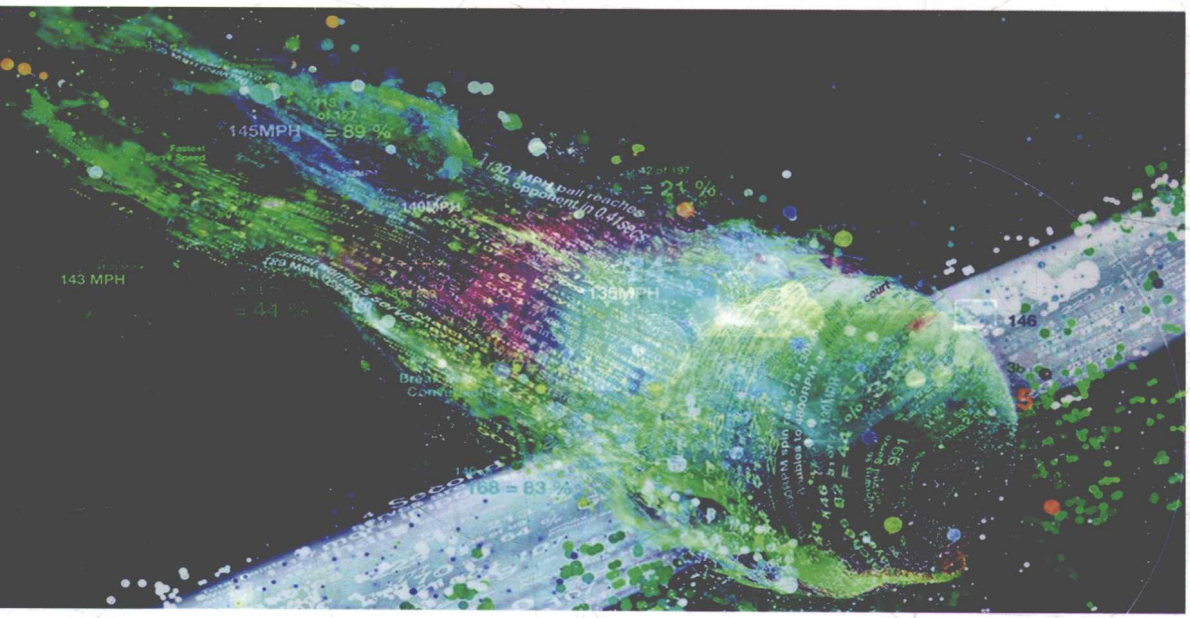
高级大数据人才培养丛书

数据挖掘

BIG DATA

刘 鹏 张 燕 ◎ 丛书主编 王朝霞 ◎ 主编



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

高级大数据人才培养丛书

数据挖掘

丛书主编： 刘　鹏　张　燕
主　　编： 王朝霞

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

2017 年年初，在中国信息协会大数据分会副会长刘鹏教授的倡议和组织下，全国上百家高校从事一线教学科研任务的教师，开展了高级大数据人才培养丛书的编撰工作。本书是丛书之一，其定位是大数据挖掘技术与应用。本书系统地介绍了数据挖掘算法理论与方法、工具和应用，包括经典数据挖掘算法，大数据环境下常用数据挖掘算法的优化，大数据新常态下催生的数据分析方法（如推荐系统、链接分析与网页排序、互联网信息抽取、日志挖掘与查询分析）、工具与应用。

本书适合作为相关专业本科生和研究生教材。高职高专学校也可以选用部分内容开展教学。本书也很适合作为大数据分析研发人员的自学书籍。

图书在版编目（CIP）数据

数据挖掘 / 王朝霞主编. —北京：电子工业出版社，2018.3
（高级大数据人才培养丛书）
ISBN 978-7-121-33531-0

Ⅰ. ①数… Ⅱ. ①王… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 012619 号

策划编辑：董亚峰
责任编辑：董亚峰
特约编辑：穆丽丽
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×1092　1/16　印张：21.75　字数：529 千字
版　　次：2018 年 3 月第 1 版
印　　次：2018 年 3 月第 1 次印刷
定　　价：58.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254694。

