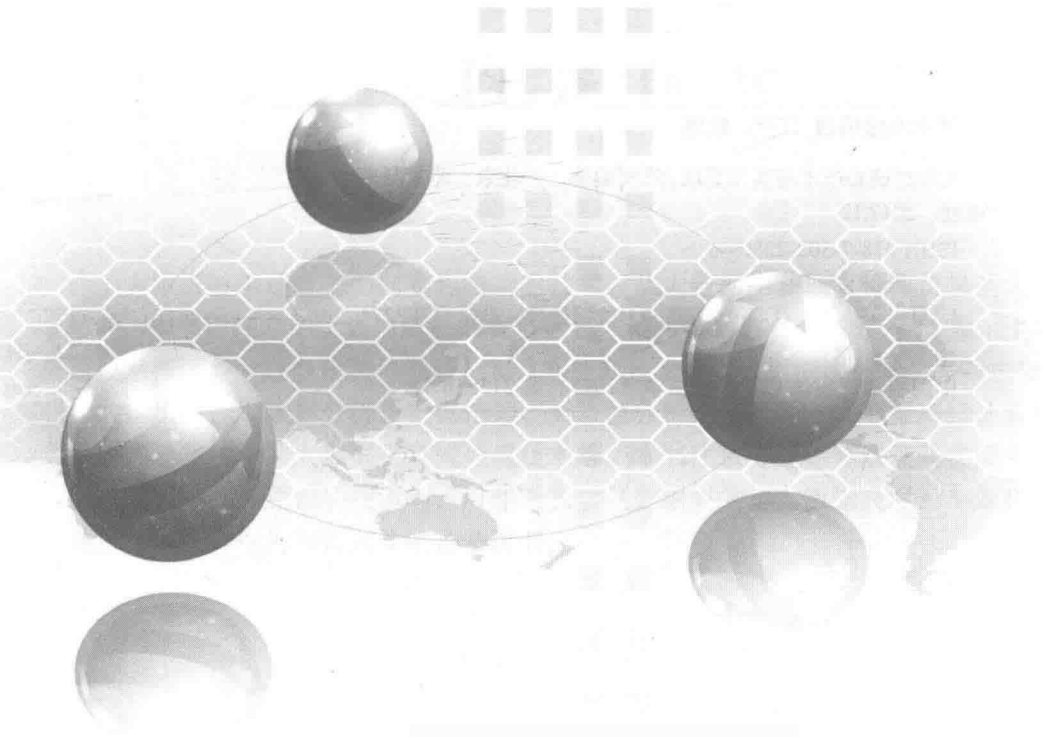
陈 明◎编著

大数据
核心技术与实用算法

- 全面介绍大数据的**核心技术**
- 列举大数据技术的几十个**实用算法**

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社



陈 明◎编著

大数据
核心技术与实用算法

DASHUJU HEXIN JISHU YU SHIYONG SUANFA

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目（CIP）数据

大数据核心技术与实用算法/陈明编著. —北京：北京师范大学出版社，2017.12

ISBN 978-7-303-22809-6

I. ①大… II. ①陈… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 215916 号

营 销 中 心 电 话 010-62978190 62979006
北师大出版社科技与经管分社网 www.jswsbook.com
电 子 信 箱 jswsbook@163.com

出版发行：北京师范大学出版社 www.bnupg.com
北京新街口外大街 19 号
邮政编码：100875

印 刷：北京京师印务有限公司
经 销：全国新华书店
开 本：787 mm×1092 mm 1/16
印 张：31
字 数：702 千字
版 次：2017 年 12 月第 1 版
印 次：2017 年 12 月第 1 次印刷
定 价：62.50 元

策划编辑：李 丹 责任编辑：李 丹
美术编辑：刘 超 装帧设计：刘 超
责任校对：赵非非 责任印制：赵非非

内 容 简 介

大数据技术是一个面向实际应用的技术。从大数据中获取有价值信息是大数据技术的精髓。本书详细介绍了当前流行的大数据技术的主要内容，全书分为12章，主要包括大数据技术概述、大数据获取与存储、大数据抽取技术、大数据清洗技术、大数据转换与约简、大数据集成、大数据分析、数据挖掘、大数据分析结果解释、大数据离线计算技术、大数据流式计算技术、大数据交互式处理技术等。本书注重方法、算法及其实现介绍，实例丰富、语言精练、逻辑层次清晰，可作为大学《数据科学与大数据技术》专业和相近专业的教材，也可以作为科技人员的参考用书。

