国之重器出版工程

网络强国建设

学术中国·大数据

Data Mining in the Era of Big Data

大数据时代的数据挖掘

李涛 著

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

国之重器出版工程

网络强国建设

学术中国·大数据

大数据时代的数据挖掘

Data Mining in the Era of Big Data

李涛 著

人民邮电出版社

北京

图书在版编目（CIP）数据

大数据时代的数据挖掘 / 李涛著. -- 北京 : 人民邮电出版社, 2019.1
（国之重器出版工程·学术中国·大数据）
ISBN 978-7-115-49239-5

Ⅰ. ①大… Ⅱ. ①李… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第199718号

内 容 提 要

本书以当前热点的数据挖掘应用贯穿全书，通过详解大数据挖掘技术在系统日志、工作票、可持续性研究、推荐系统、智能问答系统、社交媒体、生物信息学与健康医疗、隐私保护等方面的实际应用案例，阐述了如何更好地应用和学习数据挖掘技术。本书融入了数据挖掘前沿技术和典型应用，不仅适合热爱和关心数据挖掘技术的学术界和工业界人士阅读，还适合作为各大高校的数据挖掘和机器学习课堂的实践教材和参考书籍。本书有助于读者更好地理解数据挖掘技术背后的根源和本质。

◆ 著　　李　涛
责任编辑　吴娜达
责任印制　杨林杰
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷
◆ 开本：710×1000　1/16
印张：34.75　　2019年 1月第1版
字数：467千字　　2019年 1月河北第1次印刷

定价：189.00元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

《国之重器出版工程》
编辑委员会

苏哲子　中国工程院院士

李伯虎　中国工程院院士

李应红　中国科学院院士

李新亚　国家制造强国建设战略咨询委员会委员、中国机械工业联合会副会长

杨德森　中国工程院院士

张宏科　北京交通大学下一代互联网互联设备国家工程实验室主任

陆建勋　中国工程院院士

陆燕荪　国家制造强国建设战略咨询委员会委员、原机械工业部副部长

陈一坚　中国工程院院士

陈懋章　中国工程院院士

金东寒　中国工程院院士

周立伟　中国工程院院士

郑纬民　中国计算机学会原理事长

郑建华　中国科学院院士

屈贤明　国家制造强国建设战略咨询委员会委员、工业和信息化部智能制造专家咨询委员会副主任

项昌乐　“长江学者奖励计划”特聘教授，中国科协书记处书记，北京理工大学党委副书记、副校长

柳百成　中国工程院院士

闻雪友　中国工程院院士

徐德民　中国工程院院士

唐长红　中国工程院院士

黄卫东　“长江学者奖励计划”特聘教授

黄先祥　中国工程院院士

黄　维　中国科学院院士、西北工业大学常务副校长

董景辰　工业和信息化部智能制造专家咨询委员会委员

焦宗夏　“长江学者奖励计划”特聘教授

《学术中国·大数据》丛书
编辑委员会

丛书总序

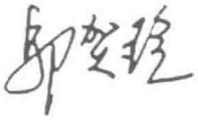
大数据、人工智能、云计算、物联网、移动互联网和产业互联网等成为新一代信息技术的特征，其中大数据与上述技术和应用都有密切关系。大数据来自于移动互联网、产业互联网和物联网等，其存储需要云计算，其挖掘依靠人工智能，而人工智能也有赖于大数据的支撑，大数据是产业互联网的重要基础。大数据不仅可以用于社会的精细化管理，更好地服务民生，大数据产业也将形成信息产业新的分支，其间接的产业影响将更大。可以说，大数据是数字经济的重要支柱。

