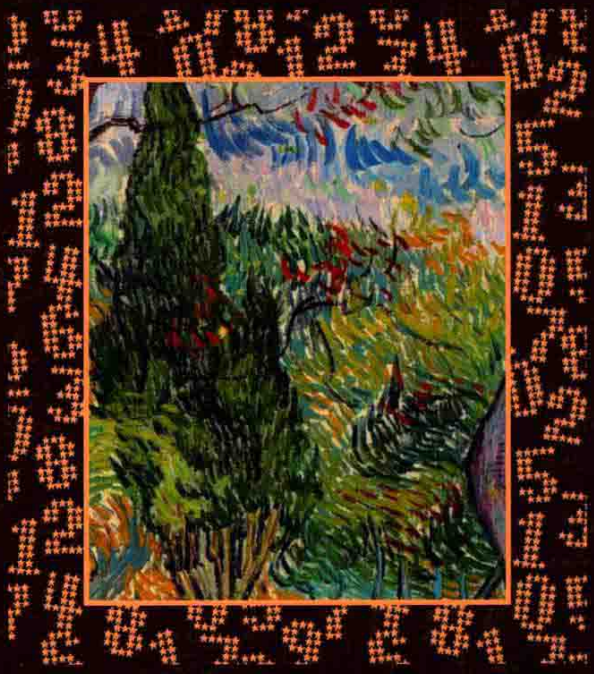
HZ BOOKS
华章IT

教育部-阿里云产学合作协同育人项目成果

数据科学与工程技术丛书

大数据分析
原理与实践

王宏志◎编著
（哈尔滨工业大学）



BIG DATA ANALYSIS
PRINCIPLE AND PRACTICE

机械工业出版社
China Machine Press

数据科学与工程技术丛书

BIG DATA ANALYSIS
PRINCIPLE AND PRACTICE

大数据分析
原理与实践

王宏志◎编著
（哈尔滨工业大学）

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据分析原理与实践 / 王宏志编著 . —北京：机械工业出版社，2017.6
（数据科学与工程技术丛书）

ISBN 978-7-111-56943-5

I. 大… II. 王… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 119356 号

本书介绍大数据分析的多种模型、所涉及的算法和技术、实现大数据分析系统所需的工具以及大数据分析的具体应用。

本书共 16 章。第 1 章为绪论，阐释大数据、大数据分析等概念，并对本书内容进行概述；第 2 ～ 7 章介绍大数据分析模型的建立方法以及关联分析模型、分类分析模型、聚类分析模型、结构分析模型和文本分析模型；第 8 ～ 11 章介绍大数据分析所涉及的技术，包括数据预处理、降维、数据仓库、各种算法等；第 12 ～ 14 章介绍三种用于实现大数据分析算法的平台，即大数据计算平台、流式计算平台和大图计算平台；第 15 章和第 16 章介绍两类大数据分析的具体应用，分别讲述社交网络和推荐系统中的大数据分析。

本书可作为高等院校大数据分析相关课程的教材，也可以作为从事大数据相关工作的工程技术人员的参考用书。

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：吴晋瑜　　责任校对：殷　虹
印　　刷：三河市宏图印务有限公司　　版　　次：2017 年 7 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：28.75
书　　号：ISBN 978-7-111-56943-5　　定　　价：79.00 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) 88378991　88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

序

当前，一场科技革命浪潮正席卷全球，这一次，IT 技术是主角之一。云计算、大数据、人工智能、物联网，这些新技术正加速走向应用。很快，它们将渗透至我们生产、生活中的每个角落，并将深刻改变我们的世界。

在这些新技术当中，云计算作为基础设施，将全面支撑各类新技术、新应用。我认为：云计算，特别是公共云，将成为这场科技革命的承载平台，全面支撑各类技术创新、应用创新和模式创新。

