULL。哎呀，我这只是夸张的说法啦，推荐序没那么糟糕，我也并非和张丹完全不熟，我看过他写的很多博客文章。好了，言归正传。在下谢益辉，身患统计学永久性脑损伤，目前在 RStudio 当码农，天天蹲 A 村村口敲代码。十年前在中国人民大学统计学院上大三（哥也年轻过），侥幸没被统计计算课上的 R 语言折磨死，按照我们疯人院（没有含沙射影的意思，请读者切勿自动匹配）的规矩：你上回没弄死我，这次你在前面跑，换我来弄死你好了。就这样 R 语言成了我快乐生活的一部分。咦，为什么有一种细思恐极的感觉？

作为一名不那么极的 R 极客，我自然乐意看到一本写给极客的书。在这个数据时代（千万别再跟我提“大”数据，否则我立刻变身纯生牛排给你看），各种新技术风起云涌，需要有人坐一段时间冷板凳，为我们收拾整理这些技术，让那些从证明三种中心极限定理的苦海中逃出来的研究生们毕了业不要立马又陷入另一片五种数据库引擎的苦海。我花了一晚上加一白天的时间才看完《 R 的极客理想》的书稿，看不懂和看得懂的地方都跳着看，看不懂主要是因为我没有计算机背景（我学习的第一门计算机语言是 VB，你们笑我可以，别笑出声就好，不过我还写过 VBScript 呢，这次你们可以笑出声了），例如 Java，我多年前曾经自学过一阵，现在都忘光了，写个“你好世界”的程序都需要搜一下；看得懂的地方主要是纯 R 的内容，不过有些地方还是慢慢看了，比如 formatR 包那里，主要是看张丹有没有黑我。算他识相没黑我，那好，我在序言里可以放心……黑他了；江湖险恶，你不中

箭谁中箭？

整本书涵盖的内容比较广，每一节的篇幅相对较短，我觉得这种写作风格挺好，每天晚上睡前打开一节看看，在口水浸湿书之前应该可以看完一节，既能学到知识又合理利用了睡前时间。如作者自己所说，这书不是写给初学者的，所以看第 1 章的不要被“R 基础”的标题给骗了，全是些奇门遁甲之术，一个基础包都没介绍！虽然我不敢讲 R 是不是最值得学习的程序语言，但我对 1.1 节的观点深表同意：R 往往用两三行代码解决问题，不会时时考虑最优与否（很多需要优化的地方都已经用 C/Fortan 等底层语言打包好了）。R“它爹”S 语言的主要作者 John Chambers 在几十年前就说了：S 语言的目的是让我们（统计学家）快速而可靠地把脑子里的想法变成软件。将数据拿到手之后甩开膀子从各个角度去分析就好了，想画图画图，想跑模型跑模型，而不必先考虑定义一个结构体以及某个成员是整型还是浮点型。书里主要用到的操作系统是 Windows 和 Ubuntu，不过苹果（OS X）用户不必担心，R 在 OS X 上跑起来也是妥妥的，个人推荐用 Homebrew 安装 R。第 1 章介绍了一系列奇特 R 包，从中我可以看出作者钻研这些包的乐趣，极客需要这种小乐趣的推动，后面书中我们还会持续看到作者钻研的迹象，他胆敢把自己失败的经历都写出来，实在是很勇敢，但这也反映了极客的真实生活嘛，哪有干什么都一帆风顺的？奇特 R 包系列里包括了我的 formatR 包，这个包有几处小小的坑，提醒广大读者注意看文档，代码格式主要还是得靠程序员自觉，不能老依靠 formatR 这样的自动整理代码的包。fortunes 包里的确有大智慧（不是炒股软件），但我感觉主要还是各种恶搞和冷笑话，码农的心思你别猜，别猜别猜。RStudio 服务器版是个好东西，未来的趋势可能是啥都存储在服务器上，浏览器控制一切，再也没有黑乎乎的 SSH 窗口。不过前两天我一个同学给我打电话哭诉我坑了他，因为 RStudio 的服务器版太好用了，他把代码都放服务器上，每天就在浏览器里写代码，不再从本地传来传去，结果服务器硬盘坏了，没备份。我感到很不好意思，这周打羽毛球我都没敢再叫他。RJSONIO 的作者本来是我很敬佩的一位大极客，可能是 R 界兴趣最广泛的极客之一，可他的代码极少写单元测试。导致我被坑了几次之后越来越不敢相信他的编码质量，哎，上天给聪明极客的惩罚就是让他们失去单元测试。不过话说回来，什么 JavaScript/JSON 啊，HTML/CSS 啊，都是非常容易入门的技术，作为统计出身的我，大力推荐大家掌握一点这种低投入高产出而且又有趣的计算机知识。Cairo 是另一位大极客的作品，我在他手下呆过两个多月，其聪明程度以及掌握的计算机知识之广总令我惊叹，例如后面章节中介绍的 Rserve、FastRWeb 以及 RCassandra 都是他单枪匹马的作品，一个比一个极，问题就是极过了以至于很少有人知道，所以也没见广泛应用。有时候我都想，你懂得这么多东西你家里人知道吗？他家里人知不知道没关系，起码现在我们知道张丹帮了个大忙，让他的几份作品至少多了一些中国读者。有志于成为极客的码农们，在忍受孤独的同时，我觉得真的应该好好想想为什么你做的东西没有广为人知，就算你不想这个问题，也该想想怎么