编 写 组

丛书主编：刘 鹏　　张 燕

主　　编：王朝霞

副 主 编：施建强　　杨慧娟　　陈建彪

编　　者：（按姓氏首字母排序）

曹 洁　　宁亚辉　　王伟嘉

袁晓东　　张卫明

基金支持：

金陵科技学院高层次人才科研启动基金（40610186）

国家自然科学基金（61472005）

江苏高校软件工程品牌专业建设工程（PPZY2015B140）

军队科研计划项目（AS214R002）

中国博士后基金（2014M562589）

江苏省高校自然科学研究重大项目（16KJA520003）

总 序

短短几年间，大数据就以一日千里的发展速度，快速实现了从概念到落地，直接带动了相关产业井喷式发展。全球多家研究机构统计数据显示，大数据产业将迎来发展黄金期：IDC 预计，大数据和分析市场将从 2016 年的 1300 亿美元增长到 2020 年的 2030 亿美元以上；中国报告大厅发布的大数据行业报告数据也说明，自 2017 年起，我国大数据产业迎来发展黄金期，未来 2~3 年的市场规模增长率将保持在 35%左右。

数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才问题的凸显。麦肯锡预测，每年数据科学专业的应届毕业生将增加 7%，然而仅高质量项目对于专业数据科学家的需求每年就会增加 12%，完全供不应求。根据《人民日报》的报道，未来 3~5 年，中国需要 180 万数据人才，但目前只有约 30 万人，人才缺口达到 150 万之多。

以贵州大学为例，其首届大数据专业研究生就业率就达到 100%，可以说“一抢而空”。急切的人才需求直接催热了大数据专业，国家教育部正式设立“数据科学与大数据技术”本科新专业。目前已经有两批共计 35 所大学获批，包括北京大学、中南大学、对外经济贸易大学、中国人民大学、北京邮电大学、复旦大学等。估计 2018 年会有几百所高校获批。

不过，就目前而言，在大数据人才培养和大数据课程建设方面，大部分高校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，尚未形成完善的大数据人才培养和课程体系，院校缺“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学科研工作缺少“原材料”。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材也正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大搜索引擎“云计算”关键词中排名第一），2010 年出版了《云计算（第 1 版）》、2011 年出版了《云计算（第 2 版）》、2015 年出版了《云计算（第 3 版）》，每一版都花费了大量成本制作并免费分享对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百

万次之多。同时，《云计算》一书也成为国内高校的首选教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。除了资料分享，在 2010 年，我也在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长等。

近几年，面对日益突出的大数据发展难题，我也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中位居第一；为了解决大数据师资匮乏的问题，我面向全国院校陆续举办多期大数据师资培训班。2016 年年末至今，在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。2017 年 5 月，还举办了全国千所高校大数据师资免费讲习班，盛况空前。

其中，为了解决大数据实验难的问题而开发的大数据实验平台，正在为越来越多高校的教学科研带去方便：我带领云创大数据（www.cstor.cn，股票代码：835305）的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册（包含 85 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校成功应用，并广受校方好评。该平台也以云服务的方式在线提供（大数据实验平台，https://bd.cstor.cn），帮助师生通过自学，用一个月左右成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

同时，为了解决缺乏权威大数据教材的问题，我所负责的南京大数据研究院，联合金陵科技学院、河南大学、云创大数据、中国地震局等多家单位，历时两年，编著出版了适合本科教学的《大数据》《大数据库》《大数据实验手册》等教材。另外，《数据挖掘》《大数据可视化》《深度学习》《虚拟化与容器》《Python 语言》等本科教材也将于近期出版。在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职、高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们还联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据

系统运维》《大数据实践》系列教材，目前也已经陆续进入定稿出版阶段。

此外，我们也将继续在中国大数据（thebigdata.cn）和中国云计算（chinacloud.cn）等网站免费提供配套 PPT 和其他资料。同时，持续开放大数据实验平台（https://bd.cstor.cn）、免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn），使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第 7 版，与时俱进且日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘　鹏

