HZ BOOKS 华章IT

[PACKT] PUBLISHING

资深数据分析咨询师多年经验结晶，内容全面而深入，为高效利用各类工具和算法进行数据分析提供最佳指导

通过大量典型数据分析案例，全面阐释分类分析、聚类分析、数据可视化及预测方面的各种技术和方法，为快速掌握并灵活运用数据分析技术提供最佳实践指南

大数据技术丛书

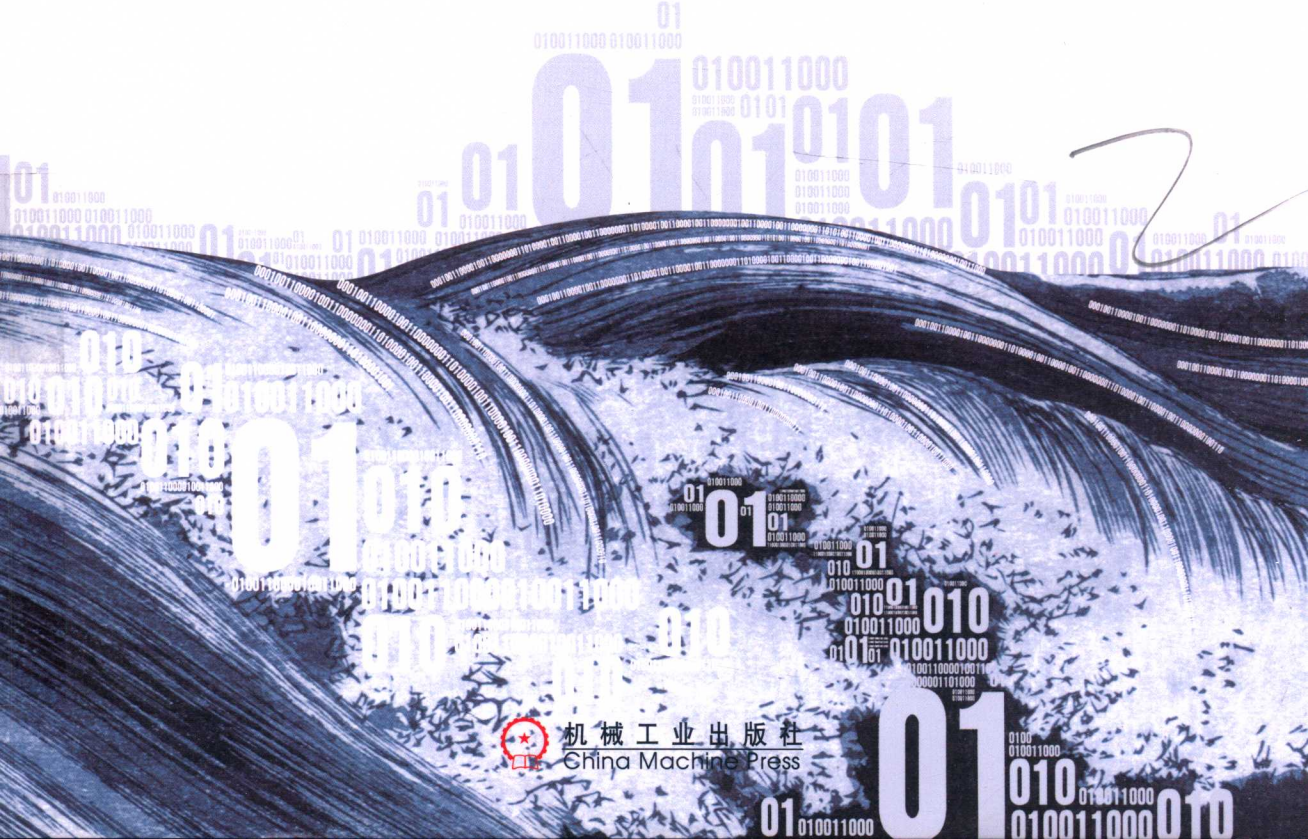
Practical Data Analysis, Second Edition

实用数据分析

（原书第2版）

[美] 赫克托·奎斯塔 (Hector Cuesta)
桑帕斯·库马尔 (Dr. Sampath Kumar) 著

刁晓纯 译



机械工业出版社
China Machine Press

大数据技术丛书

Practical Data Analysis, Second Edition

实用数据分析

（原书第2版）

[美] 赫克托·奎斯塔（Hector Cuesta）
桑帕斯·库马尔（Dr. Sampath Kumar） 著

刁晓纯 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

实用数据分析（原书第 2 版）/（美）赫克托·奎斯塔（Hector Cuesta），（美）桑帕斯·库马尔（Dr. Sampath Kumar）著；刁晓纯译．—北京：机械工业出版社，2017.8
（大数据技术丛书）
书名原文：Practical Data Analysis, Second Edition

