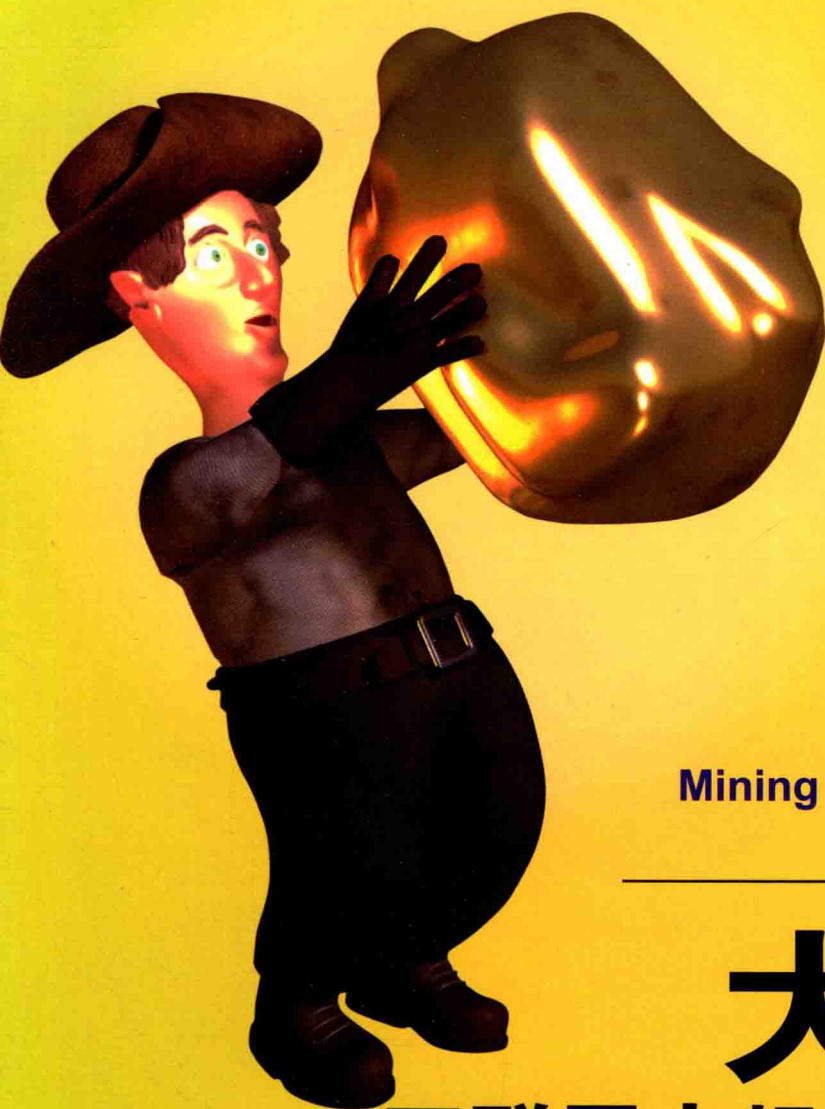




Jure Leskovec
【美】Anand Rajaraman 著
Jeffrey David Ullman
王斌 译

Mining of Massive Datasets
Second Edition

大数据

互联网大规模数据挖掘 与分布式处理(第2版)



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

TURING

图灵程序设计丛书



Jure Leskovec

【美】Anand Rajaraman 著

Jeffrey David Ullman

王斌 译



Mining of Massive Datasets

Second Edition

大数据

互联网大规模数据挖掘 与分布式处理(第2版)

人民邮电出版社

北京

图书在版编目(CIP)数据

大数据：互联网大规模数据挖掘与分布式处理：第2版 / (美) 莱斯科夫 (Leskovec, J.), (美) 拉贾拉曼 (Rajaraman, A.), (美) 厄尔曼 (Ullman, J. D.) 著；王斌译. — 2版. — 北京：人民邮电出版社，2015.7

(图灵程序设计丛书)

ISBN 978-7-115-39525-2

I. ①大… II. ①莱… ②拉… ③厄… ④王… III. ①数据处理 IV. ①TP274

中国版本图书馆CIP数据核字(2015)第123045号

内 容 提 要

本书由斯坦福大学“Web 挖掘”课程的内容总结而成，主要关注极大规模数据的挖掘。主要内容包括分布式文件系统、相似性搜索、搜索引擎技术、频繁项集挖掘、聚类算法、广告管理及推荐系统、社会网络图挖掘和大规模机器学习等。其中每一章节有对应的习题，以巩固所讲解的内容。读者更可以从网上获取相关拓展材料。

本书适合本科生、研究生及对数据挖掘感兴趣的读者阅读。

-
- ◆ 著 [美] Jure Leskovec [美] Anand Rajaraman
[美] Jeffrey David Ullman
译 王 斌
责任编辑 岳新欣
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>
北京鑫正大印刷有限公司印刷
- ◆ 开本：800×1000 1/16
印张：24.25
字数：573千字 2015年7月第2版
印数：40 001—45 000册 2015年7月北京第1次印刷
著作权合同登记号 图字：01-2015-3276号
-

定价：79.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316

反盗版热线：(010)81055315

广告经营许可证：京崇工商广字第 0021 号

版 权 声 明

Mining of Massive Datasets second edition (978-1-107-07723-2) by Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman first published by Cambridge University Press 2014.

