John R. Hubbard 著

Java数据分析

（影印版）

Java Data Analysis

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Packt

Java 数据分析（影印版）

Java Data Analysis

John R. Hubbard 著

南京　东南大学出版社

图书在版编目(CIP)数据

Java数据分析:英文/(美)约翰·R.哈伯德(John R. Hubbard)著. —影印本. —南京:东南大学出版社,2018.8

书名原文:Java Data Analysis

ISBN 978-7-5641-7736-2

Ⅰ.①J… Ⅱ.①约… Ⅲ.①JAVA语言—程序设计—英文 Ⅳ.①TP312.8

中国版本图书馆CIP数据核字(2018)第099958号

图字:10-2018-096号

Java数据分析(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼2号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787毫米×980毫米　16开本
印　　张:25.75
字　　数:504千字
版　　次:2018年8月第1版
印　　次:2018年8月第1次印刷
书　　号:ISBN 978-7-5641-7736-2
定　　价:94.00元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Author
John R. Hubbard

Reviewers
Erin Paciorkowski
Alexey Zinoviev

Commissioning Editor
Amey Varangaonkar

Acquisition Editor
Varsha Shetty

Content Development Editor
Aishwarya Pandere

Technical Editor
Prasad Ramesh

Copy Editor
Safis Editing

Project Coordinator
Nidhi Joshi

Proofreader
Safis Editing

Indexer
Tejal Daruwale Soni

Graphics
Tania Dutta

Production Coordinator
Arvindkumar Gupta

Cover Work
Arvindkumar Gupta

About the Author

John R. Hubbard has been doing computer-based data analysis for over 40 years at colleges and universities in Pennsylvania and Virginia. He holds an MSc in computer science from Penn State University and a PhD in mathematics from the University of Michigan. He is currently a professor of mathematics and computer science, Emeritus, at the University of Richmond, where he has been teaching data structures, database systems, numerical analysis, and big data.

Dr. Hubbard has published many books and research papers, including six other books on computing. Some of these books have been translated into German, French, Chinese, and five other languages. He is also an amateur timpanist.

I would like to thank the reviewers of this book for their valuable comments and suggestions. I would also like to thank the energetic team at Packt for publishing the book and helping me perfect it. Finally, I would like to thank my family for supporting me through everything.

About the Reviewers

Erin Paciorkowski studied computer science at the Georgia Institute of Technology as a National Merit Scholar. She has worked in Java development for the Department of Defense for over 8 years and is also a graduate teaching assistant for the Georgia Tech Online Masters of Computer Science program. She is a certified scrum master and holds Security+, Project+, and ITIL Foundation certifications. She was a Grace Hopper Celebration Scholar in 2016. Her interests include data analysis and information security.

Alexey Zinoviev is a lead engineer and Java and big data trainer at EPAM Systems, with a focus on Apache Spark, Apache Kafka, Java concurrency, and JVM internals. He has deep expertise in machine learning, large graph processing, and the development of distributed scalable Java applications. You can follow him at `@zaleslaw` or `https://github.com/zaleslaw`.

Currently, he's working on a Spark Tutorial at `https://github.com/zaleslaw/Spark-Tutorial` and on an Open GitBook about Spark (in Russian) at `https://zaleslaw.gitbooks.io/data-processing-book/content/`.

> Thanks to my wife, Anastasya, and my little son, Roman, for quietly tolerating the very long hours I've been putting into this book.

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1787285650`.

If you'd like to join our team of regular reviewers, you can e-mail us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Preface

> *"It has been said that you don't really understand something until you have taught it to someone else. The truth is that you don't really understand it until you have taught it to a computer; that is, implemented it as an algorithm."*
>
> *— Donald Knuth*

As Don Knuth so wisely said, the best way to understand something is to implement it. This book will help you understand some of the most important algorithms in data science by showing you how to implement them in the Java programming language.

The algorithms and data management techniques presented here are often categorized under the general fields of data science, data analytics, predictive analytics, artificial intelligence, business intelligence, knowledge discovery, machine learning, data mining, and big data. We have included many that are relatively new, surprisingly powerful, and quite exciting. For example, the ID3 classification algorithm, the K-means and K-medoid clustering algorithms, Amazon's recommender system, and Google's PageRank algorithm have become ubiquitous in their effect on nearly everyone who uses electronic devices on the web.