前　言

大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界寄予无比的厚望。大数据成为继20世纪末21世纪初互联网蓬勃发展以来的又一轮IT工业革命。

大数据技术是指从数据采集、清洗、集成、挖掘、分析与结果解释，进而从各种类型的巨量数据中快速获得有价值信息的全部技术。从数据挖掘和分析中，人们可以获取具有重要价值的信息。大数据技术的精髓是从大数据中产生新见解的能力、识别复杂关系和做出越来越精准的预测、从大数据中产生动力、获取知识和采取行动的能力。

大数据技术是现代科学与技术发展，尤其是计算机科学技术发展的重要成果和结晶，是科学发展史的又一个新的里程碑。大数据的出现对计算机科学技术的许多领域提出了挑战与冲击，推动了IT技术的发展。

大数据技术的出现凝集了多学科的研究成果，是一门多学科的交叉融合技术，随着科学技术的发展，大数据技术发展更为迅速，应用更为深入与广泛，并凸显其巨大潜力和应用价值。

本书系统地介绍了大数据技术的核心内容，全部内容说明如下。

第1章为大数据技术概述，主要包括大数据技术框架、特征、大数据的冲击、大数据研究方法论、常用的大数据计算框架。第2章为大数据获取与存储，主要包括数据获取的定义与数据获取领域、NewSQL和NoSQL、分布式文件系统、虚拟存储技术、云存储技术、分布式存储核心算法、数据仓库与数据集市、区块链技术等。第3章为大数据抽取技术，主要包括数据抽取技术概述、增量数据抽取技术、非结构化数据抽取、Web数据抽取。第4章为大数据清洗技术，主要包括数据质量与数据清洗、不完整数据的清洗方法、异常数据清洗方法、重复数据清洗方法、文本清洗和数据清洗技术的实现。第5章为大数据转换与约简，主要包括数据平滑、数据规范化、数据泛化、数据约简、数据压缩、数值约简、数值数据离散化与概念分层。第6章为大数据集成技术，主要包括数据集成技术概述、数据迁移、数据集成模式、数据集成系统、数据集成系统的构建、数据聚类集成、实时数据集成、企业信息集成。第7章为大数据分析技术，主要包括大数据分析定义与方法、数据分析的基本方法、高级数据分析方法、预测分析、大数据预测分析的应用趋势。第8章为数据挖掘，主要包括数据挖掘理论基础、关联规则挖掘、分类、聚类方法、序列模式挖掘、Web挖掘技术、空间数据挖掘、非结构化文本数据挖掘。第9章为大数据分析结果解释，主要包括数据分析结果的解释、数据的基本展现方式、大数据可视化、大数据可视分析。第10章为大数据离线计算技术，主要包括数据离线计算概述、MapReduce的体系结构、Hadoop分布式计算平台、MapReduce程序设计实例。第11章为大数据流式计算技术，主

要包括流式数据的概念与特征、大数据的流式计算模式、数据流处理、流式计算的应用场景、流式计算的系统架构、高可用技术、Storm 流式处理平台、单词计数 topology 程序设计。第 12 章为大数据交互式处理技术，主要包括交互式处理系统的问题、数据切分、数据钻取、Scala 编程语言简介、交互式大数据处理框架 Spark、交互式查询。

本书在结构上为积木状，各章内容独立地、概念性论述。出于篇幅考虑，书中所提及定理没有给出证明，如需要，可以查阅相关文献。由于作者水平有限，书中不足之处在所难免，敬请读者批评指正。