All rights reserved.

This simplified Chinese edition for the People's Republic of China is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press & Posts & Telecom Press 2015.

This book is in copyright. No reproduction of any part may take place without the written permission of Cambridge University Press and Posts & Telecom Press.

This edition is for sale in the People's Republic of China (excluding Hong Kong SAR, Macao SAR and Taiwan Province) only.

此版本仅限在中华人民共和国境内（不包括香港、澳门特别行政区及台湾地区）销售。

译者序

很高兴本书的第2版能和读者见面,和第1版相比,这一版增加了第10、11、12章。第10章介绍了近年来十分流行的社会网络分析技术,第11章对高维数据空间的降维技术进行了阐述,第12章则介绍了大规模数据下的机器学习方法。除此之外,这一版第2章也对第1版的内容进行了扩充,主要增加了对MapReduce算法的复杂度分析理论。除了内容的变化,本书的作者也从原来的两位增加到三位,在学术界(特别是社交网络挖掘领域)如日中天的斯坦福大学年轻帅哥Jure Leskovec博士,也加入到本书的作者行列。

本书第1版出版之后,获得了不少读者的积极反馈,这些反馈也在第2版中有所体现。需要指出的是,本书是一本面向大数据挖掘的技术而非概念性图书,需要反复研读认真实践才能真正理解。还有,本书主要基于MapReduce框架来介绍分布式挖掘算法的实现。目前大数据包罗万象、实现框架众多,数据挖掘并不是唯一关键技术,MapReduce也不是唯一可选框架。读者可以通过阅读其他书籍进行补充。

我曾于2009年翻译了《信息检索导论》一书。在我的理解体系下,信息检索是一门跨众多学科领域的研究方向,其主要的形式包括搜索、推荐和挖掘三种。如果说先前翻译的《信息检索导论》注重信息检索的基本理论和搜索应用,那么本书则关注了推荐和挖掘应用。在这个意义上说,这两本书可以互为补充。这也是我选择本书进行翻译的原因之一。另一个原因在于本书集中关注大数据处理这个极具研究和应用前景的话题,一想到它可以为很多人带来帮助就让我欣慰不已。

本书主要以Web上的数据为对象介绍大规模情况下的数据挖掘。除了传统的聚类、频繁项发现及链接分析等内容外,它还介绍了数据流挖掘、互联网广告、推荐系统、社会网络分析及分布式机器学习等近年来被广泛关注的话题。特别地,本书专门介绍了支持大规模数据挖掘的分布式文件系统及MapReduce分布式计算框架。和《信息检索导论》相比,本书在理论上虽然可能不如前者深入,但是它在简明扼要阐明基本原理的基础上,更侧重大数据环境下的实际算法实现。具体地,本书给出了在面对大规模数据时基于MapReduce框架的多个算法实现。换句话说,它的算法可以在大数据环境下真正“落地”,这无疑给想要或致力于大数据挖掘的读者带来理解和实现上的巨大裨益。

虽然我的很多学生都对本书内容有较深的理解,但是为了保持翻译风格的一致性并对本书翻译负全部责任,在出版社的建议下我还是与前一本书一样选择了自己独立翻译。感谢复旦大学黄萱菁教授、中科院自动化所赵军研究员、中科院软件所孙乐研究员、中科院研究生院何笨博士等

人对本书第1版及第2版提出的建设性意见和建议。对他们的无私帮助，我表示由衷的感谢。感谢图灵公司的武卫东、傅志红、李松峰、岳新欣等人为本书付出的努力，感谢人民邮电出版社杨海玲女士的大力引荐。通过翻译，我也认识了图灵公司及图灵社区的众多朋友，并从他们身上学到了很多宝贵的东西。感谢对我译书给予支持和鼓励的李锦涛研究员、孟丹研究员、郭莉研究员、刘群研究员、贺劲博士、虎嵩林博士等领导、朋友和同事。感谢我的学生们作为最早的读者给予的建议和意见，其中李佩佳、叶邦宇、洪洁等提出了许多十分宝贵的意见。感谢我的家人，他们总是无怨无悔地给我最大的支持和包容，让我能够全身心投入到工作和翻译当中。由于翻译基本在业余时间尤其是晚间完成，因此晚睡便成了家常便饭。我的儿子心心知道我要翻译便会按时睡觉不再打扰我，这让我感到欣慰并给我力量。翻译过程中，我和原书作者Jeffrey David Ullman进行了邮件交流，澄清了理解上的一些误区，并更正了原书中一些错误。我的翻译也得到了对方的热情鼓励。

