



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

编辑委员会主任：苗 圩

编辑委员会副主任：刘利华 辛国斌

编辑委员会委员：

冯长辉	梁志峰	高东升	姜子琨	许科敏
陈 因	郑立新	马向晖	高云虎	金 鑫
李 巍	李 东	高延敏	何 琼	刁石京
谢少锋	闻 库	韩 夏	赵志国	谢远生
赵永红	韩占武	刘 多	尹丽波	赵 波
卢 山	徐惠彬	赵长禄	周 玉	姚 郁
张 炜	聂 宏	付梦印	季仲华	



## 专家委员会委员（按姓氏笔画排列）：

- 于 全 中国工程院院士
- 王少萍 “长江学者奖励计划”特聘教授
- 王建民 清华大学软件学院院长
- 王哲荣 中国工程院院士
- 王 越 中国科学院院士、中国工程院院士
- 尤肖虎 “长江学者奖励计划”特聘教授
- 邓宗全 中国工程院院士
- 甘晓华 中国工程院院士
- 叶培建 中国科学院院士
- 朱英富 中国工程院院士
- 朵英贤 中国工程院院士
- 邬贺铨 中国工程院院士
- 刘大响 中国工程院院士
- 刘怡昕 中国工程院院士
- 刘韵洁 中国工程院院士
- 孙逢春 中国工程院院士
- 苏彦庆 “长江学者奖励计划”特聘教授



- 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

编辑委员会顾问：

邬贺铨      李国杰      李德毅      方滨兴

编辑委员会主任：郑纬民

编辑委员会委员（按姓氏笔画排列）：

王建民      杜跃进      李国庆      李 涛      宋 杰  
张广艳      陈 卫      陈世敏      魏哲巍

策 划：《大数据》杂志





## 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

第 1 章 数据挖掘简介	1
1.1 大数据时代的数据挖掘	2
1.1.1 大数据的特点“4V+4V”	3
1.1.2 数据挖掘	5
1.1.3 从数据挖掘应用的角度看大数据	7
1.2 数据挖掘技术的发展历史	8
1.3 十大数据挖掘算法简介	10
1.4 数据挖掘平台：FIU-Miner	21
1.4.1 FIU-Miner 平台简介	22
1.4.2 FIU-Miner 系统架构	22
1.4.3 FIU-Miner 应用实例	23
参考文献	28
第 2 章 系统日志和事件的挖掘	31
2.1 数据驱动的网络运维	32
2.1.1 网络运维 1.0 阶段：简单数据处理	33
2.1.2 网络运维 2.0 阶段：分布式大数据处理框架	34
2.1.3 网络运维 3.0 阶段：网络运维平台套件	34
2.1.4 网络运维 4.0 阶段：智能化网络运维	35
2.2 系统日志分析的目的	35
2.2.1 系统问题诊断	36
2.2.2 调试与优化	37
2.2.3 系统安全维护	37
2.3 日志数据分析管理系统的架构	38
2.3.1 日志数据的收集和预处理	39
2.3.2 历史日志数据存储	39
2.3.3 日志事件数据的分析以及对分析结果的展示和使用	39
2.4 系统日志的数据形式	40



2.4.1	无结构的日志数据	40
2.4.2	结构化与半结构化的日志数据	41
2.4.3	非结构化数据的转换	43
2.5	基于日志数据的异常检测	44
2.5.1	基于监督学习的异常检测	44
2.5.2	基于无监督学习的异常检测	48
2.6	系统故障根源跟踪	52
2.6.1	日志事件的依赖性挖掘	54
2.6.2	基于依赖关系的系统故障追踪	65
2.7	日志事件总结	65
2.7.1	事件总结算法基本要求及相关工作	66
2.7.2	基于事件发生频率变迁描述的事件总结	67
2.7.3	基于马尔可夫模型描述的事件总结	67
2.7.4	基于事件关系网络描述的事件总结	68
	参考文献	69
<b>第3章</b>	<b>工作票数据挖掘</b>	<b>75</b>
3.1	工作票简介	76
3.2	工作票产生机制和亟待解决的问题	77
3.3	研究现状	79
3.3.1	工作票分类	80
3.3.2	工作票推荐	82
3.3.3	整体解决方案和工具	84
3.4	工作票漏报和误报检测	84
3.4.1	漏报和误报	84
3.4.2	基于规则的误报识别方法	86
3.4.3	半监督的工作票漏报发现方法	89
3.4.4	评价	92
3.5	层次多标签工作票分类	96
3.5.1	问题描述	98
3.5.2	层次损失函数和期望损失最小化	98
3.5.3	算法和解决方案	102
3.5.4	实验	104
3.6	工作票解决方案推荐	108



