《R包开发》作者Hadley Wickham：数据结构“神童”（图灵访谈）

刘敏ituring2016-10-31

资讯 **Hadley Wickham**

推荐 RStudio的首席科学家，莱斯大学的助理教授，资深Ｒ社区成员，已开发了30多个R包。因在数据处理和可视化开发工具方面的卓越贡献，获得专为统计计算而设立的约翰·钱伯斯奖。

9

收藏 enter image description here

Hadley(哈德利)出生在新西兰 · 汉密尔顿的一个从事数据统计的家庭。他的父亲布莱恩•韦翰是康奈尔大学动物育种方面的数据统计博士，妹妹获得了加州大学伯克利分校数据统计的博士学位。

如果数据结构方面存在神童一说的话，Hadley应该算一个。他曾自豪地讲述自己的经历：

> "15岁时，我的第一份工作就是开发Microsoft Access数据库，很有趣。我当时做一些数据库文档，现在人们仍然在使用我写的数据库。”

Hadley第一次接触R语言是在新西兰奥克兰大学的统计专业课上。他认为R语言是“一门用于理解数据的编程语言。”同SQL和Python一样，R语言对于数据科学家来说，是最流行的编程语言。

和Hadley一样，R编程语言也来自新西兰。R语言成立于1993年，由奥克兰大学的统计学家Ross Ihaka和Robert Gentleman一起创建，主要用于数据分析，却也存在一些怪癖（如索引数据结构的方式、物理内存存储的方式等）。所以，其他开发语言的使用者大都认为R语言很奇怪。使用过Java、VBA和PHP之后，Hadley发现R“与众不同”。“（许多程序员）认为R语言荒谬、笨拙，我不这么认为，”他说，“我认为R非常有趣。”

到美国的爱荷华州立大学攻读博士之后，Hadley开始开发**R包**。用哈德利自己的话说，开发包需要涵盖“帮助人们解决问题的代码，然后必须用文档记录下这些代码，别人才可以理解怎样使用这些代码。”他创建的第一个包，作为类项目的一部分，用于生物信息学数据的可视化。虽然这个包从未公开过，这丝毫不影响他喜欢分享的态度。

2005年，他发布了**reshape包**，广受关注，也是R包开发的起点。这个包已经被下载了成千上万次。reshape的目的是减少聚合和操作数据过程中的“乏味和痛苦”。简化数据转化的过程看上去并不是什么难事儿，但对于数据科学家和统计学家来说，这往往是最耗时的工作。

显然，Hadley很享受reshape开发包的成功。他认为现有的方法并不完美，所以需要开发出新的包。这并不是吹嘘，他有足够的信心，“我坚信我掌握了正确的开发方法，”他再次强调，“要么更好，要么更糟。”

最新力作《R包开发》，着眼于将读者从R包的使用者晋升为R包的开发者，展示了R包开发的哲学。书中详细介绍了如何将可重用的 R 函数、示例数据以及文档一起打包，以便与他人分享代码、节省开发时间、组织数据分析，尽可能让工作自动化。

- 学习R包最有用的组件，包括使用指南和单元测试
- 利用devtools自动执行任务
- 掌握良好编码风格的技巧，比如如何把函数组织成文件
- 使用devtools简化开发流程
- 发现提交包到CRAN的最佳途径

点击查看英文版

一些追随者称赞您，“明明可以靠脸吃饭，为什么还要死磕技术。”所以，是什么样的原因让你这么喜欢编码？

主要有两个原因。首先，我很享受从让人生畏的表象下挖掘出背后深层原因的过程。比如，挖掘数据整理和tidyr 背后的想法就很好，我喜欢找出事物背后更深层次的理念。

其次，编程可以帮助别人。开发R包就是一个很棒的方式，它能把我的想法转换成别人可以利用的工具。我很乐意听到R社区的各种反馈。当我听说有越来越多的人在使用我的代码，并且觉得我的代码很有用的时候，我的热情更高了。

在网上可以找到《R 包开发》的免费版。你不会担心这会减少纸质版图书的销售么？或者说，你为什么还要选择出版这本书，完全没有经济动力啊？

我写书的目的并不是赚钱，而是为了接触到尽可能多的同行。既发布电子版又出售纸质版可以很好地达成这个目标。没有充足的钱花费在纸质书上的年轻人可以从网站上下载电子版，喜欢实体书，又不怎么活跃于网络的人则可以买到一本纸质书。

我了解到《R包开发》的编写方式是开放的，你能解释下这次众包编写的体验么？

写书的一大挑战就是，很难保持长久的热情和动力。写书是一项大型项目，可能需要花费一年甚至几年的时间。如果以开放的方式编写，你就能收到不断的反馈，也更容易保持长久的动力！