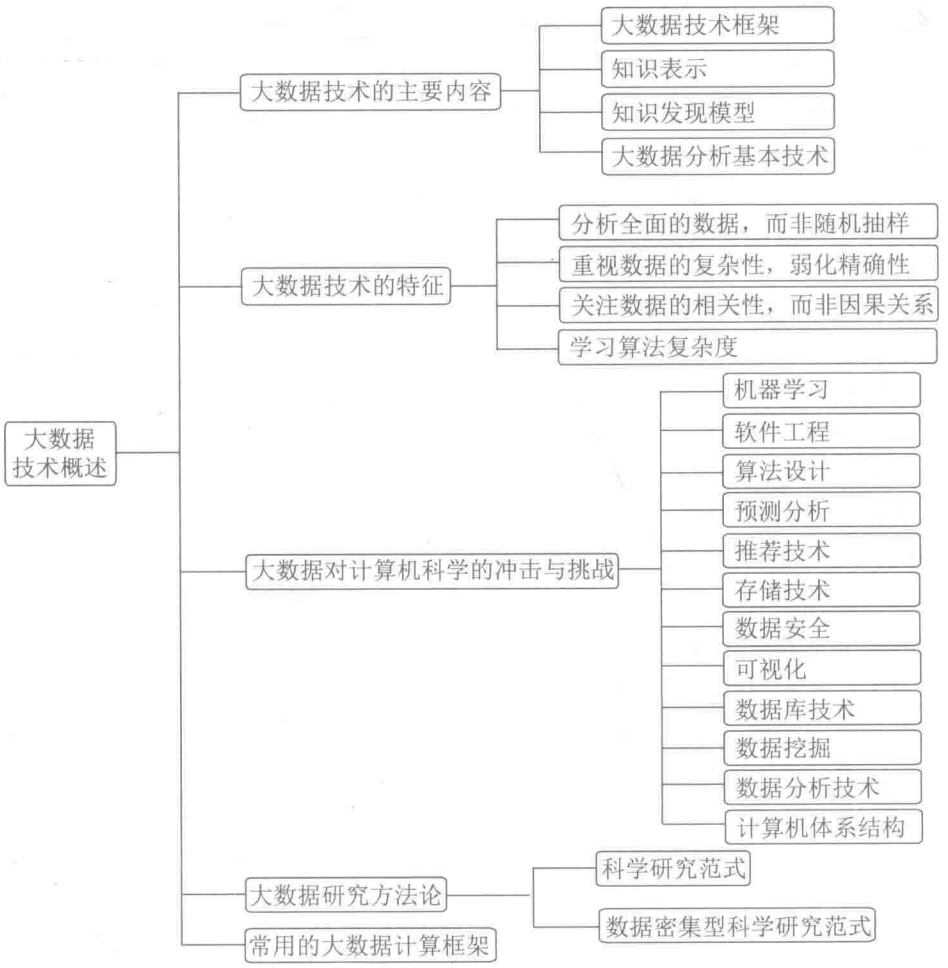
陈明

2017 年 2 月

目　录

第 1 章　大数据技术概述

本章主要内容



数据科学是关于数据的科学，定义为研究探索网络空间（Cyberspace）中数据界奥秘的理论、方法和技术。数据科学主要有两个内涵：一个是研究数据本身；另一个是为自然科学和社会科学研究提供一种科学研究的数据新方法。

1.1 大数据技术的主要内容

大数据领域主要包括大数据工程、大数据技术和大数据应用等分支领域。大数据工程是指大数据的规划建设与运营管理的系统工程；大数据技术是指通过数据获取、清洗、集成、挖掘和分析，进而从各种类型的巨量数据中快速获得有价值信息的全部技术；大数据应用是指应用大数据技术解决某一实际问题的方法与技术。基于解决问题的角度考虑，更受关注的是大数据技术与大数据应用。

在提到大数据时呈现出双重含义，一是出现了具有大数据，二是大数据技术。研究大数据技术的目的是发展大数据技术并将其应用到相关领域，通过解决大数据问题来促进突破性发展。因此，大数据带来的挑战不仅体现在如何处理大数据，并从中获取有价值的信息，也体现在如何加强大数据技术研发，抢占技术发展的前沿。

1.1.1 大数据技术框架

由于全球数据量增加遵循摩尔定律，每 18 个月翻一番，致使数据规模迅速增大，超过了当前计算机系统的存储能力与处理能力。由于大数据的 5V 特性，对处理能力的要求提高。因此，数据处理能力已成为核心竞争力。为了解决这类问题，需要多学科结合，需要研究新型数据处理的科学方法，以便在数据多样性和不确定性前提下进行数据规律和统计特征的研究。应用工具将分布的异构数据源中的数据，例如，关系数据、空间数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