作为一种普惠的公共计算资源与服务，云计算与传统 IT 计算资源相比有以下几个方面的优势：一是硬件的集约化；二是人才的集约化；三是安全的集约化；四是服务的普惠化。

公共云计算的快速发展将带动云计算产业进入一个新的阶段，我们可以称之为“云计算 2.0 时代”，云计算对行业演进发展的支撑作用将更加凸显。

云计算是“数据在线”的主要承载。“在线”是我们这个时代最重要的本能，它让互联网变成了最具渗透力的基础设施，数据变成了最具共享性的生产资料，计算变成了随时随地的公共服务。云计算不仅承载数据本身，同时也承载数据应用所需的计算资源。

云计算是“智能”与“智慧”的重要支撑。智慧有两大支撑，即网络与大数据。包括互联网、移动互联网、物联网在内的各种网络，负责搜集和共享数据；大数据作为“原材料”，是各类智慧应用的基础。云计算是支撑网络和大数据的平台，所以，几乎所有智慧应用都离不开云计算。

云计算是企业享受平等 IT 应用与创新环境的有力保障。当前，企业创新，特别是小微企业和创业企业的创新面临 IT 技术和 IT 成本方面的壁垒。云计算的出现打破了这一壁垒，IT 成为唾手可得的基础性资源，企业无须把重点放在 IT 支撑与实现上，可以更加聚焦于擅长的领域进行创新，这对提升全行业的信息化水平以及激发创新创业热情将起到至关重要的作用。

除了发挥基础设施平台的支撑作用外，2.0 时代的云计算，特别是公共云计算对产业的影响将从量变到质变。我认为，公共云将全面重塑整个 ICT 生态，向下定义数据中心、IT 设备，甚至是 CPU 等核心器件，向上定义软件与应用，横向承载数据与安全，纵向支

撑人工智能的技术演进与应用创新。

对我国来说，发展云计算产业的战略意义重大。我认为，云计算已不仅仅是“IT基础设施”，它将像电网、移动通信网、互联网、交通网络一样，成为“国家基础设施”，全面服务国家多项重大战略的实施与落地。

云计算是网络强国建设的重要基石。发展云计算产业，有利于我国实现IT全产业链的自主可控，提高信息安全保障水平，并推动大数据、人工智能的发展。

云计算是提升国家治理能力的重要工具。随着大数据、人工智能、物联网等技术应用到智慧城市、智慧政务建设中，国家及各城市的治理水平和服务能力大幅提升，这背后，云计算平台功不可没。

云计算将全面推动国家产业转型升级。云计算将支撑“中国制造2025”“互联网+”战略，全面推动“两化”深度融合。同时，云计算也为创新创业提供了优质土壤，在“双创”领域，云计算已真正成为基础设施。

在DT时代，我认为计算及计算的能力是衡量一个国家科技实力和创新能力的重要标准。只有掌握计算能力，才具备全面支撑创新的基础，才有能力挖掘数据的价值，才能在重塑ICT生态过程中掌握主导权。

接下来的几年，云计算将成为全球科技和产业竞争的焦点。目前，我国的云计算产业具备和发达国家抗衡的能力，而我们对数据的认知、驾驭能力及对资源的利用开发和人力也是与发达国家等同的。因此，我们正处在一个“黄金窗口期”。

我一直认为，支撑技术进步和产业发展的最主要力量是人才，未来世界各国在云计算、大数据、AI等领域的竞争，在某种程度上会转变为人才之争。因此，加强专业人才培养将是推动云计算、大数据产业发展的重要抓手。

由于是新兴产业，我国云计算、大数据领域的人才相对短缺。作为中国最大的云计算服务企业，阿里云希望能在云计算、大数据领域的人才培养方面做出努力，将我们在云计算、大数据领域的实践经验贡献到高校的教育中，为高校的课程建设提供支持。

与传统IT基础技术理论相比，云计算和大数据更偏向应用，而这方面恰恰是阿里云的优势。因此，我们与高校合作，优势互补，将计算机科学的理论和阿里云的产业实践融合起来，让大家从实战的角度认识、掌握云计算和大数据。

