R for Programmers

Advanced Techniques

Dan Zhang

CRC Press
Taylor & Francis Group

ISBN 978-1-138-62718-5

CRC Press
Taylor & Francis Group
an informa business

www.crcpress.com • an informa business

R FOR PROGRAMMERS

Dan Zhang

R for Programmers: Advanced Techniques

Dan Zhang

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-138-62718-5 (Hardback) 978-1-4987-3687-9 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: Zhang, Dan, 1983-
Title: R for programmers : advanced techniques / Dan Zhang.
Description: Boca Raton : CRC Press, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2016046363| ISBN 9781498736879 (pbk. : acid-free paper) | ISBN 9781138627185 (hardback : on demand) | ISBN 9781498736886 (ebook)
Subjects: LCSH: R (Computer program language)
Classification: LCC QA76.73.R3 Z43 2017 | DDC 005.13/3--dc23
LC record available at https://lccn.loc.gov/2016046363

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

R for Programmers: Advanced Techniques

This book is dedicated to my dearest family and the fans of R.

Preface

Why This Book

This book is the second one of the series *R for Programmers*. It focuses on R's core techniques, the advanced developing applications, and the cross-discipline integration of knowledge of R and other fields.

Early in my first book *R for Programmers: Mastering the Tools*, I have introduced how to work with 30+ tooling packages of R and how to efficiently use third-part R packages for applying existing IT knowledge to the learning process of R. However, due to the limited pages, the book discussed only the usage without the underlying principles.

This book will make up for the regret. I will mainly discuss the core techniques of R itself, including the environments, object-oriented, file management, mathematical calculations, R package development, and so on. I hope that by walking through this book my readers can deeply learn about the R language, master these core techniques of R, understand the characteristics of third-part packages, and even have capability to develop excellent packages with their own styles. Perhaps in the near future, I will be amazed at a lot of efforts because of the packages developed by you.

Another highlight of this book is the cross-discipline integration of knowledge of R and other fields. In the book I will show the readers without reservation how I incorporate R and other knowledge to maximize the power of R in different fields. I believe that this part of content will enlighten many readers who would be surprised at the ways how I use R. I also hope that this part of content will inspire my readers and that R can be learned and used by people from various industries and knowledge domains. Nowadays, R is no longer the laboratory language that was used by only scientists. Instead, it already has the capabilities of actual development and application. It is intelligent and creative in terms of mining data value, discovering data rule, and creating data wealth.

If we compare R to Kong Fu, then the book *R for Programmers: Mastering the Tools* is like guidelines for tool usage. It helps you promote productivity easily and effectively in short time. An obvious improvement of your R skills can be achieved quickly. But in a long run, you will encounter your bottleneck for various reasons and it is difficult for you to break through.

Well, this book is like the inner strength of Kong Fu. It brings you the core techniques of R itself that make you get the elementary logic of R. The book pays more stress on how to integrate R and other disciplines and fields in practice. With what you learn from this book, you will get a clear picture of R. Then it is possible for you to win tricks without tricks, or even to create your own Kong Fu style to become a great master in the future! (Um, I am too far from the topic ...)

Here, I must stress that this book is not an introductory of R. Anyone with blank background should learn some basic knowledge of R before reading this book. This book contains

advanced contents for R development, which require you to have experience of working with R and basic computer knowledge. Else you cannot understand the output of my experience in this book.

The content of this book is the summary of my usage of R in practice. It is a record of my working experience with R. This book puts stress on R's advanced development, involving the knowledge from fields of computer, statistics, mathematics, and finance.

The core content of this book consists of two topics: the advanced programming techniques of R and the cross-discipline application. For the first topic, this book discusses in detail the definition and usage of R's environments, the file management, and the new features of R version 3.1.1, which get you an experience to the low-level design of R. This book comprehensively introduces the design and usage of the four object-oriented architectures of R. The object-oriented programming architectures make R be able to develop complicated applications that follow rules in real world. Besides, the book introduces a complete process for developing R packages with cases of Daily China Weather app and game developments, which enlighten readers to develop own packages and open the door to productization with R.

As for the topic of cross-discipline application, R can easily handle the bothering mathematical calculations involving elementary or advanced mathematics, probability theory or statistics. Mathematics is no longer various models. The algorithms in the book include the collaborative filtering model, the PageRank model based on matrix computation, the trading strategy model used in finance, and the usage of genetic algorithms. With just several code lines and a few minutes, R turns our ideas into executable algorithm prototype.

Another thing I want to say is that although R is not adapted to develop games, no language can be competitive against R with just 200 lines of code to complete the 2048 game. Someone may ask me "why do you want to use R to develop games," "why not with Java," "is it the same way I use Java instead of R to development?" In fact, I just want to take game development as an instance to show the style of simplicity, the idea of freedom, and the creative of full of imaginations that belong to R. I hope my attitude of playing as an "R geek" can inspire your unlimited thoughts about R! Finally, we can turn our model into product and publish our own R package that can be used by people all over the world. How exciting it is!