一般说来，大数据处理的过程可以概括为 5 个步骤，分别是数据获取与存储、抽取与清洗，数据集成与聚集，数据分析与建模，结果解释。大数据技术框架如图 1-1 所示。

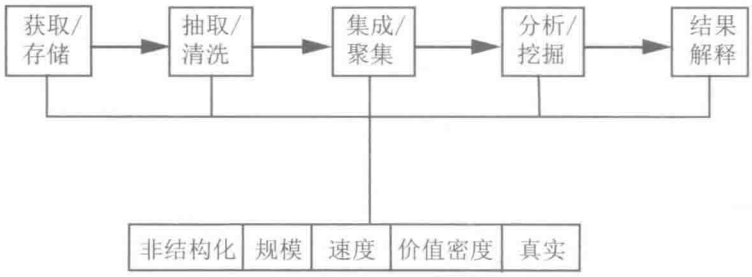


图 1-1 大数据技术框架

通过上述 5 个步骤，可以将获取的数据转换为有价值的信息，在每个阶段，需要应对大数据的 5V 特征。

1. 数据获取与存储

大数据的获取是指利用各种数据库来接收发自 Web、APP 或者传感器等数据源的数据，

并且用户可以通过这些数据库来进行简单的查询和处理工作。在大数据的获取过程中，其主要特点是并发率高，数据量巨大，因为可能有成千上万的用户同时进行访问和操作数据库系统。

2. 数据抽取与清洗

虽然在数据获取端设置大量数据库系统，但是如果要对这些数据进行有效的分析，还是应该将这些来自前端的数据抽取到一个大型分布式数据库中，或者分布式存储集群中，并且可以在抽取基础上完成数据清洗等一系列预处理工作。也有一些用户在抽取时对数据进行流式计算，来满足部分业务的实时计算需求。大数据抽取、清洗与标记过程的主要特点是抽取的数据量大，每秒的抽取数据量可达到百兆数量级，甚至千兆数量级。

3. 数据集成与聚集

数据集成技术的任务是将相互关联的分布式异构数据源集成到一起，使用户能够以透明的方式访问这些数据源。在这里，集成是指维护数据源整体上的数据一致性，提高信息共享利用的效率，透明方式是指用户不必关心如何对异构数据源进行访问，只关心用何种方式访问何种数据库。

前三步称为预处理过程，预处理过程的工作量一般可能占到全过程的 70%。

4. 数据分析与挖掘

利用分布式数据库或者分布式计算集群来对存储其内的大数据进行分析和分类汇总等，以满足分析需求。分析方法主要包括：假设检验、显著性检验、差异分析、相关分析、T 检验、方差分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、判别分析、对应分析、多元对应分析等。

在这方面，对于一些实时性需求将用到流式计算或交互计算等，而对于批处理，或者基于非结构化数据的需求可以使用离线处理技术。数据分析中涉及的数据量巨大，对系统资源，特别是 I/O 资源占用极大。

在数据分析与挖掘中，数据挖掘完成的是高级数据分析的需求，一般没有预先设定的主题，主要是对现有数据进行基于各种算法的计算，起到预测的效果。数据挖掘主要进行分类、估计、预测、相关性分组或关联规则、聚类、描述和可视化、复杂数据类型挖掘等。比较典型算法有 Kmeans 聚类算法、深度学习算法、SVM 统计学习算法和 NaiveBayes 分类算法，主要使用的工具有 Hadoop 的 Mahout 等。该过程的主要特点是挖掘的算法复杂，并且计算所涉及的数据量和计算量巨大。

应用选定的数据挖掘算法，从数据中提取出用户所需要的知识，需要用一种特定的方式表示。数据挖掘选择主要有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的、容易理解的知识，而有的用户只是希望获取预测准确度尽可能高的预测型知识，并不注重获取的知识是否易于理解。

数据挖掘阶段发现的模式，经过评估，可能存在冗余或无关的模式，这时需要将其删除。也有可能模式不满足用户要求，这时则需要整个发现过程回退到前续阶段，重新选取数据、采用新的数据变换方法、设定新的参数值，甚至更换算法等。建模的主要内容是构建预测模型、机器学习模型和建模仿真等。

5. 结果解释

由于知识发现最终是面向人类用户，因此需要对发现的模式进行可视化，或者把结果转换为用户易于理解的表示。也就是说，仅能够分析大数据，但却无法使得用户理解分析的结果，这样的效果价值不大。如果用户无法理解分析，那么需要决策者对数据分析结果进行解释。解释通常包括检查所提出的假设并对分析过程进行追踪，此外，在分析过程中，可能引入许多可能的误差来源，计算机系统可能存在缺陷。模型采用可视化展现大数据分析结果，例如，利用云计算、标签云、关系图等呈现。知识评估是知识发现的一个重要的必不可缺少的工作，不仅担负着将数据分析系统发现的知识以用户能了解的方式呈现的任务，而且要根据需要进行知识评价，如果没有达到用户的目标，就需要返回前面相应的步骤进行螺旋式处理以最终获得满意的结果。