我们希望通过这套教材，把阿里云一些经过检验的经验与成果分享给全社会，让众多计算机相关专业学生、技术开发者及所有对云计算、大数据感兴趣的企业和个人，可以与我们一起推动中国云计算、大数据产业的健康快速发展！

胡晓明

阿里云总裁

前　言

本书的缘起与成书过程

大数据经过分析能够产生高价值，这无疑已在大数据火爆的今天成为共识，从而使得大数据分析在“大数据＋”涉及的领域（如工业、医疗、农业、教育等）有了广泛的应用。大数据分析的相关知识不仅是大数据行业的从业人员应该必备的，也是和大数据相关的各行各业的从业者需要了解的。

然而，人们对大数据分析的解读有多个不同方面。从“分析”的角度解读，大数据分析可以看作统计分析的延伸；从“数据”的角度解读，大数据分析可以看作数据管理与挖掘的扩展；从“大”的角度解读，大数据分析可以看作数据密集高性能计算的具体化。

而大数据分析的有效实施也需要多个方面的知识。从分析的角度来讲，需要统计学、数据分析、机器学习等方面的知识；从数据处理的角度来讲，需要数据库、数据挖掘等方面的知识；从计算平台的角度来讲，需要并行系统和并行计算的知识。

上述多样化造成了目前大数据分析的教材和参考书的多样化：有些书重点介绍统计学或者机器学习知识，突出“分析”；有些书重点介绍实现平台和技术，突出“大”；有些书重点介绍数据挖掘知识及其应用，突出“数据”。笔者认为，这三类知识对大数据分析都是必不可少的，于是试图编写一本教材来融合这三类知识，给读者展示一个相对广阔的大数据分析图景。

也正是因为解读的角度和所需知识的多样化，本书的成书过程也比较曲折。在成书的过程中，笔者对大数据分析的认识也在不断加深，因而在编写过程中几次变换结构和体例。由于笔者主要从事数据相关工作，所以起初以大数据分析算法和相关技术为主，对数据分析模型方面的知识只是一笔带过。在和业内人士的交流中发现，对于很多读者来说，了解分析模型可能更重要，因为很多分析算法和大数据分析所需的技术都有平台实现，分析模型却需要了解业务的人来建立，于是笔者增加了较多数据分析模型方面的内容。而后通过和阿里云的合作，笔者又进一步了解了大数据分析的需求，于是增加了数据预处理等内容，并基于阿里云的技术和平台对书中的一些内容做了实现。这就是本书现在的版本。

本书的内容

本书力求系统地介绍大数据分析过程中的模型、技术、实现平台和应用。考虑到不同部分的侧重不同，故采取了不同的写作方法，尽可能使本书的内容适合更多的读者阅读。

模型部分主要突出了大数据分析模型的描述方法。通过这一部分的学习，读者可以在不考虑实现的情况下，针对应用需求建立大数据分析模型，即使不了解实现平台和具体技术，读者也可以独立学习这部分内容。在实践中，可以将分析模型表达为 R 语言，甚至像阿里云提供的可视化工具中那样分析流程，即使不掌握算法等方面的技术，同样可以进行大数据分析。

当然，如果对大数据分析相关技术有深入了解，会更加快速有效地进行分析，因而技术部分介绍了大数据分析所涉及的技术，重点在于解决大数据分析的效率和可扩展性问题。

“工欲善其事，必先利其器”，有了好的开发平台，就可以有效地实现相关的技术，因而实现平台部分介绍了多种开发大数据分析系统的实现平台。

最后两章针对“推荐系统”和“社交网络”这两个大数据分析的典型应用涉及的一些模型和技术进行了介绍，也是前面内容在应用中的具体体现。

“大数据”是一个比较宽泛的概念，本书围绕着分析过程进行讲解，突出大数据的特点，与大数据算法、大数据系统、大数据程序的编程实现、机器学习、统计学等书籍具有互补性，读者可以相互参考。