因本人各方面水平有限，现有译文中肯定存在许多不足。希望读者能够和我联系，提出疑问和勘误，以便能够不断改进本书质量。来信请联系wbxjj2008@gmail.com，本书勘误会及时公布在图灵社区网站www.ituring.com.cn上。原书的初稿电子版等信息也可以从网站<http://mmms.org>下载。

王 斌

2014年12月15日于中关村

前言

本书根据Anand Rajaraman和Jeff Ullman于斯坦福大学教授多年的一门季度课程的材料汇编而成。该课程名为“Web挖掘”(编号CS345A), 尽管它已经成为高年级本科生能接受并感兴趣的课程之一, 但其原本是一门为高年级研究生设计的课程。Jure Leskovec到斯坦福大学任职后, 我们对相关材料进行了重新组织。他开设了一门有关网络分析的新课程CS224W并为CS345A增加了一些内容, 后者重新编号为CS246。三位作者也开设了一门大规模数据挖掘的项目课程CS341。目前本书包含了所有三门课程的教学内容。

本书内容

简单来说, 本书是关于数据挖掘的。但是, 本书主要关注极大规模数据的挖掘, “极大规模”的意思是说这些数据大到无法在内存中存放。由于重点强调数据的规模, 所以本书的例子大都来自Web本身或者Web上导出的数据。另外, 本书从算法的角度来看待数据挖掘, 即数据挖掘是将算法应用于数据, 而不是使用数据来“训练”某种类型的机器学习引擎。

本书的主要内容包括:

- (1) 分布式文件系统以及已成功应用于大规模数据集并行算法构建的MapReduce工具;
- (2) 相似性搜索, 包括最小哈希和局部敏感哈希的关键技术;
- (3) 数据流处理以及面对快速到达、须立即处理、易丢失的数据的专用处理算法;
- (4) 搜索引擎技术, 包括谷歌的PageRank、链接作弊检测及计算网页导航度(hub)和权威度(authority)的HITS方法;
- (5) 频繁项集挖掘, 包括关联规则挖掘、购物篮分析、A-Priori及其改进算法;
- (6) 大规模高维数据集的聚类算法;
- (7) Web应用中的两个关键问题: 广告管理及推荐系统;
- (8) 对极大规模的图(特别是社会网络图)的结构进行分析和挖掘的算法;
- (9) 通过降维来获得大规模数据集的重要性质的技术, 包括SVD分解和隐性语义索引;
- (10) 可以应用于极大规模数据的机器学习算法, 包括感知机、支持向量机和梯度下降法。

先修课程

为了让读者完全领会本书内容, 我们推荐如下先修课程:

- (1) 数据库系统入门, 包括SQL及相关编程系统;
- (2) 大二的数据结构、算法及离散数学课程;
- (3) 大二的软件系统、软件工程及编程语言课程。

习题

本书包含大量的习题, 基本每节都有对应习题。较难的习题或其中较难的部分都用惊叹号“!”来标记, 而最难的习题则标有双惊叹号“!!”。

Web上的支持

读者可以从下列网址获得CS345A课程提供的更多材料: <http://mmds.org>。

在该网址下, 读者可以找到课件、课后作业、项目作业等材料, 某些情况下可能还有试题。

Gradiance自动化作业

本书使用www.gradiance.com/services提供的Gradiance根问题技术, 提供了一些自动化习题。学生可以在网站上通过创建账号来访问公开课, 并通过代码1EDD8A1D访问该课程。授课老师可以建立账号, 然后将登录账号、学院名称及MMDS材料请求信息发给support@gradiance.com。

致谢

本书封面由Scott Ullman设计。

感谢Foto Afrati、Arun Marathe和Rok Sasic精心阅读本书初稿并提出建设性的意见。

感谢Apoorv Agarwal、Aris Anagnostopoulos、Atilla Soner Balkir、Robin Bennett、Susan Biancani、Amitabh Chaudhary、Leland Chen、Anastasios Gounaris、Shrey Gupta、Waleed Hameid、Ed Knorr、Haewoon Kwak、Ellis Lau、Ethan Lozano、Michael Mahoney、Justin Meyer、Brad Penoff、Philips Kokoh Prasetyo、Qi Ge、Angad Singh、Sandeep Sripada、Dennis Sidharta、Krzysztof Stencel、Mark Storus、Roshan Sumbaly、Zack Taylor、Tim Triche Jr.、Wang Bin、Weng Zhen-Bin、Robert West、Oscar Wu、Xie Ke、Nicolas Zhao及Zhou Jingbo指出了本书中的部分错误。当然, 其余错误均由我们负责。