校对对我来说非常痛苦。我真的很感谢R社区成员通过github 纠正了我所有的愚蠢错误！他们甚至做出了更大的修复，指出文本里的其他问题。总之，开放的编写方式让这本书变得更完善！

R 包开发是否类似于传统编程语言的 API 设计，需要注重封装性、鲁棒性、可用性等，它还有什么独特的指标吗？

我认为存在一些一般性的原则，使包开发能够顺畅进行。但我发现这些原则大都是直观的感觉，我知道该怎么操作，但没办法向别人很好地解释出来。所以我尝试把tidyverse背后的原则写了出来，你可以在https://github.com/hadley/tidyverse/blob/master/vignettes/manifesto.Rmd.链接里找到。这些都是R包开发的重要原则，它们让API更像R，帮助工具包自然顺畅地工作。

R语言是一门专为数据分析而设计的语言，但它本身也有一些古怪的习惯，比如数据结构要索引，必须存储在物理内存。是否可以借鉴C++和Spark的内存管理方式？

R并不完美，但它很好地提高了数据分析师的工作效率。同时，R语言非常灵活，它允许特定领域语言像ggplot2和dplyr解决某些子域的数据分析问题。高度灵活的副作用是，这会导致性能减慢。不同领域应该采用不同的语言：R语言用于提高人类的数据分析效率，C++用于提高计算机的运算能力。我个人并不相信某种语言可以在两方面做的都很好。（换句话说，我赞同奥斯特豪特的二分法，https://en.wikipedia.org/wiki/Ousterhout%27s_dichotomy）

用 R 语言统计与分析数据有其独到的优势，但效率稍差。R 包开发是否考虑利用 C 语言接口，开发易用性与效率兼顾的组件？

是的，很多R包已经开始使用Rcpp和C++。随着越来越多的高级编程人员学习R语言，以及越来越多的R语言使用者变成经验丰富的编程人员，我想会出现越来越多的高效率R包。

微软和IBM都有采用R语言开发项目，还有一些商业公司提供了性能更出色的R包，比如H2o。商业公司的介入对R的发展有哪些作用？

我认为，商业公司的介入对R的持续发展以及R作为一种编程语言的不断完善都具有伟大的历史意义。R目前在很多公司的项目中充当关键作用，这意味着基于R语言开发的资源会越来越多。特别令人兴奋的是，我目前参与了R财团的工作（https://www.r-consortium.org）。作为商业公司回馈R社区的有效方式，R财团把他们的资金用于R开发，帮助更多的R语言使用者。

对于你来说，RStudio是R使用者最好的开发环境。但一些读者担心你的书可能过于关注RStudio，建议最好跟RStudio分开。

除了RStudio还有其他的R语言开发工具。就数据统计来说，仅次于RStudio的还有ESS和Emacs。这些工具也很强大，但更适合高级开发者使用。所以我的书选择使用RStudio，当然我对篇幅也进行了合理地平衡，如果你不使用RStudio，可以忽略不适用的部分。（提出这些问题的读者）很可能是经验丰富的R编程人员，所以他们可以自己找到替代方案。

您为R语言，特别是R包开发，做出了大量贡献。怎么做到如此多产？

我个人认为有以下几点原因。

写作。我努力建立了写作习惯，每天早上尝试写60-90分钟。这是我起床后干的第一件事。我认为，写作对我的帮助很大。首先，我经常参考自己写的东西。因为我并不是天天都用C++，我需要经常参考@Rcpp。写作也让我意识到自己在知识和工具方面的差距，填补差距的过程帮助我更有效地解决新问题。

阅读。我读了很多内容，浏览300多种博客，关注Twitter和Stack Overflow上每一个包含R标记的内容。不是深度阅读，大部分内容我只是简要浏览。但广泛的涉略让我紧跟技术更新、编程语言变化、和其他人处理数据的新方式。遇到新问题的时候，我也能识别出基本的名称，然后用Google搜索出可能解决的方案。如果连名称都不清楚，就很难进一步研究问题。

组块。来回切换场景费时费力，所以如果同时开发多个包，我什么也干不成。所以，大部分的R包都处于闲置状态，不断积累问题和想法。积累到一定的量，我会花几天来处理包开发。

最后，也不能忽略全职开发R的作用。自我离开Rice，超过90%的工作时间会用来思考R开发问题和用R编程。这就产生了一种复合作用，开发更出色的工具（认知方面和计算方面）时，很容易就实现了新工具的建立。我甚至可以几秒内就创建一个新包，因为我有（在大脑里储存）大量可用的技术应对新问题。

——更多访谈

##更多精彩，加入图灵访谈微信！

发表新的评论

请先 登录 后发表评论

共有1条评论热门最新

vct0r

2018-09-02 10:43:54

推荐 回复

最后一个问答之前在Quora见过[doge]

已经到底啦~