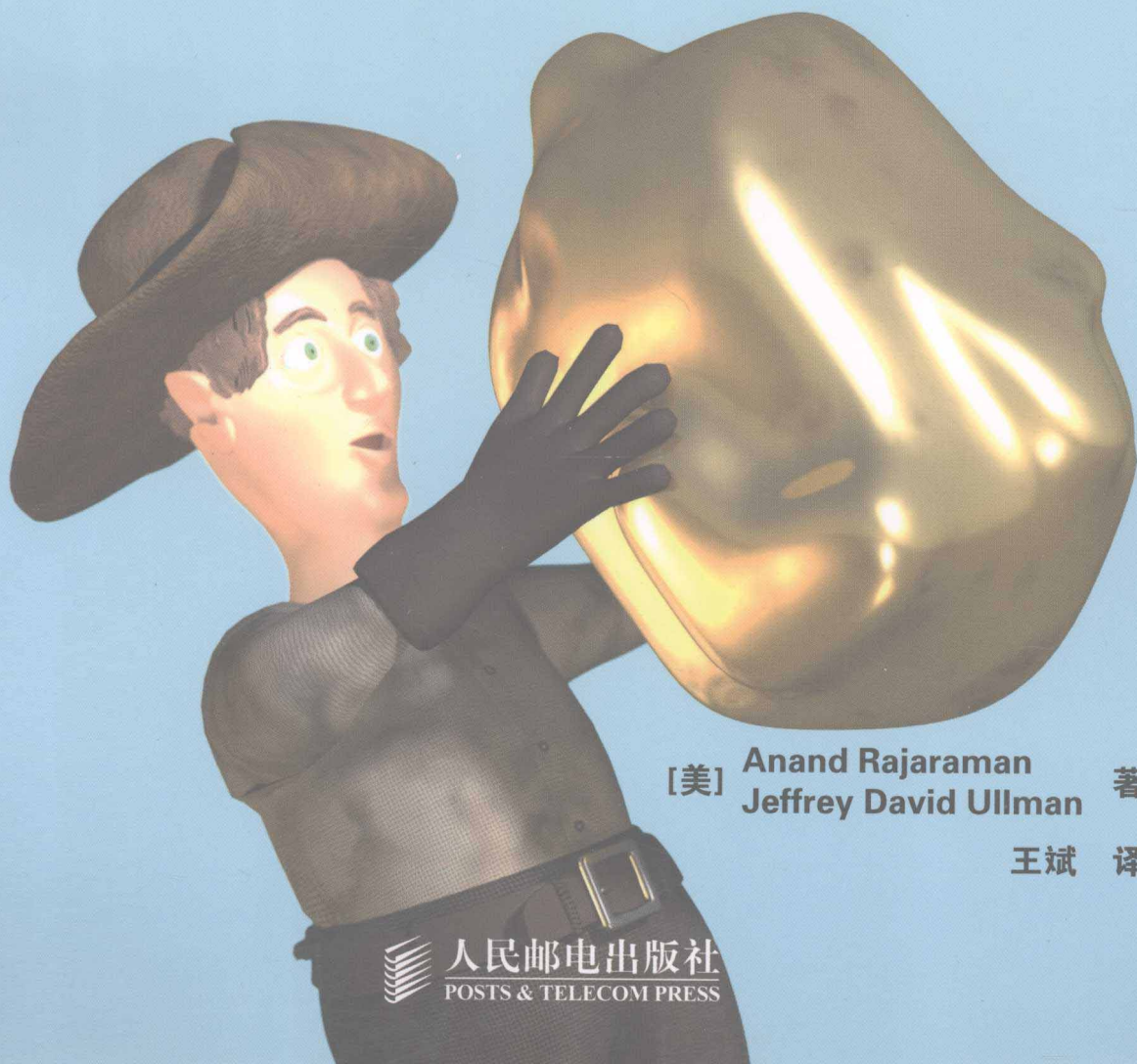


Mining of Massive Datasets

大数据

互联网大规模数据挖掘
与分布式处理



[美] Anand Rajaraman 著
Jeffrey David Ullman

王斌 译



人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Mining of Massive Datasets

大数据

互联网大规模数据挖掘
与分布式处理



[美] Anand Rajaraman 著
Jeffrey David Ullman

王斌 译

版 社

图书在版编目 (C I P) 数据

大数据 : 互联网大规模数据挖掘与分布式处理 /
(美) 拉贾拉曼 (Rajaraman, A.), (美) 厄尔曼
(Ullman, J. D.) 著 ; 王斌译. — 北京 : 人民邮电出版
社, 2012.9 (2012 11重印)
(图灵程序设计丛书)
书名原文: Mining of Massive Datasets
ISBN 978-7-115-29131-8

I. ①大… II. ①拉… ②厄… ③王… III. ①互联网
网络—数据采集②互联网络—分布式数据处理 IV.
①TP274

中国版本图书馆CIP数据核字(2012)第188384号

内 容 提 要

本书由斯坦福大学的“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。其中
相关章节有对应的习题, 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。

图灵程序设计丛书

大数据: 互联网大规模数据挖掘与分布式处理

-
- ◆ 著 [美] Anand Rajaraman [美] Jeffrey David Ullman
译 王 斌
责任编辑 卢秀丽
 - ◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮箱 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京艺辉印刷有限公司印刷
 - ◆ 开本: 800×1000 1/16
印张: 17
字数 402千字 2012年9月第1版
印数 11 001-16 000册 2012年11月北京第4次印刷
著作权合同登记号 图字: 01-2012-4002号
ISBN 978-7-115-29131-8
-

定价: 59.00元

读者服务热线: (010)51095186转604 印装质量热线: (010)67129223

反盗版热线: (010)67171154

站在巨人的肩膀上
Standing on Shoulders of Giants



www.ituring.com.cn

站在巨人的肩上
Standing on Shoulders of Giants



www.ituring.com.cn

版 权 声 明

Mining of Massive Datasets first edition (978-1-107-01535-7) by Anand Rajaraman and Jeffrey David Ullman first published by Cambridge University Press 2012.