为方便读者的学习，笔者总结了一些大数据分析常用系统和工具的安装与配置方法，读者可登录华章网站（www.hzbook.com）在本书网页中下载文档。

本书没讲什么

由于大数据分析涉及的内容过于宽泛，尽管笔者试图从多个角度介绍大数据分析，但是限于本书的写作周期和篇幅，有一些读者关心的内容并没有包括在本书之中，比如：

- 数据流分析算法
- 神经网络 / 深度学习
- 大数据可视化
- 大图分析算法
- 大数据分析技术在医疗、社会安全、教育、工业等多个领域的应用

一方面，读者可以阅读相关的书籍了解这些领域的内容；另一方面，笔者也正在筹划，期望能够在本书的再版中列入上述内容。

致使用本书的教师

本书涉及多方面内容，对于教学而言，本书适用于多门课程的教学，除了直接用于“大数据分析”或者“数据科学”课程的教学之外，还可以作为“数理统计”“数据挖掘”“机器学习”等课程的补充教材。

针对不同专业的教学，教师可以选择不同的内容。针对计算机科学专业的本科生或者研究生，可以全面讲授本书的内容，但深度和侧重点上可以有所差别。针对培养数据科学家的“数据科学”专业的学生，如果培养方案中没有计算机系统和算法相关的课程，可以重点讲授第 1 ～ 7 章的内容，第 8 ～ 11 章可以着重讲解技术的选用而不是原理，第 15 ～ 16 章着重讲解背景和模型，其中的算法部分可以略去。针对培养工程师的技术类课程或者培训，可以重点讲授第 8 ～ 14 章，第 1 ～ 7 章中对模型的介绍可以略去，仅通过例子讲授模型的形式就可以。

致使用本书的学生

笔者希望为学生提供一个大数据分析的较为全面的图景，这使得本书的不同部分有着不同的讲述方式。请读者注意，碰到并不妨碍内容理解的公式和推导，可以跳过，应着重理解其背后的原理。

由于本书涉及的内容比较多，相关的背景知识也是必不可少的，因此本书的读者应有一些“线性代数”和“概率论”方面的数学知识，因为对一些模型和算法的描述不得不用到矩阵和概率。第 6 章和第 14 章用到了一些“图论”方面的基本概念。具备“数据库系统”的相关知识对理解本书的内容会很有帮助。第 12 ～ 14 章涉及一些“计算机系统”和“分布式系统”的基本知识。如果读者有“数理统计”“机器学习”和“数据挖掘”的知识，那么在阅读本书时会轻松得多，但这三类知识并不是阅读本书所必备的。本书包含了部分“数理统计”“机器学习”和“数据挖掘”的知识，主要分布在第 2 ～ 9 章中。

此外，在撰写本书的过程中，笔者注意和已经出版的《大数据算法》[⊖]一书在内容上做了明确的区分，因而对于《大数据算法》中介绍的内容，本书不再赘述。

致使用本书的专业技术人员

本书可以作为一本大数据分析的参考书，供专业技术人员阅读。各部分针对的人群有所不同，可以单独查阅涉及的主题。

如果读者是一名数据科学家，可以根据业务需求参考第 2 ～ 7 章中的模型，其中大部

⊖ 该书已由机械工业出版社出版，ISBN978-7-111-50849-6。——编辑注

分模型都可以找到实现的源代码，一些云计算平台（如阿里云）也支持其中大部分模型的实现。在涉及大规模数据的可扩展性时，读者可以参考第 8 ～ 11 章的内容，选择合适的特征、数据缩减方法、数据管理方法和算法。第 15 ～ 16 章中的内容可以作为一些大数据分析的案例，供实际数据分析时参考。

如果读者是一名大数据方面的算法研究或者开发人员，可以参考第 2 ～ 7 章中的数据分析模型方面的知识。第 8 ～ 10 章介绍了为算法进行数据准备和数据管理方面的背景知识，第 11 章介绍了部分大数据分析算法的知识，但相对简略，读者可以进一步阅读《大数据算法》一书。第 12 ～ 14 章介绍了分析算法实现的平台。