ISBN 978-7-111-57921-2

I. 实… II. ① 赫 … ② 桑… ③ 刁… III. 统计数据－统计分析 IV. O212.1

中国版本图书馆 CIP 数据核字（2017）第 216292 号

本书版权登记号：图字：01-2017-0487

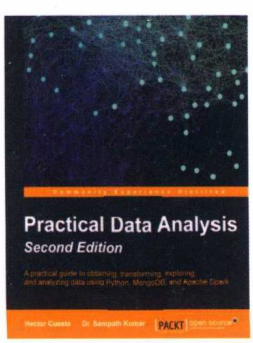
实用数据分析（原书第 2 版）

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：陈佳媛　　责任校对：李秋荣
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2017 年 9 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：15.5
书　　号：ISBN 978-7-111-57921-2　　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）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

内容简介

本书共15章：第1章探讨数据分析的基本原理和数据分析步骤；第2章解释如何清洗并准备数据；第3章展示在JavaScript可视化框架下应用D3.js来实现各类数据的可视化方法；第4章介绍如何应用朴素贝叶斯算法来区分垃圾邮件；第5章讲解应用动态时间规整方法寻找图像间的相似性；第6章介绍使用随机游走算法和可视化的D3.js动画技术模拟股票价格；第7章介绍核岭回归（KRR）的原理以及应用；第8章描述如何使用支持向量机方法进行分类分析；第9章介绍应用细胞自动机方法对传染病进行建模；第10章解释如何应用Gephi从Facebook获取社交媒体图谱并实现可视化；第11章介绍如何应用Twitter数据进行情感分析；第12章介绍如何使用MongoDB进行数据处理和聚合；第13章详细介绍如何在MongoDB数据库中应用MapReduce编程模型；第14章介绍如何应用Jupyter和Wakari开展线上数据分析；第15章介绍如何使用Apache Spark处理数据。



原书封面

作者简介

赫克托·奎斯塔(Hector Cuesta)

Dataxios（一家机器智能研发公司）的创始人及首席数据科学家，拥有信息学士及计算机科学硕士学位。他在金融、零售、金融科技、在线学习、人力资源等领域提供数据驱动产品设计的咨询服务。可以关注他的推特：https://twitter.com/hmCuesta。

桑帕斯·库马尔(Dr. Sampath Kumar)

Telangana 大学应用统计系的助理教授和系主任，拥有5年研究生教学经验，有超过4年的工作经验。他是SAS和MATLAB软件高级编程人员，擅长利用SPSS、SAS、R、Minitab、MATLAB等软件进行数据统计。他在不同的应用学科和纯统计专业，如预测建模、应用回归分析、多变量数据分析、运营管理等方面具有教学经验。

The Translator's Words 译 者 序

早在2013年7月，因为参加了“数据驱动企业分析变革商业”2013第三届大数据世界论坛，认识了机械工业出版社的编辑王春华。当时从事数据方面的工作已经3年多，身处国内大型央企，我所涉及的数据分析工作非常广泛，既跨越了银行保险、公共服务、电子商务及速递物流等行业，也包含了对客户、渠道、价格、实物网效率、经营业绩等多方面的分析。在当时的工作中遇到了很多问题，既有组织方面的，也有方法效率方面的，王编辑推荐我阅读本书第1版原著，并且我也有幸参予翻译了部分章节。令人惊喜的是，书中所介绍的广泛案例、先进的方法以及诸多便利的工具都对数据分析工作有很多帮助和值得借鉴的地方。

针对本书，我的主要体会有三方面：

第一，本书包含丰富的案例。书中介绍的案例涉及垃圾邮件的分类分析、图像匹配、流行病暴发事件分析、社交网络的数据获取和分析、对文本型数据进行情感分析、股票价格以及黄金价格走势分析等。

第二，本书所涉内容包含了数据分析的全流程，包括了数据准备和处理、多类型建模、数据可视化展示等。初次接触数据分析的读者可以由浅入深地了解分析的全貌。

第三，本书充分体现了大数据的特点，既介绍了对结构化数据的处理也介绍了对非结构化数据的处理，数据类型丰富。书中所涉数据包括时间序列数据、数值型数据、多维度数据和社交媒体数据、文本型数据等多种形式，可以帮助读者获得对数据分析的真知灼见。