J. L.

A. R.

J. D. U.

加利福尼亚州帕洛阿尔托

2014年3月

目 录

第 1 章 数据挖掘基本概念	1
1.1 数据挖掘的定义	1
1.1.1 统计建模	1
1.1.2 机器学习	1
1.1.3 建模的计算方法	2
1.1.4 数据汇总	2
1.1.5 特征抽取	3
1.2 数据挖掘的统计限制	4
1.2.1 整体情报预警	4
1.2.2 邦弗朗尼原理	4
1.2.3 邦弗朗尼原理的一个例子	5
1.2.4 习题	6
1.3 相关知识	6
1.3.1 词语在文档中的重要性	6
1.3.2 哈希函数	7
1.3.3 索引	8
1.3.4 二级存储器	9
1.3.5 自然对数的底 e	10
1.3.6 幂定律	11
1.3.7 习题	12
1.4 本书概要	13
1.5 小结	14
1.6 参考文献	15
第 2 章 MapReduce 及新软件栈	16
2.1 分布式文件系统	17
2.1.1 计算节点的物理结构	17
2.1.2 大规模文件系统的结构	18
2.2 MapReduce	19
2.2.1 Map 任务	20
2.2.2 按键分组	20
2.2.3 Reduce 任务	21
2.2.4 组合器	21
2.2.5 MapReduce 的执行细节	22
2.2.6 节点失效的处理	23
2.2.7 习题	23
2.3 使用 MapReduce 的算法	23
2.3.1 基于 MapReduce 的矩阵-向量 乘法实现	24
2.3.2 向量 v 无法放入内存时的处理	24
2.3.3 关系代数运算	25
2.3.4 基于 MapReduce 的选择运算	27
2.3.5 基于 MapReduce 的投影运算	27
2.3.6 基于 MapReduce 的并、交和差 运算	28
2.3.7 基于 MapReduce 的自然连接 运算	28
2.3.8 基于 MapReduce 的分组和聚 合运算	29
2.3.9 矩阵乘法	29
2.3.10 基于单步 MapReduce 的矩阵 乘法	30
2.3.11 习题	31
2.4 MapReduce 的扩展	31
2.4.1 工作流系统	32
2.4.2 MapReduce 的递归扩展版本	33
2.4.3 Pregel 系统	35
2.4.4 习题	35
2.5 通信开销模型	36
2.5.1 任务网络的通信开销	36
2.5.2 时钟时间	37
2.5.3 多路连接	38

2.5.4 习题	41	3.5.2 欧氏距离	71
2.6 MapReduce 复杂性理论	41	3.5.3 Jaccard 距离	72
2.6.1 Reducer 规模及复制率	41	3.5.4 余弦距离	72
2.6.2 一个例子: 相似性连接	42	3.5.5 编辑距离	73
2.6.3 MapReduce 问题的一个图模型	44	3.5.6 海明距离	74
2.6.4 映射模式	45	3.5.7 习题	74
2.6.5 并非所有输入都存在时的处理	46	3.6 局部敏感函数理论	75
2.6.6 复制率的下界	46	3.6.1 局部敏感函数	76
2.6.7 案例分析: 矩阵乘法	48	3.6.2 面向 Jaccard 距离的局部敏感函数族	77
2.6.8 习题	51	3.6.3 局部敏感函数族的放大处理	77
2.7 小结	51	3.6.4 习题	79
2.8 参考文献	53	3.7 面向其他距离测度的 LSH 函数族	80
第 3 章 相似项发现	55	3.7.1 面向海明距离的 LSH 函数族	80
3.1 近邻搜索的应用	55	3.7.2 随机超平面和余弦距离	80
3.1.1 集合的 Jaccard 相似度	55	3.7.3 梗概	81
3.1.2 文档的相似度	56	3.7.4 面向欧氏距离的 LSH 函数族	82
3.1.3 协同过滤——一个集合相似问题	57	3.7.5 面向欧氏空间的更多 LSH 函数族	83
3.1.4 习题	58	3.7.6 习题	83
3.2 文档的 shingling	58	3.8 LSH 函数的应用	84
3.2.1 k -shingle	58	3.8.1 实体关联	84
3.2.2 shingle 大小的选择	59	3.8.2 一个实体关联的例子	85
3.2.3 对 shingle 进行哈希	59	3.8.3 记录匹配的验证	86
3.2.4 基于词的 shingle	60	3.8.4 指纹匹配	87
3.2.5 习题	60	3.8.5 适用于指纹匹配的 LSH 函数族	87
3.3 保持相似度的集合摘要表示	61	3.8.6 相似新闻报道检测	88
3.3.1 集合的矩阵表示	61	3.8.7 习题	89
3.3.2 最小哈希	62	3.9 面向高相似度的方法	90
3.3.3 最小哈希及 Jaccard 相似度	62	3.9.1 相等项发现	90
3.3.4 最小哈希签名	63	3.9.2 集合的字符串表示方法	91
3.3.5 最小哈希签名的计算	63	3.9.3 基于长度的过滤	91
3.3.6 习题	66	3.9.4 前缀索引	92
3.4 文档的局部敏感哈希算法	67	3.9.5 位置信息的使用	93
3.4.1 面向最小哈希签名的 LSH	67	3.9.6 使用位置和长度信息的索引	94
3.4.2 行条化策略的分析	68	3.9.7 习题	96
3.4.3 上述技术的综合	69	3.10 小结	97
3.4.4 习题	70	3.11 参考文献	98
3.5 距离测度	70		
3.5.1 距离测度的定义	71		