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 2012.

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

这是继2009年翻译完《信息检索导论》之后，我翻译的第二本书。翻译完前一本书之后，我曾经下决心几年之内不再翻译书。这一方面是由于翻译书十分辛苦并且需要花费大量的时间，我怕时间和精力上难以保证。另一方面，书的翻译质量好坏会让有点完美主义倾向的我始终承受着巨大的心理压力。但是，我终究没能经受住诱惑。每次看到优秀的英文原版书籍时，都有尽快翻译成中文和国内同仁分享的冲动。而这次冲动的表现就是，我主动请缨提交试译稿，并有幸被出版社选中而再次开始了翻译的历程。

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、推荐和挖掘等三种。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的话，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。

同《信息检索导论》一样，本书的电子版也先于印刷版在斯坦福大学网站上公开。得到电子版书籍之后，我很快就看完了并且迫不及待地在课题组内进行了推广，我的很多学生都集中学习了本书。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、互联网广告及推荐系统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Map-Reduce分布式计算框架。和《信息检索导论》一书相比，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Map-Reduce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。整个翻译前后持续了七个多月，并历经多次修改。初稿完成后我发给本领域的一些专家审阅，并得到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、中科院研究生院何奔博士等人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。感谢图灵公司的武总、谢总、傅志红、卢秀丽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。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完成，因此加班便成了家常便饭。4岁的儿子心心在我每次出门前都嘱咐我路上小心，这让我感到幸福并给我力量。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。希望读者能够和我进行联系，以便能够不断改进。来信请联系wbxjj2008@gmail.com，本书勘误会及时公布在网站<http://ir.ict.ac.cn/~wangbin/mmd-book/>上。原书的初稿电子版等信息也可以从网站<http://infolab.stanford.edu/~ullman/mmds.html>下载。