We chose the Java programming language because it is the most widely used language and because of the reasons that make it so: it is available, free, everywhere; it is object-oriented; it has excellent support systems, such as powerful integrated development environments; its documentation system is efficient and very easy to use; and there is a multitude of open source libraries from third parties that support essentially all implementations that a data analyst is likely to use. It's no coincidence that systems such as MongoDB, which we study in *Chapter 11, Big Data Analysis with Java*, are themselves written in Java.

What this book covers

Chapter 1, Introduction to Data Analysis, introduces the subject, citing its historical development and its importance in solving critical problems of the society.

Chapter 2, Data Preprocessing, describes the various formats for data storage, the management of datasets, and basic preprocessing techniques such as sorting, merging, and hashing.

Chapter 3, Data Visualization, covers graphs, charts, time series, moving averages, normal and exponential distributions, and applications in Java.

Chapter 4, Statistics, reviews fundamental probability and statistical principles, including randomness, multivariate distributions, binomial distribution, conditional probability, independence, contingency tables, Bayes' theorem, covariance and correlation, central limit theorem, confidence intervals, and hypothesis testing.

Chapter 5, Relational Databases, covers the development and access of relational databases, including foreign keys, SQL, queries, JDBC, batch processing, database views, subqueries, and indexing. You will learn how to use Java and JDBC to analyze data stored in relational databases.

Chapter 6, Regression Analysis, demonstrates an important part of predictive analysis, including linear, polynomial, and multiple linear regression. You will learn how to implement these techniques in Java using the Apache Commons Math library.

Chapter 7, Classification Analysis, covers decision trees, entropy, the ID3 algorithm and its Java implementation, ARFF files, Bayesian classifiers and their Java implementation, support vector machine (SVM) algorithms, logistic regression, K-nearest neighbors, and fuzzy classification algorithms. You will learn how to implement these algorithms in Java with the Weka library.

Chapter 8, Cluster Analysis, includes hierarchical clustering, K-means clustering, K-medoids clustering, and affinity propagation clustering. You will learn how to implement these algorithms in Java with the Weka library.

Chapter 9, Recommender Systems, covers utility matrices, similarity measures, cosine similarity, Amazon's item-to-item recommender system, large sparse matrices, and the historic Netflix Prize competition.

Chapter 10, NoSQL Databases, centers on the MongoDB database system. It also includes geospatial databases and Java development with MongoDB.

Chapter 11, Big Data Analysis, covers Google's PageRank algorithm and its MapReduce framework. Particular attention is given to the complete Java implementations of two characteristic examples of MapReduce: WordCount and matrix multiplication.

Appendix, Java Tools, walks you through the installation of all of the software used in the book: NetBeans, MySQL, Apache Commons Math Library, javax.json, Weka, and MongoDB.

What you need for this book

This book is focused on an understanding of the fundamental principles and algorithms used in data analysis. This understanding is developed through the implementation of those principles and algorithms in the Java programming language. Accordingly, the reader should have some experience of programming in Java. Some knowledge of elementary statistics and some experience with database work will also be helpful.

Who this book is for

This book is for both students and practitioners who seek to further their understanding of data analysis and their ability to develop Java software that implements algorithms in that field.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "We can include other contexts through the use of the `include` directive."

A block of code is set as follows:

```
Color = {RED, YELLOW, BLUE, GREEN, BROWN, ORANGE}
Surface = {SMOOTH, ROUGH, FUZZY}
Size = {SMALL, MEDIUM, LARGE}
```

Any command-line input or output is written as follows:

```
mongo-java-driver-3.4.2.jar
mongo-java-driver-3.4.2-javadoc.jar
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "Clicking the **Next** button moves you to the next screen."

> Warnings or important notes appear in a box like this.

> Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for this book from your account at `http://www.packtpub.com`. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

You can download the code files by following these steps:

1. Log in or register to our website using your e-mail address and password.
2. Hover the mouse pointer on the **SUPPORT** tab at the top.
3. Click on **Code Downloads & Errata**.
4. Enter the name of the book in the **Search** box.
5. Select the book for which you're looking to download the code files.
6. Choose from the drop-down menu where you purchased this book from.
7. Click on **Code Download**.

You can also download the code files by clicking on the **Code Files** button on the book's webpage at the Packt Publishing website. This page can be accessed by entering the book's name in the **Search** box. Please note that you need to be logged in to your Packt account.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR / 7-Zip for Windows
- Zipeg / iZip / UnRarX for Mac
- 7-Zip / PeaZip for Linux

The code bundle for the book is also hosted on GitHub at `https://github.com/PacktPublishing/Java-Data-Analysis`. We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.

Table of Contents