“十二五”国家重点图书出版规划项目

大数据技术与应用

丛书主编
朱扬勇 吴俊伟

Big Data
Technology and Application Series

熊赟 朱扬勇 陈志渊
编著

大数据挖掘

上海科学技术出版社

大数据技术与应用

大数据挖掘

熊 赟 朱扬勇 陈志渊
编著

上海科学技术出版社

图书在版编目(CIP)数据

大数据挖掘 / 熊赟,朱扬勇,陈志渊编著.
—上海：上海科学技术出版社,2016.4
(大数据技术与应用)
ISBN 978-7-5478-2961-5

Ⅰ.①大… Ⅱ.①熊… ②朱… ③陈… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 010840 号

大数据挖掘

熊　赟　朱扬勇　陈志渊　编著

上海世纪出版股份有限公司
上海科学技术出版社　出版
(上海钦州南路 71 号　邮政编码 200235)
上海世纪出版股份有限公司发行中心发行
200001　上海福建中路 193 号　www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092　1/16　印张 20
字数 420 千字
2016 年 4 月第 1 版　2016 年 4 月第 1 次印刷
ISBN 978-7-5478-2961-5/TP·40
定价：75.00 元

内容提要

本书系统介绍了大数据挖掘的概念、原理、技术和应用，具体内容包括：认识和理解大数据；大数据挖掘需要的相关技术（大数据获取技术、大数据存储管理技术和大数据可视化技术等）；大数据计算框架；大数据挖掘任务（关联分析、聚类分析、分类分析、异常分析、特异群组挖掘和演变分析）；大数据应用实现。

本书对大数据挖掘技术进行了全面而细致的定义和归纳，并向读者展现了该领域最新研究热点和技术。

本书可供数据科学专业的高等学校学生及教师，从事数据领域工作的研究人员、技术人员、管理人员和决策人员参考阅读。

本书相关研究工作得到了以下项目的支持资助：
21 世纪复旦大学研究生用书
国家自然科学基金重点项目(71331005,大数据环境下的管理决策创新研究)

大数据技术与应用

学术顾问

大数据技术与应用

编撰委员会

丛书序

我国各级政府非常重视大数据的科研和产业发展，2014 年国务院政府工作报告中明确指出要“以创新支撑和引领经济结构优化升级”，并提出“设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展”。2015 年 8 月 31 日，国务院印发了《促进大数据发展行动纲要》，明确提出将全面推进我国大数据发展和应用，加快建设数据强国。前不久，党的十八届五中全会公报提出要实施“国家大数据战略”，这是大数据第一次写入党的全会决议，标志着大数据战略正式上升为国家战略。

上海的大数据研究与发展在国内起步较早。上海市科学技术委员会于 2012 年开始布局，并组织力量开展大数据三年行动计划的调研和编制工作，于 2013 年 7 月 12 日率先发布了《上海推进大数据研究与发展三年行动计划（2013—2015 年）》，又称“汇计划”，寓意“汇数据、汇技术、汇人才”和“数据‘汇’聚、百川入‘海’”的文化内涵。

“汇计划”围绕“发展数据产业，服务智慧城市”的指导思想，对上海大数据研究与发展做了顶层设计，包括大数据理论研究、关键技术突破、重要产品开发、公共服务平台建设、行业应用、产业模式和模式创新等大数据研究与发展的各个方面。近两年来，“汇计划”针对城市交通、医疗健康、食品安全、公共安全等大型城市中的重大民生问题，逐步建立了大数据公共服务平台，惠及民生。一批新型大数据算法，特别是实时数据库、内存计算平台在国内独树一帜，有企业因此获得了数百万美元的投资。

为确保行动计划的实施，着力营造大数据创新生态，“上海大数据产业技术创新战略联盟”（以下简称“联盟”）于 2013 年 7 月成立。截至 2015 年 8 月底，联盟共有 108 家成员单位，既有从事各类数据应用与服务的企业，也有行业协会和专业学会、高校和研究院所、大数据技术和产品装备研发企业，更有大数据领域投资机构、产业园区、非 IT