王 斌

2012年7月20日于中关村

前言

本书是在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基础上总结而成的。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。

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,也就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Map-Reduce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。

先修课程

尽管从编号CS345A看,本课程属于高年级研究生课程,但是我们发现高年级本科生和低年级硕士生也能接受该课程。该课程将来可能会分配一个介于高年级研究生和低年级硕士生水平之间的编号。

CS345A的先修课程包括:

- (1) 数据库系统的首期课程,包括基于SQL及其他数据库相关语言(如XQuery)的应用编程;
- (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该课程过去提供的材料: <http://infolab.stanford.edu/~ullman/mining/mining.html>。

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

致谢

本书封面由Scott Ullman设计。感谢Foto Afrati和Arun Marathe精心阅读本书初稿并提出建设性的意见。感谢Leland Chen、Shrey Gupta、Xie Ke、Haewoon Kwak、Brad Penoff、Philips Kokoh Prasetyo、Mark Storus、Tim Triche Jr.及Roshan Sumbaly指出了本书中的部分错误。当然,剩余错误均由我们负责。

A. R.

J. D. U.

加利福尼亚州帕洛阿尔托

2011年6月

目 录

第 1 章 数据挖掘基本概念	1	2.2.2 分组和聚合	20
1.1 数据挖掘的定义	1	2.2.3 Reduce 任务	20
1.1.1 统计建模	1	2.2.4 组合器	21
1.1.2 机器学习	1	2.2.5 Map-Reduce 的执行细节	21
1.1.3 建模的计算方法	2	2.2.6 节点失效的处理	22
1.1.4 数据汇总	2	2.3 使用 Map-Reduce 的算法	22
1.1.5 特征抽取	3	2.3.1 基于 Map-Reduce 的矩阵-向量 乘法实现	23
1.2 数据挖掘的统计限制	4	2.3.2 向量 v 无法放入内存时的处理	23
1.2.1 整体情报预警	4	2.3.3 关系代数运算	24
1.2.2 邦弗朗尼原理	4	2.3.4 基于 Map-Reduce 的选择运算	26
1.2.3 邦弗朗尼原理的一个例子	5	2.3.5 基于 Map-Reduce 的投影运算	26
1.2.4 习题	6	2.3.6 基于 Map-Reduce 的并、交和差 运算	27
1.3 相关知识	6	2.3.7 基于 Map-Reduce 的自然连接 运算	27
1.3.1 词语在文档中的重要性	6	2.3.8 一般性的连接算法	28
1.3.2 哈希函数	7	2.3.9 基于 Map-Reduce 的分组和聚合 运算	28
1.3.3 索引	8	2.3.10 矩阵乘法	29
1.3.4 二级存储器	10	2.3.11 基于单步 Map-Reduce 的矩阵 乘法	29
1.3.5 自然对数的底 e	10	2.3.12 习题	30
1.3.6 幂定律	11	2.4 Map-Reduce 的扩展	31
1.3.7 习题	12	2.4.1 工作流系统	31
1.4 本书概要	13	2.4.2 Map-Reduce 的递归扩展版本	32
1.5 小结	14	2.4.3 Pregel 系统	34
1.6 参考文献	14	2.4.4 习题	35
第 2 章 大规模文件系统及 Map- Reduce	16	2.5 集群计算算法的效率问题	35
2.1 分布式文件系统	16	2.5.1 集群计算的通信开销模型	35
2.1.1 计算节点的物理结构	17	2.5.2 实耗通信开销	36
2.1.2 大规模文件系统的结构	18		
2.2 Map-Reduce	18		
2.2.1 Map 任务	19		