When communicating with R users with different backgrounds, I find that on one hand, users with programming background are able to write clean and efficient code but due to lack of statistics, they have no idea on how to optimize the model. On the other, those with statistics background are able to design and optimize a model, but they don't know how to implement the model into product.

This book introduces several cases where I not only design the model from the perspective of academy but also implement the model into product. By studying practical cases, user with different discipline backgrounds can think from each other's perspective to find new ways for solving problem. This is another highlight!

For most programmers, it is easy to learn R but difficult to use. R has no complex programming syntax like C/C++, no need for consideration of global architecture like Java, and no flexible usage like Javascript. However, the data-oriented programming of R is totally different from other languages, which makes many programmers confusing how to use R although they have mastered its syntax.

In my opinion, learning the R language is to customize yourself, to find your real position and to make cross-discipline innovation by integrating your knowledge. It is not to copy other's idea. To use R across disciplines, you are required to combine knowledge of basic disciplines (elementary/advanced mathematics, linear algebra, probability theory, statistics) and IT

technologies (R syntax, R packages, database, algorithms). So you cannot master R until you promote your comprehensive knowledge level. In other words, once you have mastered the R language, you are outstanding.

Again, I have to emphasize that this book is not an introductory of R. It is an advanced development book. Neither the introductory syntax nor the usage of third-part packages is introduced in this book. Instead, if you have had certain basics on R and want to productize your R model, then I will tell you how to enhance the reliability and extensibility of your R program and how to publish your own packages.

This book is the second one of the series *R for Programmers*, while the third book *Quantitative Investments* will introduce the applications of R in finance. That book uses the R language to create trading models and to implement the process of automatic trading, which makes technical engineers turn their knowledge into real values.

The development environments of this book include Linux Ubuntu and Windows 7, which are declared in each section. All the programs have passed the test against R version 3.1.1.

The R language is being improved and updated. It will lead a revolution of data. Cross-discipline integration is the trend of the times and it is also the opportunity for us to grasp!

Potential Readers of This Book

This book will be helpful to the following people working with R:

- Software engineers with a computer background
- Advanced users of R
- Data scientists with a data analysis background
- Scientific researchers with a statistical background
- Students in universities and colleges

How to Read This Book

The content of this book is divided into three sections:

Section I discusses R's application to calculations and algorithms (Chapters 1 and 2), which introduces R's knowledge system and R' support for basic disciplines. By implementing various algorithms of basic disciplines with R, this part helps readers easily learn the approaches of mathematical calculations and the development of customized modeling algorithms.

Section II discusses the in-depth development of R (Chapters 3 and 4), which introduces programming related to R's kernel-related programming skills including the definition and usage of environments, and the design and application of object-oriented programming. The aim of this section is to help readers to have learn in depth R's low-level knowledge and to design complicated application structure using object-oriented programming skills.

Section III is about developing own R packages (Chapters 5 and 6). A complete development process of R package is introduced in this section. The section provides cases including Daily China Weather app and games development, which show readers how to establish their own R packages to open the door to productization using R.

There many cases that incorporate different knowledge, so it is best for readers to follow the chapters in sequence.

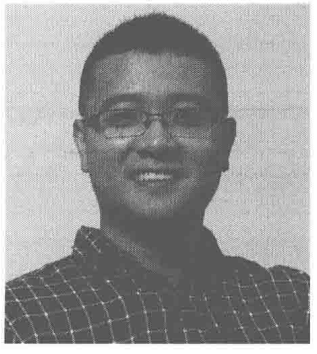
Correction and Support

Because the time spent writing this book and the author's knowledge of R are both limited, there will inevitably be some errors or incorrect viewpoints in this book. I sincerely hope that readers will point out and comment on any errors. For this purpose, I have created an online communication website for readers of this book (http://fens.me) to use to communicate. If you encounter any problems reading this book, please heave notes on this website, and I will try my best to provide a satisfactory solution. All of the source code of this book can be downloaded from the official website of CRC Press (https://www.crcpress.com) or from the online communication website, where I will update the codes in time. This book is printed in black and white, so all the colored pictures can only be achieved by running the codes of this book. I sincerely hope that you can send your valuable feedback and advice on this book to bsspirit@gmail.com.

Acknowledgments

I wish to thank He Ruijun, the acquiring editor at CRC Press, who helped promote the publication of this book. Thanks are also extended to the translator Wang Tao for his efforts on this book. I give special thanks to my parents and my wife for their support and care.

About the Author



Dan Zhang is the founder and former CTO of Qutke.com and he now works at China Minsheng Bank Corp. Ltd. as a data scientist in the company's Big Data Center. Dan has 10 years of programming experience and obtained a number of technical certificates from Sun and IBM. With rich knowledge of developing Internet application architectures, he is proficient with languages including R, Java, Nodejs. Dan masters the techniques of big data and data mining and he has accumulated lots of knowledge on statistics and finance. He is the author of *R for Programmers: Mastering the Tools and R for Programmers: Advanced Techniques*. His blog is available at http://fens.me, with global Alexa ranking of 70 k.