大数据处理至少应该经过上述 5 个基本步骤，才能成为一个完整的大数据处理过程。

1.1.2 知识表示

数据挖掘的目的是发现知识，知识表示引人关注，下面介绍几种常用的知识表示模式，主要包括广义知识、关联知识、类知识、预测型知识和特异型知识。

1. 广义知识

广义知识是描述类别特征的概括性知识。在数据源（如数据库等）中存放的是细节性的具体数据，但是需要从高层视图处理、观察这些数据，这就可以通过对数据进行不同层次的泛化来寻找数据所蕴含的概念或逻辑，以适应数据分析的要求。数据挖掘的目的之一就是根据数据微观特性发现有普适性的、高层次概念的中观和宏观的知识。因此，数据挖掘过程是对数据所蕴含的概念特征信息、汇总信息和比较信息等的概括、精练和抽象的过程。挖掘出的广义知识可以应用可视化技术，以直观的图表，如饼图、柱状图、曲线图、立方体等形式展示给用户，也可以作为其他应用，如分类、预测等的基础知识。

（1）概念

概念描述是对某类对象特征的概括，概念描述可分为特征性描述和区别性描述。概念描述是描述某类对象的共同特征、描述异类对象之间的区别。概念归纳是概念描述的最典型方法。

（2）多维数据存储

数据聚集是数据分析的基本工作，如计数、求和、平均、最大值等。由于很多聚集函数需经常重复计算，而且这类操作的计算量一般又特别大，因此，可以将这些汇总的操作结果预先计算并存储起来，以便于高级数据分析使用。最流行的存储汇集数据的方法是多

维数据库技术。

（3）多层次概念

由数据归纳出的各层次的概念是对原始数据的各种粒度的抽象。概念分层技术就是低层概念集到高层概念集的映射技术。在任何形式的源数据组织形式下，存储的细节数据总是作用在一个特定的范畴内，更高层次的数据综合与分析是进行决策的基础。

2. 关联知识

关联知识反映事件之间的关联，对其挖掘的目的就是找出数据库中隐藏的关联信息。数据之间存在的关联不仅是在数据模型中的关联，大部分是蕴藏的关联。关联可分为简单关联、时序关联、因果关联、数量关联等。关联事先不明确，而是通过数据的关联分析获得。

关联规则挖掘是关联知识发现的最常用方法，最为著名的是 Apriori 及其改进算法。为了发现有意义的关联规则，设定两个阈值：最小支持度和最小可信度。挖掘出的关联规则必须满足用户规定的最小支持度，它表示了一组项目关联在一起需要满足的最低联系程度。挖掘出的关联规则也必须满足用户规定的最小可信度，最小可信度标志了一个关联规则的最低可靠度。

3. 类知识

类知识刻画了一类事物与其他类事物区别的某种共同特征，主要分为分类知识和聚类知识。

（1）分类

分类是一个经典的技术，分类的主要工作是构建分类模型，该模型能把数据库中的数据项映射到给定类别中。需要有一个训练样本数据集作为输入。分类知识也必须来自于源数据，应该通过对源数据的过滤、抽取（抽样）、压缩以及概念提取等而获得。从机器学习的角度，分类技术是一种有指导的学习，即每个训练样本的数据对象已经有类标识，通过学习可以形成表达数据对象与类标识间对应的知识。数据挖掘的目标就是根据样本数据形成的类知识并对源数据进行分类，进而也可以预测非训练样本数据、未来数据的归类。常用决策树、贝叶斯分类、神经网络、遗传算法、类比学习等学习方法。

（2）聚类

聚类是把一组个体按照相似性归成若干类别，其目的是使得属于同一类别的个体之间的差别尽可能小，而不同类别上的个体间的差别尽可能大。通过聚类可以对源数据划分为一系列有意义的子集，进而实现数据分析。聚类和分类不同，前者总是在特定的类标识下寻求新元素属于哪个类，而后者则是通过对数据的分析比较生成新的类标识。聚类分析生成的类标识描述了数据所蕴含的类知识，聚类技术是一种无指导的学习。

4. 预测型知识

预测型知识是指由历史数据和当前数据而产生的能够推测未来趋势的知识。这类知识可以是以时间为关键属性的关联知识，这时可以应用到以时间为关键属性的源数据挖掘。