2.5.3 多路连接	37	3.6.2 面向 Jaccard 距离的局部敏感 函数族	66
2.5.4 习题	40	3.6.3 局部敏感函数族的放大处理	66
2.6 小结	40	3.6.4 习题	68
2.7 参考文献	42	3.7 面向其他距离测度的 LSH 函数族	68
第 3 章 相似项发现	44	3.7.1 面向海明距离的 LSH 函数族	69
3.1 近邻搜索的应用	44	3.7.2 随机超平面和余弦距离	69
3.1.1 集合的 Jaccard 相似度	44	3.7.3 梗概	70
3.1.2 文档的相似度	45	3.7.4 面向欧氏距离的 LSH 函数族	71
3.1.3 协同过滤——一个集合相似 问题	46	3.7.5 面向欧氏空间的更多 LSH 函数族	72
3.1.4 习题	47	3.7.6 习题	72
3.2 文档的 Shingling	47	3.8 LSH 函数的应用	73
3.2.1 k -Shingle	47	3.8.1 实体关联	73
3.2.2 shingle 大小的选择	48	3.8.2 一个实体关联的例子	74
3.2.3 对 shingle 进行哈希	48	3.8.3 记录匹配的验证	74
3.2.4 基于词的 shingle	49	3.8.4 指纹匹配	75
3.2.5 习题	49	3.8.5 适用于指纹匹配的 LSH 函数族	76
3.3 保持相似度的集合摘要表示	49	3.8.6 相似新闻报道检测	77
3.3.1 集合的矩阵表示	50	3.8.7 习题	78
3.3.2 最小哈希	50	3.9 面向高相似度的方法	79
3.3.3 最小哈希及 Jaccard 相似度	51	3.9.1 相等项发现	79
3.3.4 最小哈希签名	52	3.9.2 集合的字符串表示方法	79
3.3.5 最小哈希签名的计算	52	3.9.3 基于长度的过滤	80
3.3.6 习题	54	3.9.4 前缀索引	81
3.4 文档的局部敏感哈希算法	55	3.9.5 位置信息的使用	82
3.4.1 面向最小哈希签名的 LSH	56	3.9.6 使用位置和长度信息的索引	83
3.4.2 行条化策略的分析	57	3.9.7 习题	85
3.4.3 上述技术的综合	58	3.10 小结	85
3.4.4 习题	59	3.11 参考文献	87
3.5 距离测度	59	第 4 章 数据流挖掘	89
3.5.1 距离测度的定义	59	4.1 流数据模型	89
3.5.2 欧氏距离	60	4.1.1 一个数据流管理系统	89
3.5.3 Jaccard 距离	60	4.1.2 流数据源的例子	90
3.5.4 余弦距离	61	4.1.3 流查询	91
3.5.5 编辑距离	62	4.1.4 流处理中的若干问题	92
3.5.6 海明距离	63	4.2 流当中的数据抽样	92
3.5.7 习题	63	4.2.1 一个富于启发性的例子	93
3.6 局部敏感函数理论	64	4.2.2 代表性样本的获取	93
3.6.1 局部敏感函数	65		