第 4 章 数据流挖掘	100	4.7.1 最常见元素问题	121
4.1 流数据模型	100	4.7.2 衰减窗口的定义	122
4.1.1 一个数据流管理系统	100	4.7.3 最流行元素的发现	123
4.1.2 流数据源的例子	101	4.8 小结	123
4.1.3 流查询	102	4.9 参考文献	124
4.1.4 流处理中的若干问题	103		
4.2 流当中的数据抽样	103	第 5 章 链接分析	126
4.2.1 一个富于启发性的例子	104	5.1 PageRank	126
4.2.2 代表性样本的获取	104	5.1.1 早期的搜索引擎及词项作弊	126
4.2.3 一般的抽样问题	105	5.1.2 PageRank 的定义	128
4.2.4 样本规模的变化	105	5.1.3 Web 结构	130
4.2.5 习题	106	5.1.4 避免终止点	132
4.3 流过滤	106	5.1.5 采集器陷阱及“抽税”法	134
4.3.1 一个例子	106	5.1.6 PageRank 在搜索引擎中的 使用	136
4.3.2 布隆过滤器	107	5.1.7 习题	136
4.3.3 布隆过滤方法的分析	107	5.2 PageRank 的快速计算	137
4.3.4 习题	108	5.2.1 转移矩阵的表示	137
4.4 流中独立元素的数目统计	109	5.2.2 基于 MapReduce 的 PageRank 迭代计算	138
4.4.1 独立元素计数问题	109	5.2.3 结果向量合并时的组合器 使用	139
4.4.2 FM 算法	109	5.2.4 转移矩阵中块表示	140
4.4.3 组合估计	110	5.2.5 其他高效的 PageRank 迭代 方法	141
4.4.4 空间需求	111	5.2.6 习题	142
4.4.5 习题	111	5.3 面向主题的 PageRank	142
4.5 矩估计	111	5.3.1 动机	142
4.5.1 矩定义	111	5.3.2 有偏的随机游走模型	143
4.5.2 二阶矩估计的 AMS 算法	112	5.3.3 面向主题的 PageRank 的使用	144
4.5.3 AMS 算法有效的原因	113	5.3.4 基于词汇的主题推断	144
4.5.4 更高阶矩的估计	113	5.3.5 习题	145
4.5.5 无限流的处理	114	5.4 链接作弊	145
4.5.6 习题	115	5.4.1 垃圾农场的架构	145
4.6 窗口内的计数问题	116	5.4.2 垃圾农场的分析	147
4.6.1 精确计数的开销	116	5.4.3 与链接作弊的斗争	147
4.6.2 DGIM 算法	116	5.4.4 TrustRank	148
4.6.3 DGIM 算法的存储需求	118	5.4.5 垃圾质量	148
4.6.4 DGIM 算法中的查询应答	118	5.4.6 习题	149
4.6.5 DGIM 条件的保持	119	5.5 导航页和权威页	149
4.6.6 降低错误率	120		
4.6.7 窗口内计数问题的扩展	120		
4.6.8 习题	121		
4.7 衰减窗口	121		