于南京大数据研究院

前 言

21 世纪初，人类迈入大数据时代，各行各业拥抱大数据，希冀借大数据挖掘与分析来促进产业升级与变革。因此，大数据人才的需求呈现井喷之势。

中国信息协会大数据分会副会长刘鹏教授顺势而为，周密思考，提出高级大数据人才培养课程体系，并邀请全国上百家高校中从事一线教学科研任务的教师一起，编撰高级大数据人才培养丛书。本书即该套丛书之一。

本书的定位是大数据挖掘技术与应用。以“让学习变得轻松”为根本出发点，本书努力回答：数据挖掘是什么？发展如何？经典的数据挖掘算法有哪些？大数据环境下数据挖掘有哪些新特点和新延展？如何分析实际问题，如何应用？本书编写的指导思想有三：一是理论与应用相呼应。从数据挖掘算法理论与方法、工具和应用两方面进行阐述，既注重理论，同时贴近实战，解行结合，希望学习者既能很快将理论应用于实际领域的数据分析中，同时也具备厚积薄发的能力；二是基础与发展一脉相承。大数据新常态下经典数据挖掘的基本原理仍然适用，不同之处在于，根据现有分布式、并行环境，对原有算法进行优化。本书循序渐进地介绍经典数据挖掘算法，以及大数据环境下数据挖掘算法的新特点和新延展，有助于学习者全面掌握数据挖掘理论；三是局部与全局整体联动。本书属于高级大数据人才培养丛书系列教材，因此，在本书内容组织上，需要考虑与丛书其他教材的关系，既紧密联系又自成一体，共同组成高级大数据人才培养课程体系，例如，考虑到丛书中《深度学习》一书对神经网络技术有详尽的介绍，故本书没有再介绍神经网络相关技术。

基于上述指导思想，本书内容分为四部分：一是概念与基础（第 1、2 章）；二是经典的数据挖掘算法（第 3～6 章）；三是大数据挖掘技术，其中，第 7 章重点介绍了大数据环境下经典数据挖掘算法的优化与改进，第 8 章介绍了推荐系统的理论与方法，第 9 章则对链接分析与网页排序、互联网信息抽取、日志挖掘与查询分析等技术进行了介绍；四是常用数据挖掘工具（包），见附录。

本书成稿过程中得到了丛书总主编刘鹏教授和金陵科技学院张燕副校长的大力支持，在书稿提纲和内容组织上提出了诸多建设性意见。同时，两轮审稿评审专家对本书给予了全面指导和帮助，在此一并致谢。

当前，大数据挖掘技术仍处在高速发展的历史阶段，其概念内涵、技术方法、应用模式还在不断创新演化之中，由于时间和水平所限，本书还存在缺点和不足，欢迎大家不吝赐教。

王朝霞

于重庆市陆军勤务学院

目 录

第1章 绪 论

计算机技术、数据库技术和传感器技术的飞速发展，使人们获取数据和存储数据变得越来越容易。社会信息化水平的不断提高和数据库应用的日益普及，使人类积累的数据量正在以指数方式增长。与日趋成熟的数据管理技术与软件工具相比，数据分析技术与工具所提供的功能，却无法有效地为决策者提供其决策支持所需的有效知识，从而形成了一种“丰富的数据，贫乏的知识”的现象。为有效解决这一问题，自 20 世纪 80 年代开始，数据挖掘技术逐步发展起来，人们迫切希望能对海量数据进行更加深入的分析，发现并提取隐藏在其中的有价值信息，以便更好地利用这些数据。数据挖掘技术的迅速发展，得益于目前全世界所拥有的巨大数据资源，以及对其中有价值的信息和知识的巨大需求。在这种背景下，数据挖掘的理论和方法获得了飞速的发展，其技术和工具已经广泛应用到互联网、金融、电商、管理、生产、决策等各个领域。