如果读者是一名大数据分析方面的系统工程师，可以参考第 2 ～ 11 章介绍的系统实现原理，第 12 ～ 14 章介绍了系统实现的平台并给出了一些例子，可以和相关的程序设计书籍对照阅读。本书中涉及的平台的安装和配置方法，读者可登录华章网站下载相关文档。

致谢

感谢哈尔滨工业大学的李建中教授、高宏教授以及国际大数据计算研究中心的诸位同事对本书的编写给出的指导和建议，以及在专业上对我的帮助。

在本书的撰写过程中，哈尔滨工业大学的李东升、袁芳怡、孙铭、韩姗珊、张浩然、王雅萱、黎竹平、苏钰、李佳红、马妍娇、孙芳媛等同学在资料翻译、搜集、整理、文本校对、制图等多个方面提供了帮助和支持，孟凡山、石乾坤、王鹤澎、李斯泽、窦隆绪等同学基于阿里云平台对本书中的部分模型和算法进行了实现，在此对他们表示感谢。

非常感谢我的爱人黎玲利副教授，感谢她一如既往地对我工作的支持，不但对书稿提出了许多有益的建议，还在本书的写作期间为我们的家添了可爱的宝宝——壮壮。感谢我的母亲和岳母，她们悉心帮助我们料理家务、照顾壮壮，使得我有时间专注于本书的写作。

本书的编写得到了阿里云公司的大力协助，入选“教育部 – 阿里云产学合作协同育人云计算大数据系列教材改革项目”（编号 2016 01001007）。感谢阿里云的李妹芳女士，她在本书成书的过程中提供了大量有益的建议，同时协助我和阿里云的工程师及时沟通，使得在实现过程中遇到的问题得以快速解决。感谢阿里云公司的张良模、宁尚兵、王勇、石立勇、李博、王晓斐等同仁在本书编写过程中给予的帮助和支持。

在本书的成书过程中，我和机械工业出版社保持着愉快的合作，感谢机械工业出版社朱劼编辑对我的帮助和支持。

还要感谢在哈尔滨工业大学选修“大数据管理与分析”课程的同学，你们所提出的意见和建议对本书的写作大有裨益。

最后，作者关于大数据管理和分析方面的研究以及本书的写作还得到了国家自然科学基金项目（编号：U1509216，61472099）、国家科技支撑计划项目（编号：2015BAH10F01）、哈尔滨工业大学研究生教育教学改革研究项目（编号：JGYJ-201527）、黑龙江省留学回国人员基金（编号：LC2016026）和微软－教育部语言语音重点实验室经费的资助，在此表示感谢。

本书涉及的内容比较多，且跨越了多个快速发展的领域，而有些领域并不是笔者的专长，尽管笔者尽力去学习，但由于水平有限，在内容安排、表述、推导等方面难免会有不当之处，敬请读者在阅读本书的过程中不吝提出宝贵的建议，以期改进本书。

从大数据产生开始，对其基本概念、分析方法、计算平台等方面的争论就一直没有停止过，本书并没有试图回避这些争论，而是在一些地方将争论的不同观点和笔者的观点都列出来，请读者做出自己的判断，也欢迎读者发表自己的观点。读者若有意见和建议，请与笔者联系：wangzh@hit.edu.cn。