5.5.1 HITS 的直观意义	150	6.6 小结	184
5.5.2 导航度和权威度的形式化	150	6.7 参考文献	186
5.5.3 习题	153		
5.6 小结	153	第 7 章 聚类	187
5.7 参考文献	155	7.1 聚类技术介绍	187
第 6 章 频繁项集	157	7.1.1 点、空间和距离	187
6.1 购物篮模型	157	7.1.2 聚类策略	188
6.1.1 频繁项集的定义	157	7.1.3 维数灾难	189
6.1.2 频繁项集的应用	159	7.1.4 习题	190
6.1.3 关联规则	160	7.2 层次聚类	190
6.1.4 高可信度关联规则地发现	161	7.2.1 欧氏空间下的层次聚类	191
6.1.5 习题	162	7.2.2 层次聚类算法的效率	194
6.2 购物篮及 A-Priori 算法	163	7.2.3 控制层次聚类的其他规则	194
6.2.1 购物篮数据的表示	163	7.2.4 非欧空间下的层次聚类	196
6.2.2 项集计数中的内存使用	164	7.2.5 习题	197
6.2.3 项集的单调性	165	7.3 k-均值算法	198
6.2.4 二元组计数	166	7.3.1 k-均值算法基本知识	198
6.2.5 A-Priori 算法	166	7.3.2 k-均值算法的簇初始化	198
6.2.6 所有频繁项集上的 A-Priori 算法	168	7.3.3 选择正确的 k 值	199
6.2.7 习题	169	7.3.4 BFR 算法	200
6.3 更大数据集在内存中的处理	170	7.3.5 BFR 算法中的数据处理	202
6.3.1 PCY 算法	171	7.3.6 习题	203
6.3.2 多阶段算法	172	7.4 CURE 算法	204
6.3.3 多哈希算法	174	7.4.1 CURE 算法的初始化	205
6.3.4 习题	175	7.4.2 CURE 算法的完成	206
6.4 有限扫描算法	177	7.4.3 习题	206
6.4.1 简单的随机化算法	177	7.5 非欧空间下的聚类	207
6.4.2 抽样算法中的错误规避	178	7.5.1 GRGPF 算法中的簇表示	207
6.4.3 SON 算法	179	7.5.2 簇表示树的初始化	207
6.4.4 SON 算法和 MapReduce	179	7.5.3 GRGPF 算法中的点加入	208
6.4.5 Toivonen 算法	180	7.5.4 簇的分裂及合并	209
6.4.6 Toivonen 算法的有效性分析	181	7.5.5 习题	210
6.4.7 习题	181	7.6 流聚类及并行化	210
6.5 流中的频繁项计数	182	7.6.1 流计算模型	210
6.5.1 流的抽样方法	182	7.6.2 一个流聚类算法	211
6.5.2 衰减窗口中的频繁项集	183	7.6.3 桶的初始化	211
6.5.3 混合方法	183	7.6.4 桶合并	211
6.5.4 习题	184	7.6.5 查询应答	213
		7.6.6 并行环境下的聚类	213
		7.6.7 习题	214

7.7 小结	214	9.1.4 效用矩阵的填充	241
7.8 参考文献	216	9.2 基于内容的推荐	242
第 8 章 Web 广告	218	9.2.1 项模型	242
8.1 在线广告相关问题	218	9.2.2 文档的特征发现	242
8.1.1 广告机会	218	9.2.3 基于 Tag 的项特征获取	243
8.1.2 直投广告	219	9.2.4 项模型的表示	244
8.1.3 展示广告的相关问题	219	9.2.5 用户模型	245
8.2 在线算法	220	9.2.6 基于内容的项推荐	246
8.2.1 在线和离线算法	220	9.2.7 分类算法	247
8.2.2 贪心算法	221	9.2.8 习题	248
8.2.3 竞争率	222	9.3 协同过滤	249
8.2.4 习题	222	9.3.1 相似度计算	249
8.3 广告匹配问题	223	9.3.2 相似度对偶性	252
8.3.1 匹配及完美匹配	223	9.3.3 用户聚类 and 项聚类	253
8.3.2 最大匹配贪心算法	224	9.3.4 习题	254
8.3.3 贪心匹配算法的竞争率	224	9.4 降维处理	254
8.3.4 习题	225	9.4.1 UV 分解	255
8.4 adwords 问题	225	9.4.2 RMSE	255
8.4.1 搜索广告的历史	226	9.4.3 UV 分解的增量式计算	256
8.4.2 adwords 问题的定义	226	9.4.4 对任一元素的优化	259
8.4.3 adwords 问题的贪心方法	227	9.4.5 一个完整 UV 分解算法的构 建	259
8.4.4 Balance 算法	228	9.4.6 习题	261
8.4.5 Balance 算法竞争率的一个 下界	228	9.5 NetFlix 竞赛	262
8.4.6 多投标者的 Balance 算法	230	9.6 小结	263
8.4.7 一般性的 Balance 算法	231	9.7 参考文献	264
8.4.8 adwords 问题的最后论述	232	第 10 章 社会网络图挖掘	265
8.4.9 习题	232	10.1 将社会网络看成图	265
8.5 adwords 的实现	232	10.1.1 社会网络的概念	265
8.5.1 投标和搜索查询的匹配	233	10.1.2 将社会网络看成图	266
8.5.2 更复杂的匹配问题	233	10.1.3 各种社会网络的例子	267
8.5.3 文档和投标之间的匹配算法	234	10.1.4 多类型节点构成的图	268
8.6 小结	235	10.1.5 习题	269
8.7 参考文献	237	10.2 社会网络图的聚类	269
第 9 章 推荐系统	238	10.2.1 社会网络图的距离计算	269
9.1 一个推荐系统的模型	238	10.2.2 应用标准的聚类算法	270
9.1.1 效用矩阵	238	10.2.3 中介度	271
9.1.2 长尾现象	239	10.2.4 Girvan-Newman 算法	271
9.1.3 推荐系统的应用	241	10.2.5 利用中介度来发现社区	274

