异步图书

Java 数据分析指南

Java Data Analysis

[美] 约翰·哈伯德（John R. Hubbard）著　高蓉 李茂 译

数据挖掘 • 大数据分析 • NoSQL • 数据可视化

Packt>

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

Java 数据分析指南

Java Data Analysis

[美] 约翰·哈伯德（John R. Hubbard）著　高蓉 李茂 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Java数据分析指南 / （美）约翰·哈伯德
(John R. Hubbard) 著 ; 高蓉, 李茂译. -- 北京 : 人民邮电出版社, 2018.12
ISBN 978-7-115-49486-3

Ⅰ. ①J… Ⅱ. ①约… ②高… ③李… Ⅲ. ①JAVA语言－程序设计 Ⅳ. ①TP312.8

中国版本图书馆CIP数据核字(2018)第228012号

版权声明

◆ 著　[美] 约翰·哈伯德（John R. Hubbard）
译　高　蓉　李　茂
责任编辑　胡俊英
责任印制　焦志炜

◆ 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
固安县铭成印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：21.75
字数：510 千字　2018 年 12 月第 1 版
印数：1－2 400 册　2018 年 12 月河北第 1 次印刷

著作权合同登记号　图字：01-2017-9210 号

定价：79.00 元

读者服务热线：(010)81055410　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

内容提要

当今，数据科学已经成为一个热门的技术领域，例如数据处理、信息检索、机器学习、自然语言处理、数据可视化等都得到了广泛的应用和发展。而 Java 作为一门经典的编程语言，在数据科学领域也有着卓越的表现。

本书旨在通过 Java 编程来引导读者更好地进行数据分析。本书包含 11 章内容，详细地介绍了数据科学导论、数据预处理、数据可视化、统计、关系数据库、回归分析、分类分析、聚类分析、推荐系统、NoSQL 数据库以及 Java 大数据分析等重要主题。

本书适合想通过 Java 解决数据科学问题的读者，也适合数据科学领域的专业人士以及普通的 Java 开发者阅读。通过阅读本书，读者将能够对数据分析有更加深入的理解，并且掌握实用的数据分析技术。

译者简介

高蓉，博士，毕业于南开大学，现任教于杭州电子科技大学。研究领域包括资产定价、实证金融、数据科学应用，已出版著作多部，发表论文数篇。感谢浙江省教育厅科研项目（编号 Y201840396）、浙江省自然科学基金项目（编号 LY17G030033）、南开大学基本科研费（编号 63185010）对翻译本书的支持。

李茂，毕业于北京师范大学，现任教于天津理工大学。热爱数据科学，从事与统计和数据分析相关的教学和研究工作。

作者简介

约翰·哈伯德（John R. Hubbard）任教于宾夕法尼亚州和弗吉尼亚州的高校，从事计算机数据分析工作长达 40 余年。他拥有宾州州立大学的计算机科学硕士学位和密歇根大学的数学博士学位。目前，他在里士满大学担任数学和计算机科学的名誉教授，他在该校讲授数据结构、数据库系统、数值分析和大数据。

哈伯德博士出版了许多著作并发表过多篇论文，除了本书，他还出版过 6 本计算领域的著作。其中某些著作已经翻译为德文、法文、中文和其他 5 种语言。此外，他还是一位业余音乐家。

我要感谢本书审阅者的宝贵意见与建议。此外，我还要感谢 Packt 出版社的强大团队出版并完善本书。最后，我要感谢我的家庭对我无微不至的支持。

审阅者简介

埃林·帕奇奥可夫斯基（Erin Paciorkowski）是乔治亚理工学院的计算机科学专家，一位优秀学者。她在国防部从事 Java 开发超过 8 年，也是乔治亚理工学院在线计算机硕士项目的研究生助教。她是一位经过认证的敏捷专家，同时持有 Security+、Project+和 ITIL 基金会证书。在 2016 年，她获得了格蕾丝·霍珀荣誉奖学金。她的兴趣包括数据分析和信息安全。