王宏志

2017 年 2 月 7 日于哈尔滨

教学建议

教学内容	学习要点及教学要求	课时安排		
		计算机（软件工程）专业本科生	计算机（软件工程）专业研究生	数据科学专业
第 1 章 绪论	了解大数据的基本概念、来源、大数据分析的概念、大数据分析过程中的关键技术与难点	2	2	2
第 2 章 大数据分析模型	掌握大数据分析模型的建立方法，了解大数据分析的基本统计量，掌握推断统计的方法及其实现方法	2	6～10	2
第 3 章 关联分析模型	了解回归分析的基本概念和评估方法，掌握回归分析的建模方法和实现方法，了解关联规则分析和相关分析的基本概念，掌握关联规则分析和相关分析的实现方法	2～4		6
第 4 章 分类分析模型	了解分类分析的定义，了解多种判别分析和机器学习分类的方法，掌握分类模型的建立和实现方法	2～4		6
第 5 章 聚类分析模型	了解聚类分析的定义、分类、评价方法、实现方法和应用，掌握聚类分析模型的建立和实现方法	2		2
第 6 章 结构分析模型	了解结构分析的定义以及最短路径、链接排名、结构计数、结构聚类和社团发现等结构分析模型的建立和实现方法	2		2
第 7 章 文本分析模型	了解文本分析的定义及分词、词频统计、TF-IDF、PLDA、Word2Vec 等文本分析模型的建立和实现方法	2		2（选讲）
第 8 章 大数据分析的数据预处理	了解数据抽样和过滤、数据标准化与归一化、数据清洗等支持大数据分析的数据预处理过程	2～4	4	2
第 9 章 降维	了解特征工程的基本概念和方法，掌握主成分分析、因子分析和压缩感知等主要的降维方法，了解面向神经网络的降维、基于特征散列的维度缩减和基于 Lasso 算法的降维方法	2～4	4	2～4
第 10 章 面向大数据的数据仓库系统	了解数据仓库的基本概念、内涵、基本组成、体系结构和建立方法，了解面向大数据特征的数据仓库系统和内存数据仓库系统	2	4	2

（续）

教学内容	学习要点及教学要求	课时安排		
		计算机（软件工程）专业本科生	计算机（软件工程）专业研究生	数据科学专业
第 11 章 大数据分析算法	了解大数据分析算法的需求和类型，掌握基于 MapReduce 的回归算法、关联规则算法、分类算法和聚类算法等大数据分析算法	2	2	2（选讲）
第 12 章 大数据计算平台	了解 Spark、Hyracks、DPark、HaLoop、MaxCompute 等大数据计算平台的系统结构和其上分析算法的实现方法	2	2	
第 13 章 流式计算平台	了解流式计算平台的定义、应用和发展，了解 Storm、Samza、Cloud Dataflow 等平台的系统结构和其上分析算法的实现方法	2	2	
第 14 章 大图计算平台	了解大图计算的基本概念，了解 GraphLab、Giraph、Neo4j、Apache Hama、MaxCompute Graph 等平台的系统结构和其上分析算法的实现方法	2	2	
第 15 章 社交网络	了解社交网络的建模方法、社交网络的结构，了解社交网络中基于社交网络语义分析的利益冲突发现、社区发现、关联分析、影响力预测等关键技术	2	4	2
第 16 章 推荐系统	了解推荐系统的基本概念以及协同过滤、基于用户评价的推荐、基于“人”的推荐、基于标记的推荐和社交网络中的推荐等关键技术	2	4	2
	教学总学时建议	32 ～ 40	36 ～ 40	30 ～ 36

课堂教学建议：

1）本教材讲授大数据分析的概念、模型、方法、实现和应用，涉及内容较多，基于本教材的教学组织可以根据课程的需求选择重点内容，无须面面俱到。

2）限于篇幅，本教材对一些概念和技术介绍比较简略，笔者已经提供一些相关书籍和文献作为参考，教师在讲授的时候，可以根据需要拓展相关的内容，而不囿于本书中的内容。

3）本书第 12 ～ 14 章介绍了若干系统，可以选取其中一部分作为实验课程内容，每一章中系统的功能类似，可以选取 1 个重点讲授和进行实现方面的实验，也可以选取几个讲授并进行对比。

4）大数据分析领域发展很快，尽管本教材试图加入了一些比较前沿的内容，但是由于写作周期及知识的扩充发展，难以包含最新的内容，所以在本课程的学习过程中，教师可以选取每年大数据分析方面最新的文献加以介绍。

目　录

⊖ 附录见华章网站 www.hzbook.com。——编辑注