时隔几年，机械工业出版社联系上我，询问我是否愿意翻译本书第2版，我二话不说接下了这个任务，这几年随着数据工作方面的积累，对于本书，除了有更深的体会，也重新回顾、整理了当年翻译的内容。随着“大数据”技术的发展，本书最后一章也新增了对Cloudera VM和Apache Spark的介绍，使读者了解其在大数据领域的地位，并掌握一些常见的方法和操作。这又是一次温故而知新的历程。

书中部分内容是按照原文直译的，难免有不完整或者偏颇的地方，请读者批评指正，也欢迎广大读者与我交流沟通，我的邮箱是jacqueline_dut@hotmail.com。

刁晓纯

2017年6月

作者简介 *About the Author*

Hector Cuesta Dataxios（一家机器智能研发公司）的创办人及首席数据科学家，拥有信息学士及计算机科学硕士学位。他在金融、零售、金融科技、在线学习、人力资源等领域提供数据驱动产品设计的咨询服务。在空闲时间，他热衷于研究机器人。可以关注他的推特：https://twitter.com/hmCuesta。

本书献给我的妻子 Yolanda 和我可爱的孩子 Damian 和 Issac，他们为我的生活带来了无比的快乐。同时把本书献给我的父母 Elena 和 Miguel，感谢他们对我的支持和爱护。

Dr. Sampath Kumar Telangana 大学应用统计系的助理教授和系主任，他拥有理学硕士、哲学硕士和统计学博士学位，拥有 5 年研究生教学经验，有超过 4 年的工作经验。他是 SAS 和 MATLAB 软件高级程序员，专长是利用 SPSS、SAS、R、Minitab、MATLAB 等软件进行数据统计。他在不同的应用学科和纯统计专业（如预测建模、应用回归分析、多变量数据分析、运营管理等）方面具有教学经验。

About the Reviewers 审校者简介

Chandana N. Athauda 目前是 BAG（Brunei Accenture Group）的员工，他在 Brunei 是一名技术顾问。他主要关注商务智能、大数据和数据可视化工具技术。他已经在 IT 行业从业超过 15 年（曾获前微软最有价值员工和微软 TFS 管理员）。他对 IT 行业充满了工作热情，他的工作职业从程序员贯穿到技术顾问。

如果有兴趣与 Chandana 讨论本书，请发送邮件到 info@inzeek.net 或是上推特 @inzeek。

Mark Kerzner 大数据架构师及培训师。他是 Elephant Scale 的创始人及负责人，这家企业为不同领域提供大数据的顾问咨询及培训。同时他也是《 HBase Design Patterns 》一书的作者。

我要感谢我的联合创始人 Sujee Maniyam 和他的同事 Tim Fox，还要感谢所有的老师及学生。最后同样重要的是，感谢家人对我的帮助。

前　言 Preface

本书提供了一系列将数据转化为重要结论的现实案例。书中覆盖了广泛的数据分析工具和算法，用于进行分类分析、聚类分析、数据可视化、数据模拟以及预测。本书旨在帮助读者了解数据从而找到相应的模式、趋势、相互关系以及重要结论。

书中所包括的实用项目充分利用了 MongoDB、D3.js 和 Python 语言，并采用代码片段和详细描述的方式呈现本书的核心概念。

本书主要内容

第 1 章探讨数据分析的基本原理和数据分析步骤。

第 2 章解释如何清洗并准备好数据来开展分析，同时介绍数据清洗工具 OpenRefine 的使用方法。

第 3 章展示在 JavaScript 可视化框架下应用 D3.js 语言来实现各类数据的可视化方法。

第 4 章介绍应用朴素贝叶斯（Naive Bayes）算法来区分垃圾文本的一种二元分类法。

第 5 章展示一个应用动态时间规整方法来寻找图像间相似性的项目。

第 6 章解释如何使用随机漫步算法和可视化的 D3.js 动画技术来模拟股票价格。

第 7 章介绍核岭回归（Kernel Ridge Regression，KRR）的原理以及如何使用此方法和时间序列数据来预测黄金价格。

第 8 章描述如何使用支持向量机的方法进行分类分析。

第 9 章介绍对流行病进行模拟计算的基本概念并解释如何应用细胞自动机方法、D3.js 和 JavaScript 语言来模拟流行病爆发。

第 10 章解释如何应用 Gephi 从 Facebook 获取社交媒体图谱并使之实现可视化。

第 11 章解释如何应用 Twitter 的应用程序编程接口（API）来获取 Twitter 的数据。读者也将看到如何改进文本分类分析方法并将其应用于情感分析。这一过程在自然语言工具包（Natural Language Toolkit，NLTK）中应用了朴素贝叶斯算法。