3.6.1	背景 .....	108
3.6.2	基于 KNN 的推荐方法 .....	109
3.6.3	划分方法 .....	111
3.6.4	概率融合方法 .....	112
3.6.5	度量学习方法 .....	113
3.6.6	实验 .....	116
参考文献	.....	126
<b>第 4 章</b>	<b>大数据与计算可持续性研究 .....</b>	<b>131</b>
4.1	大数据与可持续发展 .....	132
4.1.1	可持续发展 .....	132
4.1.2	大数据时代可持续发展面临的机遇和挑战 .....	133
4.2	计算可持续性 .....	133
4.2.1	计算可持续性数据及其特征 .....	134
4.2.2	大数据环境下计算可持续性研究现状 .....	137
4.3	研究案例 .....	142
4.3.1	基于数据驱动的气象分析 .....	142
4.3.2	基于数据驱动的建筑能耗分析 .....	145
参考文献	.....	155
<b>第 5 章</b>	<b>推荐系统 .....</b>	<b>159</b>
5.1	个性化推荐系统概述 .....	160
5.2	推荐技术 .....	163
5.2.1	基于内容的推荐系统 .....	163
5.2.2	基于协同过滤的推荐系统 .....	164
5.2.3	基于知识的推荐系统 .....	165
5.2.4	基于混合技术的推荐系统 .....	165
5.2.5	基于计算智能的推荐系统 .....	166
5.2.6	基于社交网络的推荐系统 .....	167
5.2.7	基于上下文敏感的推荐系统 .....	169
5.2.8	基于组群的推荐系统 .....	170
5.3	推荐系统评测 .....	170
5.3.1	推荐系统评测环境 .....	171
5.3.2	推荐系统评测指标 .....	174



5.4 推荐系统实例	181
5.4.1 新闻推荐	181
5.4.2 POI 推荐	190
参考文献	198
<b>第6章 智能问答系统</b>	<b>203</b>
6.1 发展历史	204
6.2 句法分析	205
6.2.1 中文分词技术	205
6.2.2 词的分类和兼类	207
6.2.3 汉语句法分析	208
6.3 问题理解	210
6.3.1 词法分析	210
6.3.2 问题分类	210
6.3.3 关键词扩展与抽取	211
6.3.4 答案抽取	212
6.4 问题检索	212
6.4.1 基于词法的问句检索	212
6.4.2 基于句法的问句检索	213
6.4.3 基于语义的问句检索	213
6.4.4 常见问题集的问候检索	213
6.5 信息抽取	214
6.5.1 抽取的对象	214
6.5.2 抽取的种类	215
6.5.3 抽取的方法	215
6.6 知识库构建	217
6.6.1 基本概念	217
6.6.2 体系结构	218
6.6.3 关键技术	219
6.7 知识推理	223
6.7.1 线索挖掘	223
6.7.2 关系推理	224
6.7.3 关系预测	225
6.8 案例分析	225





6.8.1 限定域系统的现有案例分析 .....	225
6.8.2 开放域系统的现有案例分析 .....	233
参考文献 .....	238
<b>第7章 文本挖掘 .....</b>	<b>245</b>
7.1 文本表示 .....	246
7.2 话题挖掘 .....	248
7.2.1 非负矩阵分解 .....	248
7.2.2 概率潜在语义分析 .....	249
7.2.3 潜在狄利克雷分配模型 .....	250
7.2.4 分析与实例比较 .....	251
7.3 多文档自动文摘 .....	253
7.3.1 目标函数选择: 句子重要性评价 .....	253
7.3.2 优化方法 .....	257
7.3.3 其他的自动文摘问题 .....	258
7.3.4 实例分析 .....	259
7.4 情感分析和摘要 .....	262
7.4.1 基于频繁项集的方法 .....	264
7.4.2 实例分析 .....	266
7.4.3 基于方面的话题模型分析方法 .....	267
7.5 数据挖掘在专利分析中的应用 .....	272
7.5.1 专利分析的内容、流程与方法 .....	273
7.5.2 数据挖掘在专利分析中的应用方向 .....	278
参考文献 .....	284
<b>第8章 多媒体数据挖掘 .....</b>	<b>291</b>
8.1 多媒体技术的特点 .....	292
8.1.1 数字化 .....	292
8.1.2 多样性 .....	293
8.1.3 集成性 .....	293
8.1.4 交互性 .....	293
8.1.5 非线性 .....	294
8.1.6 实时性 .....	294
8.2 多媒体数据挖掘概述 .....	294