很多国家都将大数据作为新时期的国家发展战略。2015 年，国务院印发大数据发展的首个权威性、系统性文件《促进大数据发展行动纲要》，2016 年国家发展和改革委员会批复了 13 个大数据领域的国家工程实验室，我国一些省市也纷纷制定大数据发展战略与规划。当前，我国在大数据共享开放、大数据资源开发、大数据技术研发、大数据挖掘应用、大数据产业培育、大数据安全管理、大数据人才培养和大数据法规研究等方面全面部署，为我国实现供给侧结构性改革，促进产业升级和转型，提升国家竞争力，争取在国际领域的话语权和实现跨越式发展起到了不可或缺的作用。

然而，我国的大数据发展也面临一些亟待解决的问题，例如基础研究薄弱、创新能力不强、产业链条缺口、数据资源封闭、法律法规滞后、数据安全不力、数据人才短缺和数据设施布局不合理及利用率不高等。为了使我国的大数据应用与产业可持续健康发展，需要多管齐下，其中普及大数据科学是重要的一环。为此，《学术中国·大数据》丛书编委会组织多个大数据领域优秀的研究团队的专家，基于国家

“973”计划、“863”计划、国家自然科学基金、国家重点研究计划等科研项目的创新研究成果和国内外大数据应用的成功实践，编写了这套丛书，内容涵盖大数据存储、数据管理、数据挖掘、分析平台、优化算法等核心技术领域。

本丛书的出版对传播大数据科学知识、推动大数据的学术探讨、鼓励大数据领域的产学研用协同创新、促进大数据标准化研究、加快大数据核心技术研发、培训大数据技术人才、引导大数据应用与产业化发展以及完善大数据有关的制度建设，都将起到积极作用。



2017 年 12 月

前 言

互联网技术的迅猛发展，催化数据量呈现指数级增长，一座座数据金山堆积在我们面前。然而，从实际的角度考虑，大数据的一个关键特征就是数据量巨大、知识贫瘠。于是，当人们面对 TB 级别甚至 PB 级别的数据量时，再也无法通过人工手段对数据进行知识提取，此时数据挖掘技术大显身手。

数据挖掘是大数据中最关键和最有价值的工作。2016 年 12 月，麦肯锡全球研究院（McKinsey Global Institution，MGI）发表了一份名为《分析的时代：在大数据的世界竞争（The Age of Analytics: Competing in a Data-Driven World）》的报告。该报告指出近年来数据量呈指数型增长，从而发展出更复杂的算法，计算机的存储能力也得到提升，随着技术日新月异的变化，商业模式也受到颠覆式的影响。

在这样的背景下，利用先进的数据挖掘技术，迎合各领域实际的需求痛点，才是和谐发展之道。大数据解决方案能够给企业带来巨大的资金效率和生产效率提升。IBM、谷歌、微软、阿里巴巴等 IT 巨头也将大数据描述成一种颠覆性的技术，其力量在将来足以影响和改变我们每一个人，甚至一个行业和一个国家。若想充分发挥大数据的巨大潜力，数据的产生和收集是基本，数据挖掘（知识发现）是工具和手段，是大数据应用中最关键和最有价值的工作。

作者长期从事数据挖掘研究和教学工作，经历了从最初数据挖掘基础研究的兴起到如今数据挖掘应用百花齐放这样一个时代的变迁，深刻体会到研究和应用两者间不可分割的联系：数据挖掘研究源于实践中的实际应用需求，以具体的应用数据为驱动，以方法、工具和系统为支撑，最终将发现的知识和信息运用到实践中，从

而提供量化的、合理的、可行的、能够产生巨大价值的信息。

大数据挖掘技术提供智能决策依据，在技术进步和人类生活的方方面面大显身手。本书针对大数据挖掘技术的不同应用场景，分别介绍了大数据技术在系统日志和事件的挖掘、工作票数据挖掘、大数据与计算可持续性研究、推荐系统、隐私保护等方面的应用。

本书既通俗易懂，又比较全面，融入了最新前沿技术和应用，适合不同背景的读者阅读，也欢迎各大高校的师生把此书作为数据挖掘和机器学习课堂的实践教材和参考书籍。

目 录