10.2.6	习题	275	10.8.5	智能传递闭包	303
10.3	社区的直接发现	275	10.8.6	基于图归约的传递闭包	304
10.3.1	团的发现	276	10.8.7	邻居规模的近似计算	305
10.3.2	完全二部图	276	10.8.8	习题	306
10.3.3	发现完全二部子图	277	10.9	小结	307
10.3.4	完全二部子图一定存在的 原因	277	10.10	参考文献	310
10.3.5	习题	279	第 11 章	降维处理	312
10.4	图划分	280	11.1	特征值和特征向量	312
10.4.1	图划分的好坏标准	280	11.1.1	定义	312
10.4.2	归一化割	280	11.1.2	特征值与特征向量计算	313
10.4.3	描述图的一些矩阵	281	11.1.3	基于幂迭代方法的特征对 求解	315
10.4.4	拉普拉斯矩阵的特征值	282	11.1.4	特征向量矩阵	317
10.4.5	其他图划分方法	284	11.1.5	习题	317
10.4.6	习题	284	11.2	主成分分析	318
10.5	重叠社区的发现	285	11.2.1	一个示例	318
10.5.1	社区的本质	285	11.2.2	利用特征向量进行降维	321
10.5.2	极大似然估计	286	11.2.3	距离矩阵	322
10.5.3	关系图模型	287	11.2.4	习题	323
10.5.4	避免成员隶属关系的离散式 变化	288	11.3	奇异值分解	323
10.5.5	习题	290	11.3.1	SVD 的定义	323
10.6	Simrank	290	11.3.2	SVD 解析	325
10.6.1	社会网络上的随机游走者	290	11.3.3	基于 SVD 的降维	326
10.6.2	带重启的随机游走	291	11.3.4	将较低奇异值置为 0 后有 效的原因	327
10.6.3	习题	293	11.3.5	使用概念进行查询处理	328
10.7	三角形计数问题	293	11.3.6	矩阵 SVD 的计算	329
10.7.1	为什么要对三角形计数	294	11.3.7	习题	330
10.7.2	一个寻找三角形的算法	294	11.4	CUR 分解	331
10.7.3	三角形寻找算法的最优性	295	11.4.1	CUR 的定义	331
10.7.4	基于 MapReduce 寻找三 角形	295	11.4.2	合理选择行和列	332
10.7.5	使用更少的 Reduce 任务	297	11.4.3	构建中间矩阵	333
10.7.6	习题	297	11.4.4	完整的 CUR 分解	334
10.8	图的邻居性质	298	11.4.5	去除重复行和列	335
10.8.1	有向图和邻居	298	11.4.6	习题	335
10.8.2	图的直径	299	11.5	小结	336
10.8.3	传递闭包和可达性	300	11.6	参考文献	337
10.8.4	基于 MapReduce 的传递闭包 求解	301	第 12 章	大规模机器学习	338
			12.1	机器学习模型	338

12.1.1	训练集	338	12.3.2	超平面归一化	356
12.1.2	一些例子	339	12.3.3	寻找最优逼近分界面	357
12.1.3	机器学习方法	341	12.3.4	基于梯度下降法求解 SVM	359
12.1.4	机器学习架构	342	12.3.5	随机梯度下降	363
12.1.5	习题	344	12.3.6	SVM 的并行实现	363
12.2	感知机	344	12.3.7	习题	363
12.2.1	训练阈值为 0 的感知机	344	12.4	近邻学习	364
12.2.2	感知机的收敛性	347	12.4.1	近邻计算的框架	364
12.2.3	Winnow 算法	347	12.4.2	最近邻学习	365
12.2.4	允许阈值变化的情况	349	12.4.3	学习一维函数	365
12.2.5	多类感知机	350	12.4.4	核回归	367
12.2.6	变换训练集	351	12.4.5	处理高维欧氏空间数据	368
12.2.7	感知机的问题	351	12.4.6	对非欧距离的处理	369
12.2.8	感知机的并行实现	353	12.4.7	习题	369
12.2.9	习题	354	12.5	各种学习方法的比较	370
12.3	支持向量机	354	12.6	小结	371
12.3.1	支持向量机的构成	354	12.7	参考文献	372