到今天还没有女朋友吧，这两个问题背后肯定有共同原因。如张丹引用我的话所说，Cairo包的高质量在现今的R版本中已经不算太大优势了，因为R本身已经支持Cairo库，例如其PNG图片输出和Cairo包的输出质量上几乎是一样的，但有一个特例除外，就是散点图中点的形状为某一个类型的圆点时（看不懂这句话的人请查阅points函数的帮助），基础R画出来有锯齿，而Cairo包没有，至于pch取值多少时有锯齿，这问题就留给你们自己去探索吧。caTools这个神奇到莫名其妙的包，某种程度上反映了一些R包作者广泛的兴趣以及没有正规计算机背景的特点，这事情难说好坏。

也许是缘于对金融的兴趣，张丹在第2章介绍了时间序列数据的处理。金融是才疏学浅的我不太能理解的行业（注定了我穷困一生的命运），时间序列也是我比较弱的功课，除了R自带的ts()函数以及简单的ARMA模型，我脑子里剩下的时间序列的知识已经没多少了。zoo和xts是我听过无数遍但从来没使过的R包，从此也可以看出R的应用领域太广泛了，我用了10年R也没用过这两个流行的包，主要是因为我不太做时间序列相关的工作。看完这一章我觉得读者不妨也研究一下R的基础图形，主要是graphics包，掌握一点基本的画点画线技术，对图形的灵活应用会很有帮助，如今R的图形天下基本被ggplot2占据，但我还是老土的基础“图形党”（没办法，我学R的时候ggplot2还没出生），我觉得不是所有的数据都适合ggplot2的。这一章介绍了时间序列数据处理以及可视化，如果是做时间序列的预测，我听说过无数遍forcast包，但至今未得一试。

R的性能一直以来都是计算机专业人士对R的槽点，张丹和我都表示没有压力，但这不表示性能不是问题，当然谁都希望自己的程序跑得快，好省下几分钟时间去写另外300行C代码提升下一个程序的速度。第3章提供了提速以及找代码瓶颈的工具。对于memoise包本身我其实没太大兴趣，但它的源代码是值得一看的，R里面的函数（或称闭包，Closure）和环境是很有意思的话题，路远坑多，慢走不送。性能监控也是一个优化代码的重要手段，作者介绍了基础工具Rprof()以及酷炫工具lineprof包，让我们知道自己的代码的瓶颈所在。最后作者讲，R语言需要更多IT人的推动，我实在不能同意更多。R作为统计学家写给统计学家的语言，总是会有些坑，需要专业人士来帮忙填补；另一方面，这帮顽固的统计学家完全无视界面的重要性，看看R官方网站有多朴素就知道了，简直是土得连渣都掉不下来，你跑去跟他反映，他只是一道冰冷眼神就可以杀死你。话说做网页前端的IT人士你们在哪儿呢？

第二部分的两章介绍R的服务器应用，前面说了我是Java外行，所以不敢乱点评。只记得那时候Urbanek大人一路上手舞足蹈给我解释FastRWeb的原理，然而我回去看到/var/FastRWeb/这个目录的时候就已经暗自决定抛弃它了。我个人对那些需要sudo才能运行且存放在非标准位置的程序有抵触心理，因为我下次一定会忘记怎么运行它以及配置文件在什么位置。Rserve和FastRWeb在它们被发明的年代里绝对都是划时代的，在服务器