1.1 数据挖掘概述

1.1.1 数据挖掘的概念

数据挖掘（Data Mining，DM），是从大量的、有噪声的、不完全的、模糊和随机的数据中，提取出隐含在其中的、人们事先不知道的、具有潜在利用价值的信息和知识的过程[1]。这个定义包括几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求放之四海皆准的知识，仅支持特定的发现问题。所提取到的知识的表示形式可以是概念、规律、规则与模式等。数据挖掘能够对将来的趋势和行为进行预测，从而帮助决策者做出科学和合理的决策。比如，通过对公司数据库系统的分析，数据挖掘可以回答诸如“哪些客户最有可能购买我们公司的什么产品？”“客户有哪些常见的消费模式和消费习惯？”等类似问题。

与数据挖掘相似的概念是知识发现（Knowledge Discovery in Databases，KDD），知识发现是指用数据库管理系统来存储数据、用机器学习方法来分析数据、挖掘大量数据背后隐藏的知识的过程。数据挖掘是整个知识发现流程中的一个具体步骤，也是知识发现过程中最重要的核心步骤。

数据挖掘是一个交叉学科，涉及数据库技术、人工智能、数理统计、机器学习、模式识别、高性能计算、知识工程、神经网络、信息检索、信息的可视化等众多领域，其中数据库技术、机器学习、统计学对数据挖掘的影响最大。对数据挖掘而言，数据库为其提供数据管理技术，机器学习和统计学提供数据分析技术[2]。数据挖掘所采用的算

法，一部分是机器学习的理论和方法，如神经网络、决策树等；另一部分是基于统计学习理论，如支持向量机、分类回归树和关联分析等。但传统的机器学习和统计学研究往往并不把海量数据作为处理对象，因此数据挖掘要把这两类技术用于海量数据中的知识发现，需要对算法进行改造，使得算法性能和空间占用达到实用的地步。

常见的数据挖掘对象有以下七大类：

（1）关系型数据库、事务型数据库、面向对象的数据库。

（2）数据仓库/多维数据库。

（3）空间数据（如地图信息）。

（4）工程数据（如建筑、集成电路信息）。

（5）文本和多媒体数据（如文本、图像、音频、视频数据）。

（6）时间相关的数据（如历史数据或股票交换数据）。

（7）万维网（如半结构化的 HTML、结构化的 XML 以及其他网络信息）。

1.1.2 大数据环境下的数据挖掘

继互联网、物联网、云计算的不断发展及智能终端的普及，海量复杂多样的数据呈现出爆炸式的增长，标记着“大数据”时代的到来。作为重要的生产因素，大数据已成为蕴含巨大潜在价值的战略资产，推动着产业升级和崛起，影响着科学思维与研究方法的变革。然而，大数据在依托其丰富的资源储备和借助强大的计算技术发挥优势的同时，也带来了极大的挑战。海量、动态及不确定的数据使得传统数据处理系统面临着存储和计算瓶颈，同时，就如何从复杂的大数据中实时快速地挖掘出有价值的信息和知识，传统的数据挖掘技术自身受限的功能已无法满足用户的需求。因此，大数据环境下需要一种适用技术，即“大数据挖掘”，来应对面临的挑战[3]。

大数据挖掘是从体量巨大、类型多样、动态快速流转及价值密度低的大数据中挖掘有巨大潜在价值的信息和知识，并以服务的形式提供给用户。与传统数据挖掘相比，大数据挖掘同样是以挖掘有价值的信息和知识为目的，然而就技术发展背景、所面临的数据环境及挖掘的广度深度而言，两者存在很多差异：

1. 技术背景差异

传统数据挖掘在数据库、数据仓库及互联网发展等背景下，实现了从独立、横向到纵向数据挖掘的发展。而大数据挖掘是在大数据背景下得益于云计算、物联网、移动智能终端等技术产生与发展，具备了充实环境技术条件，基于云计算等相关技术集成实现海量数据的挖掘。

2. 处理对象的差异

传统数据挖掘的数据来源主要是以某个特定范围的管理信息系统被动数据的产生为主，外加少数的 Web 信息系统中由用户产生的主动数据，数据类型以结构化数据为主，外加少量的半结构化或非结构化数据。相比于传统数据挖掘，大数据挖掘的数据来源更广、体量巨大、类型更加复杂；采集方式不再局限于被动，采集范围更为全面，吞吐量高，处理实时且快速，但由于对数据的精确度要求不高致使数据的冗余度和不确定