第 12 章介绍在 MongoDB 数据库中进行基本操作以及分组、过滤和聚合的方法。

第 13 章详细介绍如何在 MongoDB 数据库中应用 MapReduce 编程模型。

第 14 章解释如何使用 Wakari 平台，同时介绍在 IPython 中运用 pandas 进行数据处理和使用 PIL 图像处理库的方法。

第 15 章介绍如何在 Cloudera VM 上使用分布式文件系统及数据环境。最后，利用实际案例介绍 Apache Spark 的主要特征。

阅读准备

使用本书需要掌握如下技术：

- Python
- OpenRefine
- D3.js
- mlpy
- NLTK
- Gephi
- MongoDB

读者对象

本书主要面向那些希望能够实际开展数据分析和数据可视化的软件开发人员、分析人员、计算机科学家。同时，本书也希望能够为读者提供包含时间序列数据、数值型数据、多维度数据和社交媒体数据、文本型数据等多种数据形式的实际案例，以帮助读者获得对数据分析的真知灼见。

读者不需要具备数据分析的经验，但仍需要对统计学和 Python 编程有基础性的了解。

下载本书相关资源

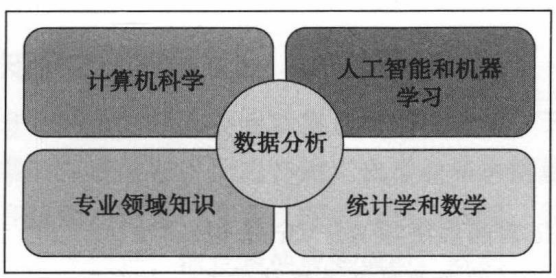
读者可登录华章网站（http://www.hzbook.com）下载本书的相关资源。

目　录 *Contents*

第 1 章 Chapter 1

开　始

数据分析是将原始数据进行排序和组织的过程，是用于帮助解释过去和预测未来的一系列方法。数据分析不只针对数字，而是关于如何设定或提出问题，演化解释，以及验证假设的过程。数据分析（Data Analysis）横跨了计算机科学、人工智能和机器学习、统计学和数学以及专业领域知识等多个领域，如下图所示。



这些技能都是很重要的，为了更好地了解问题及优化的解答，让我们来定义这些领域。

1.1　计算机科学

计算机科学为数据分析提供了工具。大规模数据的产生使计算机分析变得至关重要，对编程、数据库管理、网络管理和高性能计算的需求也逐步增多。有基本的 Python（或其他任何一种高级编程语言）编程经验对于理解本章的内容也是必不可少的。

1.2 人工智能

根据 Stuart Russell 和 Peter Norvig 二人的定义：

“人工智能（AI）一定与聪明的程序有关，那么让我们开始着手去编写这样的聪明程序吧。”

换句话说，人工智能研究那些可以模拟智能行为的算法。在数据分析中，我们应用人工智能来实施那些需要推理、相似性搜索或者无监督分类的智能活动。像是深度学习的领域就会依赖人工智能演算法，现在正使用的场景有聊天机器人、推荐引擎、图像分类，等等。

1.3 机器学习

机器学习（Machine Learning，ML）是对计算机的算法进行研究，从而使机器学会如何在特定情况或既定模式下进行反应。根据 Arther Samuel（1959）的定义：

“机器学习的研究领域是指在没有明确编程的情况下，赋予计算机进行学习的能力。”

机器学习所对应的一系列算法，根据该算法的训练过程，大致可以划分为三组：

- 有监督学习（supervised learning）
- 无监督学习（unsupervised learning）
- 强化学习（reinforcement learning）

这些算法贯穿本书，且伴随一些实际例子，引导读者了解从初始数据问题到编程化解决方案的全过程。

1.4 统计学

2009 年 1 月，Google 首席经济师 Hal Varian 说：

“我坚信未来十年最热门的职业将会是统计学家。人们以为我在开玩笑，但是谁会猜到在 20 世纪 90 年代计算机工程师曾是最热门的职业呢？”

统计学是对获取、分析和解读数据的方法加以发展和应用。数据分析包括了一系列统计学的技巧，例如：模拟、贝叶斯方法、预测、回归、线性分析及分类等。

1.5 数学

数据分析利用诸多种数学技术，如线性代数（向量和矩阵、因式分解以及特征值）、数值法和条件概念算法等。在本书中，每一章都是独立的并包含必要的数学知识。