端跑基于 R 的服务是很多码农的梦想，还记得那些年我们一起追的 Rweb 吗？如今回头看看，还有多少人记得并用着 CGI？可喜的是，张丹在第 5 章也介绍了 WebSocket 技术，这也是一个相对较新、很有趣且有用的话题，建议读者好好研究。书中提到 websockets 包已经被移出 CRAN，什么原因我也不知道，不过我基本上确定 httpuv 包可以取代它，也许是 websockets 的作者看到 httpuv 的工作之后决定不要重复劳动了。httpuv 包是 shiny 包的核心技术之一，如今捣鼓 R 的服务器端应用，怎么能忽略 shiny？ shiny 比 Rserve/FastRWeb 出现晚了近十年，为什么前者迅速流行起来，而后者尽管带着划时代的思想和技术，却被广大用户忽略，极客们应该再次好好想想。有读者可能会说，切，shiny 是你们厂（RStudio）的产品，你当然自卖自夸啦！是不是这样呢，我们且留给时间检验。

第三部分的两章就是数据库八仙过海各显神通了，我仅仅粗浅了解一点 MySQL，请 Hadoop/NoSQL 同行们不要笑出声。作为入门教程，这些章节都不错，从安装到“你好世界”的例子都有介绍，十八般武艺入门之后的事情大家都知道，遇到问题搜索就是了。

极客不是一种身份，而是一种态度。在我眼里，这个词是中性的，极客不代表一个人有多牛，而是他的钻研态度、好奇心以及对新技术的识别和接受能力。有些很牛很聪明的人，未必能把聪明才智转化为生产力（请勿对号入座）。张丹这本书给大家提供了一条通向 R 的极客之路，但这绝对不是终点。技术人士容易沉迷于技术，就像科学研究人士迷信某一种科学一样，唉，我就是这样浑身负能量。希望读者通过这本书能感受到作者探索的乐趣，保持开放心态，积极学习，然后找到适合自己的极客理想（以及女朋友！相信我，后者会让前者更快实现）。写序似乎应该说点鼓励的话吧，我没写过也不清楚贵圈的规矩，那么就引用麦太的话好了：从前有一位小朋友他很努力学习，后来他发财了。

谢益辉

2014 年 6 月 23 日于 A 村

（吝啬的房东一直不给我修空调，已热哭，决定在最后这个黄金广告位狠狠黑她一把，叫她随便得罪码农！）

# Foreword 序　二

我有时会问自己会不会因为名字而买一本书？当我看到“R的极客理想”时，我就找到了答案。把这个书名做个分词，去除停止词之后有三个词：“R”、“极客”和“理想”。这是耐人寻味的三个词。

R是我现在谋生的工具，我对它有着十多年的感情。我亲历了R从和GAUSS进行比较的时代走过和SAS进行比较的时代、来到了和Python及Julia这样的语言进行比较的时代，从最开始的无人问津到如今的炙手可热。R的资料在互联网上可以说汗牛充栋，但是中文书籍仍然很少。张丹是圈内著名的博主，明永玲是圈内著名的编辑，有幸被邀约写这篇序言时，我对这个组合非常看好。

阅读内容之后，我发现这本书有太多和其他R语言书籍不一样的地方。传统的R语言书籍大多是基于统计的思维展开的，通过介绍统计方法在R中的实现来学习R，这就使得很多统计出身的用户可以很容易地和其他统计软件进行类比，从而加速学习的过程。进入大数据的时代后，R作为数据处理的神器也越来越受关注，R语言的书籍也开始以数据为中心，从数据的获取、处理、分析一直到可重复研究和可视化展现，在应用层面进行全方位的介绍。但是，作为一种编程语言，程序员视角的R书籍还是非常少的。张丹的这本书刚好可以覆盖这部分的内容。

极客是R圈中比较少见的一种生物，尤其是来自极客之乡IT界的正宗极客。R的最初用户基本上都是统计圈的，但是最近几年R能够在国内越来越火，主要得益于IT界的贡献。R在欧美火的时间更早，但发展的趋势也大抵如此。从R的本性来说，它本不该是极客关注的语言，因为其对外部功能延伸的追求远甚于对内部语言完美的追求，从S语言设计之初就声明了人的时间远比计算机的时间宝贵，尤其是分析建模人员的时间。这种比较乡愿的风格最初是不为极客所喜的。但是因为其极端易用，在惜时如金的产业界快速地流行了起来，自然而然地产生了大量的难题。于是高贵冷艳、魅惑狂狷的极客们就参与进来了。我认识的张丹，就是这样一位极客。