第1章

数据挖掘基本概念



本章为全书的导论部分,首先阐述数据挖掘的本质,并讨论其在多个相关学科中的不同理解。接着介绍邦弗朗尼原理 (Bonferroni's principle), 该原理实际上对数据挖掘的过度使用提出了警告。本章还概述了一些非常有用的思想,它们未必都属于数据挖掘的范畴,但是却有利于理解数据挖掘中的某些重要概念。这些思想包括度量词语重要性的TF.IDF权重、哈希函数及索引结构的性质、包含自然对数底 e 的恒等式等。最后,简要介绍了后续章节所要涉及的主题。

1.1 数据挖掘的定义

最广为接受的定义是,数据挖掘 (data mining) 是数据“模型”的发现过程。而“模型”却可以有多种含义。下面介绍在建模方面最重要的几个方向。

1.1.1 统计建模

最早使用“data mining”术语的人是统计学家。术语“data mining”或者“data dredging”最初是贬义词,意指试图抽取数据本身不支持的信息的过程。1.2节给出了这种挖掘情况下可能犯的几类错误。当然,现在术语“data mining”的意义已经是正面的了。目前,统计学家认为数据挖掘就是统计模型 (statistical model) 的构建过程,而这个统计模型指的就是可见数据所遵从的总体分布。

例1.1 假定现有的数据是一系列数字。这种数据相对于常用的挖掘数据而言显得过于简单,但这只是为了说明问题而采用的例子。统计学家可能会判定这些数字来自一个高斯分布 (即正态分布), 并利用公式来计算该分布最有可能的参数值。该高斯分布的均值和标准差能够完整地刻画整个分布,因而成为上述数据的一个模型。

1.1.2 机器学习

有些人将数据挖掘看成是机器学习的同义词。毫无疑问,一些数据挖掘方法中适当使用了机器学习算法。机器学习的实践者将数据当成训练集来训练某类算法,比如贝叶斯网络、支持向量机、决策树、隐马尔可夫模型等。

某些场景下上述的数据利用方式是合理的。机器学习擅长的典型场景是人们对数据中的寻找目标几乎一无所知。比如，我们并不清楚到底是影片的什么因素导致某些观众喜欢或者厌恶该影片。因此，在Netflix竞赛要求设计一个算法来预测观众对影片的评分时，基于已有评分样本的机器学习算法获得了巨大成功。在9.4节中，我们将讨论此类算法的一个简单形式。

另一方面，当挖掘的目标能够更直接地描述时，机器学习方法并不成功。一个有趣的例子是，WhizBang!实验室^①曾试图使用机器学习方法在Web上定位人们的简历。但是不管使用什么机器学习算法，最后的效果都比不过人工设计的直接通过典型关键词和短语来查找简历的算法。由于看过或者写过简历的人都对简历包含哪些内容非常清楚，Web页面是否包含简历毫无秘密可言。因此，使用机器学习方法相对于直接设计的简历发现算法而言并无任何优势。

1.1.3 建模的计算方法

近年来，计算机科学家已将数据挖掘看成一个算法问题。这种情况下，数据模型仅仅就是复杂查询的答案。例如，给定例1.1中的一系列数字，我们可以计算它们的均值和标准差。需要注意的是，这样计算出的参数可能并不是这组数据的最佳高斯分布拟合参数，尽管在数据集规模很大时两者非常接近。

数据建模有很多不同的方法。前面我们已经提到，数据可以通过其生成所可能遵从的统计过程构建来建模。而其他的大部分数据建模方法可以描述为下列两种做法之一：

- (1) 对数据进行简洁的近似汇总描述；
- (2) 从数据中抽取出最突出的特征来代替数据并将剩余内容忽略。

在接下来的内容中，我们将探究上述两种做法。

1.1.4 数据汇总

一种最有趣的数据汇总形式是PageRank，它也是使谷歌成功的关键算法之一，我们将在第5章对它进行详细介绍。在这种形式的Web挖掘当中，Web的整个复杂结构可由每个页面所对应的一个数字归纳而成。这种数字就是网页的PageRank值，即一个Web结构上的随机游走者在任意给定时刻处于该页的概率（这是极其简化的一种说法）。PageRank的一个非常好的特性就是它能够很好地反映网页的重要性，即典型用户在搜索时期望返回某个页面的程度。

另一种重要的数据汇总形式是聚类，第7章将予以介绍。在聚类中，数据被看成是多维空间下的点，空间中相互邻近的点将被赋予相同的类别。这些类别本身也会被概括表示，比如通过类别质心及类别中的点到质心的平均距离来描述。这些类别的概括信息综合在一起形成了全体数据集合的数据汇总结果。

^① 该初创实验室试图使用机器学习方法来进行大规模数据挖掘，并且雇用了大批机器学习高手来实现这一点。遗憾的是，该实验室并没有能够生存下来。