4.2.3 一般的抽样问题.....	94	5.1.2 PageRank 的定义.....	117
4.2.4 样本规模的变化.....	94	5.1.3 Web 结构.....	119
4.2.5 习题.....	95	5.1.4 避免终止点.....	121
4.3 流过滤.....	95	5.1.5 采集器陷阱及“抽税”法.....	123
4.3.1 一个例子.....	95	5.1.6 PageRank 在搜索引擎中的 使用.....	125
4.3.2 布隆过滤器.....	96	5.1.7 习题.....	125
4.3.3 布隆过滤方法的分析.....	96	5.2 PageRank 的快速计算.....	126
4.3.4 习题.....	97	5.2.1 转移矩阵的表示.....	127
4.4 流中独立元素的数目统计.....	98	5.2.2 基于 Map-Reduce 的 PageRank 迭代计算.....	128
4.4.1 独立元素计数问题.....	98	5.2.3 结果向量合并时的组合器 使用.....	128
4.4.2 FM 算法.....	98	5.2.4 转移矩阵中块的表示.....	129
4.4.3 组合估计.....	99	5.2.5 其他高效的 PageRank 迭代 方法.....	130
4.4.4 空间需求.....	100	5.2.6 习题.....	131
4.4.5 习题.....	100	5.3 面向主题的 PageRank.....	131
4.5 矩估计.....	100	5.3.1 动机.....	131
4.5.1 矩定义.....	100	5.3.2 有偏的随机游走模型.....	132
4.5.2 二阶矩估计的 AMS 算法.....	101	5.3.3 面向主题的 PageRank 的使用.....	133
4.5.3 AMS 算法有效的原因.....	102	5.3.4 基于词汇的主题推断.....	134
4.5.4 更高阶矩的估计.....	103	5.3.5 习题.....	134
4.5.5 无限流的处理.....	103	5.4 链接作弊.....	135
4.5.6 习题.....	104	5.4.1 垃圾农场的架构.....	135
4.6 窗口内的计数问题.....	105	5.4.2 垃圾农场的分析.....	136
4.6.1 精确计数的开销.....	105	5.4.3 与链接作弊的斗争.....	137
4.6.2 DGIM 算法.....	105	5.4.4 TrustRank.....	137
4.6.3 DGIM 算法的存储需求.....	107	5.4.5 垃圾质量.....	137
4.6.4 DGIM 算法中的查询应答.....	107	5.4.6 习题.....	138
4.6.5 DGIM 条件的保持.....	108	5.5 导航页和权威页.....	139
4.6.6 降低错误率.....	109	5.5.1 HITS 的直观意义.....	139
4.6.7 窗口内计数问题的扩展.....	109	5.5.2 导航度和权威度的形式化.....	139
4.6.8 习题.....	110	5.5.3 习题.....	142
4.7 衰减窗口.....	110	5.6 小结.....	143
4.7.1 最常见元素问题.....	110	5.7 参考文献.....	145
4.7.2 衰减窗口的定义.....	111		
4.7.3 最流行元素的发现.....	111		
4.8 小结.....	112		
4.9 参考文献.....	113		
第 5 章 链接分析.....	115	第 6 章 频繁项集.....	146
5.1 PageRank.....	115	6.1 购物篮模型.....	146
5.1.1 早期的搜索引擎及词项作弊.....	115	6.1.1 频繁项集的定义.....	146