领域的数据资源拥有单位，显现出强大的吸引力，勾勒出上海数据产业的良好生态。同时，依托复旦大学筹建成立了“上海市数据科学重点实验室”，开展数据科学和大数据理论基础研究、建设数据科学学科和开展人才培养、解决大数据发展中的基础科学问题和技术问题、开展大数据发展战略咨询等工作。

在“汇计划”引领下，由联盟、上海市数据科学重点实验室、上海产业技术研究院和上海科学技术出版社于2014年初共同策划了“大数据技术与应用”丛书。本丛书第一批已于2015年初上市，包括了《汇计划在行动》《大数据测评》《数据密集型计算和模型》《城市发展的数据逻辑》《智慧城市大数据》《金融大数据》《城市交通大数据》《医疗大数据》共八册，在业界取得了广泛的好评。今年进一步联合北京中关村大数据产业联盟共同策划本丛书第二批，包括《大数据挖掘》《制造业大数据》《航运大数据》《海洋大数据》《能源大数据》《大数据治理与服务》等。从大数据的共性技术概念、主要前沿技术研究和当前的成功应用领域等方面向读者做了阐述，作者希望把上海在大数据领域技术研究的成果和应用成功案例分享给大家，希望读者能从中获得有益启示并共同探讨。第三批的书目也已在策划、编写中，作者将与大家分享更多的技术与应用。

大数据对科学研究、经济建设、社会发展和文化生活等各个领域正在产生革命性的影响。上海希望通过“汇计划”的实施，同时也是本丛书希望带给大家一个理念：大数据所带来的变革，让公众能享受到更个性化的医疗服务、更便利的出行、更放心的食品，以及在互联网、金融等领域创造新型商业模式，让老百姓享受到科技带来的美好生活，促进经济结构调整和产业转型。

上海市科学技术委员会副主任

2015年11月

前 言

数据挖掘已经有20多年历史了，20年前，“尿布和啤酒的故事”像童话一样被许多应用领域的信息主管认为是不靠谱的幻想（很多地方称为营销神话）。如今，大数据来了，却有很多人将该故事当作大数据的典型案例来讲。“尿布和啤酒的故事”是说“通过对所有购物单进行数据挖掘，发现尿布和啤酒经常被同时购买”，数据挖掘告诉我们一种关联现象，而不是因果。

大数据来了，之前众多基于内存的数据挖掘算法不能再使用，于是需要新的能够运行在大数据计算架构上的数据挖掘算法。另一方面，数据来源和类型也更复杂了，而之前的数据挖掘算法主要是针对单一数据类型的，于是需要新的数据挖掘算法——能够分析复杂数据类型的数据挖掘算法。这些应该是大数据挖掘的主要内容。本书取名《大数据挖掘》，主要介绍了两件事情：传统的数据挖掘算法及其面向大数据的改进；面向复杂数据集的数据挖掘算法，如异质网络挖掘。

本书各章内容如下：第1章介绍大数据挖掘的基本概念、大数据挖掘的任务，以及与大数据挖掘相关的技术。第2章介绍大数据计算框架。第3章到第8章介绍数据挖掘的主要任务，包括关联分析、聚类分析、分类分析、异常分析、特异群组挖掘、演变分析，介绍了这些算法面对大数据的改进。其中，特异群组挖掘是一类典型的面向大数据高价值、低密度特征的数据挖掘任务。第9章介绍了异质数据网络挖掘，这是针对大数据的复杂性而设计的一类新的大数据挖掘方法。第10章以推荐系统为例，介绍了大数据挖掘的应用。第11章对大数据隐私技术进行了介绍。

大数据挖掘是新问题、新应用，作为应用层，算法和软件设计都依赖于计算机基础架构、计算框架和数据管理系统。大数据目前还是一个大问题，计算机基础架构、计算框架、大数据管理都还在摸索中前进，Hadoop只是一个不得已的选择，但不是一个理想

的大数据挖掘计算框架。因此，我们非常期待大数据的基础研究能够取得突破，这些基础包括大数据的数学基础、计算基础、数据基础、分析基础和应用基础。

本书的疏漏和错误完全是作者的水平有限所致，如能获得读者的指正是我们的荣幸。

作 者

目 录

第1章

绪论