R 是一个很奇怪的东西，没有编程基础的人可以很容易地入门，但是很难有信心觉得自己成了高手。编程高手初学时常常是破口大骂，但是很快就中了 R 的迷毒。说到底，还是因为 R 本质上是统计学家发明的语言，和模型打交道的能力比和计算机打交道的能力更强大。虽然如此，随着 R 的扩展包越来越丰富，用户在享受便利性的同时也增加了理解上的风险，函数背后的机制不再是黑箱，那么，R 中的 IT 高手的见解就变得非常重要。从他们的视角来看 R、使用 R，无论是对于统计背景还是 IT 背景的用户都有很强的借鉴意义。

R 流的是实用主义的血，看上去和理想是背道而驰的。但 R 毕竟是没有灵魂的工具，它的性格应该取决于用它的人。在张丹的这本书里，我看到了理想的光辉。我看过不少书也自己写书，感觉介绍知识是最简单的事，但是表明观点是最困难的。对于作者来说，观点越不鲜明就越能避免犯错。但是对于读者来说，尤其是初学的读者，很容易陷入对自己的怀疑中。而张丹的这本书会直接告诉读者应该怎样做，照着代码操作一遍就能解决问题。虽然有些建议可能不是最好的解决方案，但至少是足够好的，在实际的应用中可以解决问题。在这一点上，理想主义的作者和实用主义的 R 实现了完美的结合。

当然，整本书在“R”、“极客”、“理想”之间实现了更加完美的结合。

李舰

2014 年 8 月 1 日于上海

# Preface 前　言

## 为什么要写这本书

我是一名程序员，前后做了10年的程序开发工作。在这10年间，我从程序员一路做到架构师，经历了太多的系统和应用。我做过手机游戏，写过编程工具；做过大型Web应用系统，写过公司内部CRM；做过SOA的系统集成，写过基于Hadoop的大数据工具；做过外包，做过电商，做过团购，做过支付，做过SNS，也做过移动SNS。以前只用Java，然后开始用PHP……如同其他程序员一样，我一度陶醉于追求各种技术的创新，但始终有一个问题困扰着我，那就是如何才能够将我所掌握的技术转变成价值？这就好比我面对着一座金山，我拥有先进的技术，可以制作各种性能稳定、功能卓越的挖掘机器，但我不懂如何将矿石提纯，变成金子！每每看到别人利用我的技术挖掘出金子时，我只能满脸的羡慕，心中无限的不甘。

直到遇见R语言，我豁然开朗。R语言为我从另外一个角度开启了宝藏的大门，也让我对自己的职业重新思考、规划，最后坚定了我向统计、金融行业的转型。如果你也存在以上的问题，不如随着本书一起进入R语言的世界，领略R语言特有的魅力，通过对R语言的学习，重新认识大数据的价值，更深一步地提升个人价值。

随着我与统计、金融领域的朋友交流地逐步深入，我深刻地体会到，他们对R语言的实际使用也存在着很大的问题和困惑。比如，他们在某些实验室环境下，使用R语言可以很轻松、很顺利地实现预期效果，但是移植到真实环境下，面对浩瀚繁复的大数据，在使用R语言的过程中出现了很多问题。这就好比面对一座金山，他们掌握着先进的提纯技术，但他们所使用的挖掘、采集工具却还停留在石器时代！使用工具的落后，使他们要面对大量R语言之外的问题，这让他们应接不暇，甚至崩溃！有的人甚至因此认为，R语言只是一种实验室语言，至少以现在的技术水平无法将它运用到现实生活中，R语言在现实生活中广泛应用，简直是天方夜谭！

