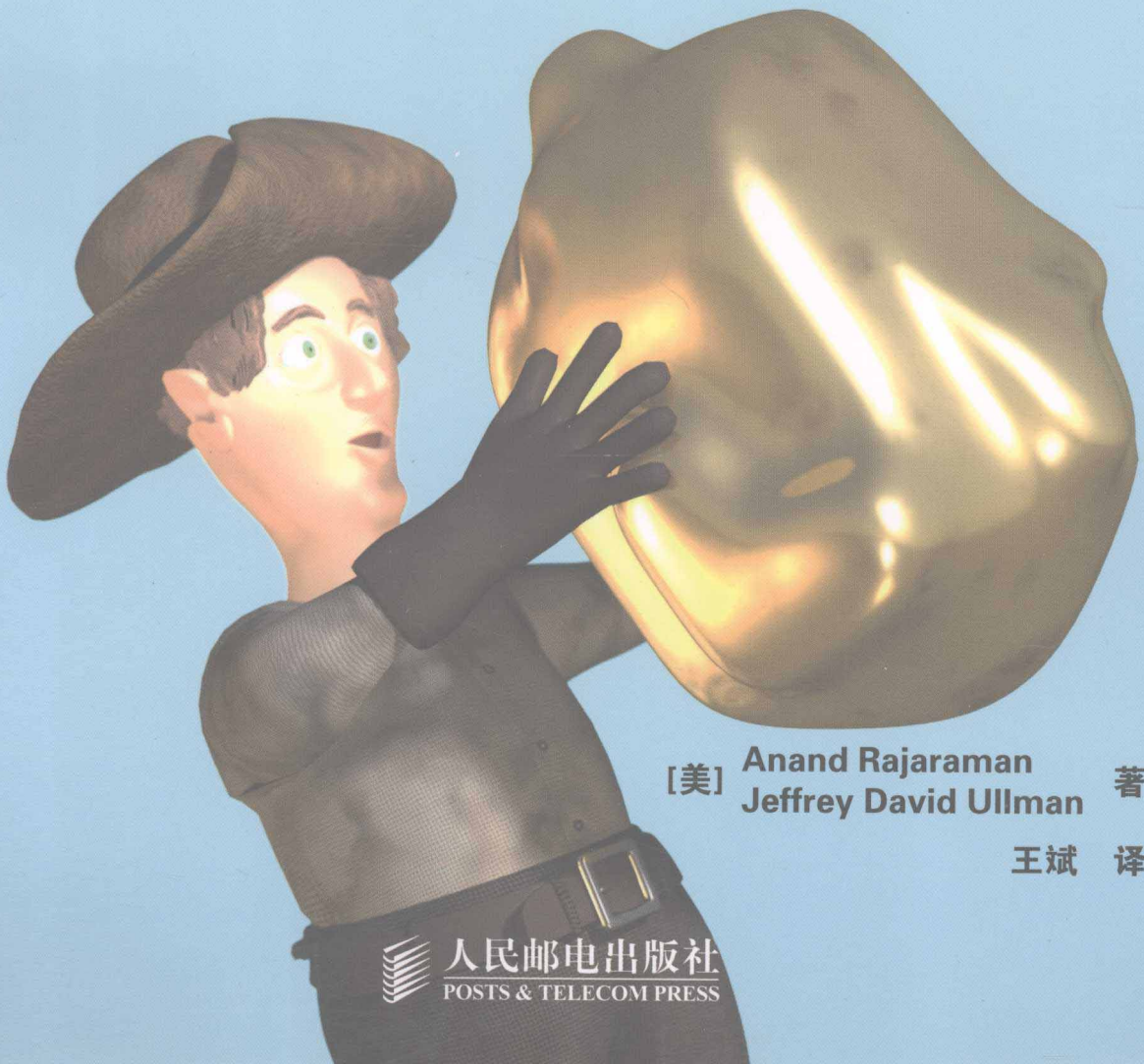
TURING 图灵程序设计丛书

CAMBRIDGE

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
Jeffrey David Ullman 著

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
Jeffrey David Ullman 著

王斌 译

版 社

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 /
(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
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 电子邮件 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

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，其主要的应用形式包括搜索、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、中科院研究生院何苯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

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。