阿列克谢·季诺维也夫（Alexey Zinoviev）是 EPAM 系统的首席工程师、Java 和大数据培训师，精通 Apache Spark、Apache Kafka、Java 并发以及 JVM 内部机制。他在机器学习、大型图形处理以及分布式可扩展 Java 应用开发领域具有深厚经验。你可以通过 @zaleslaw 关注他，或通过 GitHub 关注他。

感谢我的妻子阿娜斯塔娅和可爱的儿子罗曼，在相当长的时间里包容我专心审阅本书。

前言

“有人说，技术只有教给别人，自己才能真正理解。而真相是，除非教给计算机，即作为算法实现，否则依然无法真正理解。”

——高德纳（Donald Knuth）

正如高德纳的名言，理解某种技术的最佳方法是实现。本书通过展示 Java 编程语言的实现，帮助你理解一些最重要的数据科学算法。

本书介绍的算法和数据管理技术通常归入以下领域：数据科学、数据分析、预测分析、人工智能、商业智能、知识发现、机器学习、数据挖掘，以及大数据。其中很多新颖的方法都会令人震惊！例如，ID3 分类算法、K-均值和 K-中心点聚类算法、亚马逊的推荐系统以及谷歌的 PageRank 算法，这些技术几乎影响着每一个使用网络电子设备的人。

本书之所以选择 Java 编程语言，一方面是因为这种语言使用广泛，另一方面是因为它的易获得性。此外，Java 是面向对象语言；它有优秀的支持系统，比如强大的集成开发环境；它的文档系统高效易用；它有大量开源的第三方库，这些库基本可以支持数据分析师所有会用到的实现。比如 MongoDB 系统就是使用 Java 编写的，这并非巧合，我们将在第 11 章“Java 大数据分析”中学习 MongoDB 系统。

本书内容

第 1 章，“数据科学导论”，介绍本书主题，概述了数据分析的历史发展及其在解决社会关键问题方面扮演的重要角色。

第 2 章，“数据预处理”，阐述存储数据的多种格式、数据集的管理，以及基本的预处理技术，比如排序、合并和散列。

第 3 章，“数据可视化”，包含图形、图表、时间序列、移动平均、正态和指数分布以及 Java 应用。

第 4 章，“统计”，概述基本的概率和统计原理，包括随机性、多元分布、二项分布、条件概率、独立性、列联表、贝叶斯定理、协方差和相关性、中心极限定理、置信区间以及假设检验。

第 5 章，“关系数据库”，介绍关系数据库的开发与访问，包括外键、SQL、查询、JDBC、批处理、数据库视图、子查询以及索引。你将学习如何使用 Java 和 JDBC 分析存储在关系数据库中的数据。

第 6 章，“回归分析”，讲述预测分析的一个重要部分，包括线性回归、多项式回归以及多元线性回归。你将学习在 Java 中如何使用 Apache Commons Math Library 实现这些技术。

第 7 章，“分类分析”，包括决策树、熵、ID3 算法及其 Java 实现、ARFF 文件、贝叶斯分类器及其 Java 实现、支持向量机（SVM）算法、logistic 回归、K-最近邻以及模糊分类算法。你将学习在 Java 中如何使用 Weka 库实现这些算法。

第 8 章，“聚类分析”，包括分层聚类、K-均值聚类、K-中心点聚类以及仿射传播聚类。你将学习如何在 Java 中使用 Weka 库实现这些算法。

第 9 章，“推荐系统”，包括效用矩阵、相似性度量、余弦相似性、亚马逊的 item-to-item 推荐系统、大型稀疏矩阵以及具有历史意义的网飞大奖赛（Netflix Prize competition）。