6.1.2 频繁项集的应用	148	7.2 层次聚类	179
6.1.3 关联规则	149	7.2.1 欧氏空间下的层次聚类	180
6.1.4 高可信度关联规则的发现	150	7.2.2 层次聚类算法的效率	183
6.1.5 习题	151	7.2.3 控制层次聚类的其他规则	183
6.2 购物篮及 A-Priori 算法	152	7.2.4 非欧空间下的层次聚类	185
6.2.1 购物篮数据的表示	152	7.2.5 习题	186
6.2.2 项集计数中的内存使用	153	7.3 k -均值算法	187
6.2.3 项集的单调性	154	7.3.1 k -均值算法基本知识	187
6.2.4 二元组计数	155	7.3.2 k -均值算法的簇初始化	187
6.2.5 A-Priori 算法	155	7.3.3 选择 k 的正确值	188
6.2.6 所有频繁项集上的 A-Priori 算法	157	7.3.4 BFR 算法	189
6.2.7 习题	158	7.3.5 BFR 算法中的数据处理	191
6.3 更大数据集在内存中的处理	159	7.3.6 习题	192
6.3.1 PCY 算法	160	7.4 CURE 算法	193
6.3.2 多阶段算法	161	7.4.1 CURE 算法的初始化	194
6.3.3 多哈希算法	163	7.4.2 CURE 算法的完成	195
6.3.4 习题	164	7.4.3 习题	195
6.4 有限扫描算法	166	7.5 非欧空间下的聚类	196
6.4.1 简单的随机化算法	166	7.5.1 GRGPF 算法中的簇表示	196
6.4.2 抽样算法中的错误规避	167	7.5.2 簇表示树的初始化	196
6.4.3 SON 算法	168	7.5.3 GRGPF 算法中的点加入	197
6.4.4 SON 算法和 Map-Reduce	168	7.5.4 簇的分裂及合并	198
6.4.5 Toivonen 算法	169	7.5.5 习题	199
6.4.6 Toivonen 算法的有效性分析	170	7.6 流聚类及并行化	199
6.4.7 习题	170	7.6.1 流计算模型	199
6.5 流中的频繁项计数	171	7.6.2 一个流聚类算法	200
6.5.1 流的抽样方法	171	7.6.3 桶的初始化	200
6.5.2 衰减窗口中的频繁项集	172	7.6.4 桶合并	200
6.5.3 混合方法	172	7.6.5 查询应答	202
6.5.4 习题	173	7.6.6 并行环境下的聚类	202
6.6 小结	173	7.6.7 习题	203
6.7 参考文献	175	7.7 小结	203
		7.8 参考文献	205
第 7 章 聚类	176	第 8 章 Web 广告	207
7.1 聚类技术介绍	176	8.1 在线广告相关问题	207
7.1.1 点、空间和距离	176	8.1.1 广告机会	207
7.1.2 聚类策略	177	8.1.2 直投广告	208
7.1.3 维数灾难	178	8.1.3 展示广告的相关问题	208
7.1.4 习题	179	8.2 在线算法	209

8.2.1 在线和离线算法	209	9.1.2 长尾现象	228
8.2.2 贪心算法	210	9.1.3 推荐系统的应用	230
8.2.3 竞争率	211	9.1.4 效用矩阵的填充	230
8.2.4 习题	211	9.2 基于内容的推荐	231
8.3 广告匹配问题	212	9.2.1 项模型	231
8.3.1 匹配及完美匹配	212	9.2.2 文档的特征发现	231
8.3.2 最大匹配贪心算法	213	9.2.3 基于 Tag 的项特征获取	232
8.3.3 贪心匹配算法的竞争率	213	9.2.4 项模型的表示	233
8.3.4 习题	214	9.2.5 用户模型	234
8.4 Adwords 问题	214	9.2.6 基于内容的项推荐	235
8.4.1 搜索广告的历史	215	9.2.7 分类算法	235
8.4.2 Adwords 问题的定义	215	9.2.8 习题	237
8.4.3 Adwords 问题的贪心方法	216	9.3 协同过滤	238
8.4.4 Balance 算法	217	9.3.1 相似度计算	238
8.4.5 Balance 算法竞争率的一个下界	217	9.3.2 相似度对偶性	241
8.4.6 多投标者的 Balance 算法	219	9.3.3 用户聚类 and 项聚类	242
8.4.7 一般性的 Balance 算法	220	9.3.4 习题	243
8.4.8 Adwords 问题的最后论述	221	9.4 降维处理	243
8.4.9 习题	221	9.4.1 UV 分解	244
8.5 Adwords 的实现	221	9.4.2 RMSE	244
8.5.1 投标和搜索查询的匹配	222	9.4.3 UV 分解的增量式计算	245
8.5.2 更复杂的匹配问题	222	9.4.4 对任一元素的优化	247
8.5.3 文档和投标之间的匹配算法	223	9.4.5 一个完整 UV 分解算法的构建	248
8.6 小结	224	9.4.6 习题	250
8.7 参考文献	226	9.5 NetFlix 竞赛	250
第 9 章 推荐系统	227	9.6 小结	251
9.1 一个推荐系统的模型	227	9.7 参考文献	253
9.1.1 效用矩阵	227	索引	254

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,意指试图抽取出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。