是的，如果你是一名没有计算机背景的R语言使用者，你在实际使用中也同样会遇到

许多这样或那样的问题，面对这些棘手的问题寝食难安，尝试着通过各种方式寻求解决方案。其实，在计算机领域，这些问题已经早就有了成熟、有效的解决方案。

本书的内容来自我在R语言实际使用过程中的经验总结，基本都是我在工作中使用R语言的真实记录，其中涉及计算机、互联网、数据库、大数据、统计、金融等领域，详细总结了R语言在实际使用时与Java、MySQL、Redis、MongoDB、Cassandra、Hadoop、Hive、HBase等技术综合运用的解决方案，具有实战性，可操作性强。如果你与R语言接触时间不长，本书可以让你看到R语言在各行业、各领域所散发的魅力；如果你在某行业使用R语言已经有一段时间了，可能在使用R语言的过程中遇到了瓶颈，本书将让你看到R语言在与其他计算机语言结合后所迸发的强大活力；如果你是技术人员，本书中有全局观的案例实施，也许会给你带来新的启发，甚至跟我一样，重新规划自己的职业生涯，找到学习、奋斗的新方向；如果你是企业的中高层管理者，在本书中可以看到我们已经实现的技术成果，如果需要，你甚至可以按照书中记录的详细操作步骤，直接在企业环境中实施，直接获利！

在此，我不得不强调，本书不是入门书，不讲R的语法，如果你想学习R语言的基础语言入门知识，那么，你来错地方了。但是，如果你已经具备了一定的R语言基础，但不一定具有计算机语言背景，我将告诉你R语言在真实环境下到底都能够做什么，并且详细地告诉你怎样一步一步地实施。

在与各界R语言初学者的交流中，我发现，入门后，学习R语言最大的问题，在于如何使用R语言的众多软件包，而介绍这方面的图书很难找到，只有一些网上流传的小册子。本书涉及了30个R语言包，并结合我的使用心得及案例分析，相信会解决大家R语言入门后的困扰。

本书是“R的极客理想”系列图书的第一本，姊妹篇《R的极客理想——高级开发篇》将深入介绍R语言底层原理，并使用R语言开发出企业级的应用。

本书的使用环境涉及Linux Ubuntu和Windows 7两种操作系统，R语言包的2.15.3和3.0.1两个版本，在每一节中都有明确的标识。

R语言还在不断地进步和更新，它将引导一场数据的革命，跨学科的结合是时代趋势，也是我们的机遇！

## 读者对象

本书适合以下R语言工作者：

- ❑ 计算机背景的软件工程师；
- ❑ 数据库背景的DBA；
- ❑ 数据分析背景的数据科学家；
- ❑ 统计背景的科研工作者；
- ❑ 大专院校相关专业的学生。

## 如何阅读本书

本书的内容分为四个部分。

第一部分是R基础（第1 ~ 3章），介绍了为什么要学习R语言，R语言不同版本的安装，以及R语言中常用的12个软件包。帮助读者快速了解R语言的工具包、时间序列包和性能监控包。

第二部分是R服务器（第4 ~ 5章），介绍了R语言与其他编程语言的通信，以及R语言作为服务器的应用。帮助读者打通R语言与其他编程语言的通道，并实现R语言的服务器应用。

第三部分是数据库和大数据（第6 ~ 7章），介绍了R语言与各种数据库的通信，以及R语言与Hadoop集成。帮助读者打通R语言与各种数据库层的通道，并实现R语言对基于Hadoop大数据的处理。

第四部分是附录，介绍了Java、各种数据库以及Hadoop的安装方式。笔者希望读者可以在不借助其他参考书的情况下，完成书中所有实例。

本书为工具书，每节之间没有特别的顺序要求，你可以选择任何你感兴趣的章节进行阅读。如果你是一名初学者，想全面掌握R语言，请按顺序阅读全部的章节。

## 勘误和支持

由于笔者的水平有限，加之编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。为此，笔者创建一个在线的图书交流网站（https://onbook.me），方便与读者进行沟通。如果读者在阅读过程中遇到问题，也可以在网站中留言，我将尽量在线上为你提供最满意的解答。书中的全部源代码都可以从华章公司网站（www.hzbook.com）或本书交流网站下载，我也会及时更新代码。本书为黑白印刷，关于更绚丽的彩色图片，读者运行源代码即可看到。如果你有什么宝贵意见，欢迎发送邮件至bsspirit@gmail.com，期待能够得到你真挚的反馈。

## 致谢

感谢我的团队，林伟林、林伟平、邓一硕，让我们因R语言走到一起。感谢机械工业出版社华章公司的编辑明永玲，帮助我审阅全部章节，引导我顺利完成书稿。感谢我的爸爸、妈妈和爱人，感谢你们对我工作上的支持和生活上的照顾！

谨以此书献给我最亲爱的家人以及众多R语言爱好者们！

张丹

2014年5月于中国北京

# 目　录 *Contents*

## 第三部分 数据库和大数据