第 10 章，“NoSQL 数据库”，主要介绍 MongoDB 数据库系统，同时介绍地理空间数据库和基于 MongoDB 的 Java 开发。

第 11 章，“Java 大数据分析”，包括谷歌的 PageRank 算法及其 MapReduce 框架。需要特别注意两个典型示例——WordCount 和矩阵操作，它们都是 MapReduce 的完整 Java 实现。

附录，“Java 工具”，指导你安装本书用到的所有软件：NetBeans、MySQL、Apache Commons Math Library、javax.json、Weka 以及 MongoDB。

阅读准备

本书的重点是帮助读者理解数据分析的基本应用原理和算法，方式是指导读者通过Java编程不断深化对基本原理和算法的理解。因此，读者需要具备一定Java编程经验。如果读者还具有一些基础的统计学知识和数据库工作经验，那么学习起来会感觉更轻松。

目标读者

如果你是一名学生或者从业者，希望深入理解数据分析，并希望在该领域提高Java的算法开发能力，本书正是为你而写！

排版约定

本书通过不同的文本样式区分不同种类的信息。这里举例说明这些样式并解释其含义。

书中的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、虚拟URL、用户输入，都会按这样的字体显示："我们可以通过使用`include`指令包含其他环境。"

代码块设置如下：

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

命令行的输入和输出格式如下：

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

新术语和**重要词汇**以黑体表示。例如，你会在屏幕上看到在菜单或对话框中出现这样的文字："点击**下一个**按钮可以移到下一屏。"

注释

警告或重要的注释形式如下。

提示

提示和技巧形式如下。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

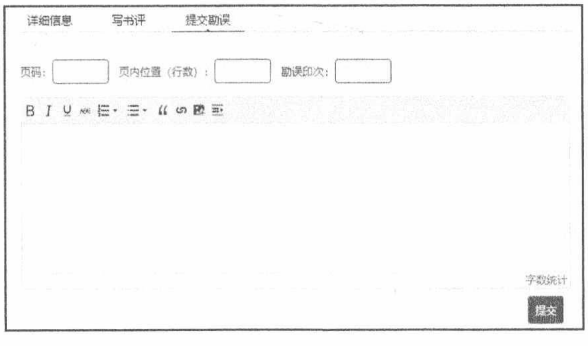
本书提供配套源代码，要获得该配套资源，请在异步社区本书页面中单击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

第 1 章
数据科学导论

数据分析是对数据进行组织、清洗、转换和建模的过程，目的是获取有价值的信息和新知识。数据分析、商业分析、数据挖掘、人工智能、机器学习、知识发现和大数据，这些术语也可以用来描述相似的过程。这些领域之间的区别更多体现在应用领域，而非基础本质。有人认为，这些领域都是数据科学新学科的一部分。

在从组织化数据中获取有效信息的过程中，关键步骤是应用计算机科学算法进行管理。而本书的重点就是这些算法。

数据分析是一个历久弥新的领域。它起源于数值方法和统计分析的数学领域，可以追溯至 18 世纪。近年来，随着互联网愈加普遍和海量数据逐渐可得，许多数据科学方法受到越来越多的关注，随后我们将研究这些算法。

在第 1 章中，我们来讲述数据分析史上的一些著名案例。这些案例可以帮助我们理解这门科学的重要性和未来前景。

1.1 数据分析起源

数据与文明一样历史悠久，甚至年代更为古老。1.7 万年前，法国拉斯科的原始居民为了纪念他们最伟大的狩猎胜利，尝试以洞穴壁画的形式记录这些胜利。这些记录为我们提供了旧石器时代人类活动的数据。从现代意义上讲，这些数据并没有被分析，也没有为我们提供新知识。但是，这些数据的存在本身就证明了人类需要使用数据保存自己的思想。

5000 年前，美索不达米亚的苏美尔人在泥板上记录了更为重要的数据。那些楔形文字记录了与日常商业交易相关的大量会计数据。为了运用数据，苏美尔人不仅发明了文字，还发明了人类文明史上的第一个数字系统。