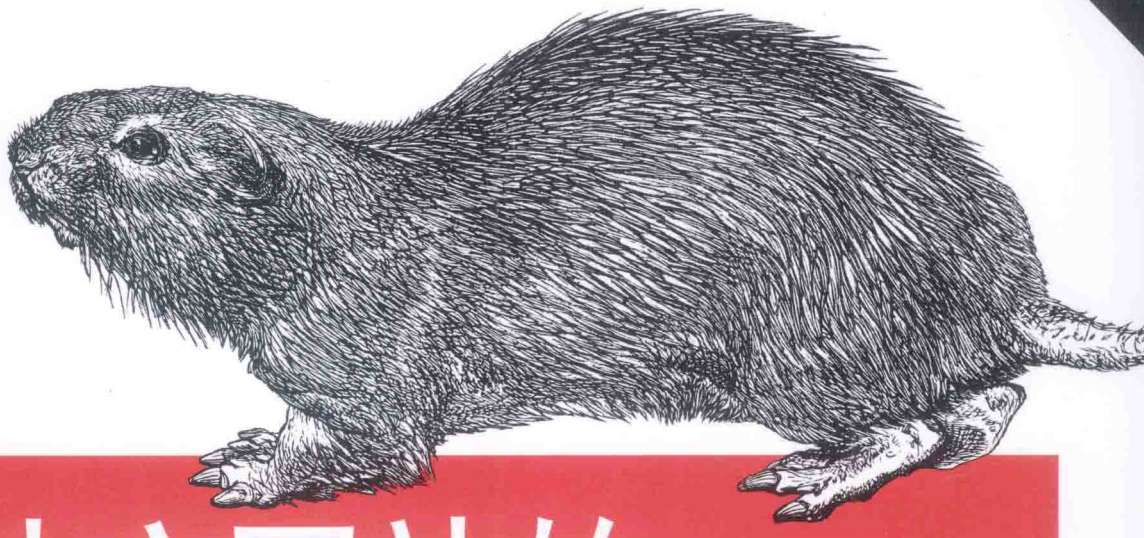
O'REILLY®

原书第2版

HZ BOOKS
华章IT



社交网站的数据挖掘与分析

MINING THE SOCIAL WEB

Matthew A. Russell 著
苏统华 魏通 赵逸雪 王烁行 刘智月 译

机械工业出版社
China Machine Press

社交网站的数据挖掘与分析

原书第2版

社交网站的数据挖掘与分析

Matthew A. Russell 著

苏统华 魏通 赵逸雪 王烁行 刘智月 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

社交网站的数据挖掘与分析（原书第2版）/（美）拉塞尔（Russell, M. A.）著；苏统华等译. —北京：机械工业出版社，2015.1

（O'Reilly精品图书系列）

书名原文：Mining the Social Web, Second Edition

ISBN 978-7-111-48699-2

I. 社… II. ①拉… ②苏… III. 数据采集 IV. TP274

中国版本图书馆CIP数据核字（2014）第281411号

北京市版权局著作权合同登记

图字：01-2014-0341号

Copyright ©2014 by Matthew A. Russell.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and China Machine Press, 2015. Authorized translation of the English edition, 2014 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由O'Reilly Media, Inc. 出版2014。

简体中文版由机械工业出版社出版 2015。英文原版的翻译得到O'Reilly Media, Inc.的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc.的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

封底无防伪标均为盗版

本书法律顾问

北京大成律师事务所 韩光/邹晓东

书　　名/　社交网站的数据挖掘与分析（原书第2版）

书　　号/　ISBN 978-7-111-48699-2

责任编辑/　秦健

封面设计/　Karen Montgomery，张健

出版发行/　机械工业出版社

地　　址/　北京市西城区百万庄大街22号（邮政编码 100037）

印　　刷/　藁城市京瑞印刷有限公司

开　　本/　178毫米×233毫米　16开本　24.25印张

版　　次/　2015年1月第1版　2015年1月第1次印刷

定　　价/　79.00元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010)88378991；88361066

购书热线：(010)68326294；88379649；68995259

投稿热线：(010)88379604

读者信箱：hzjsj@hzbook.com

原书第2版

社交网站的数据挖掘与分析

此为试读,需要完整PDF请访问: www.ertongbook.com

斧子钝了，其刃不再锋利，必多费力气；但得智慧指教，方可奏效。

——《传道书》

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

让你决定捧起这本书的，极可能是这本书的主题。毋庸置疑，社交网络已经深入人心。社交网络引发了交往方式的历史性变化，这是当今时代最大的一次变革。Facebook、Twitter、LinkedIn、Google+、GitHub等社交网站的非凡意义在于改变了互联网的生态，将自发、无序状态的互联网引入到一个有组织、有阶层甚至实名制的社交网络之中。

社交网站数据蕴藏着大量的价值和见解。这些数据随着岁月流逝，会如同醇酒一般越发芳香。社交网络与物联网也在不断融合，大数据为统计学习方法提供了广阔的舞台。随着数据爆炸不断升级，庞杂的大数据中夹杂的噪声有可能被放大。只有利用合适的工具，才能准确而快速地挖掘出感兴趣的或者有意义的知识。

Matthew A. Russell是挖掘社交网络数据的资深专家。他深谙数据挖掘的各种工具和技术，同时他对数据挖掘初学者的所需所求了如指掌。你正在阅读的这本书是在第1版基础上作了“重大更新”的第2版，其中吸收了第1版读者的大量有建设性的意见。如果你对社交网站数据感兴趣，那么本书是帮助你快速入门的法宝。本书体例完备、特色鲜明，从实用的角度出发，对主流的各种社交网站做了较全面覆盖。本书各个章节之间也保持着一定的独立性，如果你只对特定章节的技术感兴趣，也可以直接跳到对应内容阅读。不论你关注的是Facebook、Twitter、LinkedIn、Google+、GitHub、邮箱、网页还是语义网，本书都可以传授给你爬取数据、分析数据以及展示数据的技术。特别值得一提的是，本书配套的代码借助了IPython Notebook，让你可以快速配置自己的开发环境并享受交互式学习的乐趣。强烈建议你配合书中的虚拟机来学习！

本书的翻译由苏统华全程组织。魏通、赵逸雪、王烁行以及刘智月协助苏统华完成了全

书的译文初稿。其中本书的前言、前四章的初稿由魏通、王烁行和苏统华共同完成，接着的第5章到第7章的初稿由赵逸雪和苏统华共同翻译，第8章的初稿由刘智月和苏统华完成，第9章的初稿由赵逸雪和苏统华协作完成，最后的附录由王烁行和苏统华共同翻译。在初稿的基础上，翻译团队进行了交叉核对，并最终由苏统华统一定稿。

本书从启动翻译到进入出版流程历时整整一年，在此过程中，得到很多同事、朋友和编辑团队的热心帮助，在此表达我们深深的谢意。另外，本书的翻译还得到了多个项目的资助，在此一并致谢。国家自然科学基金（资助号：61203260）、黑龙江省科研启动基金（资助号：LBH-Q13066）、哈尔滨工业大学科研创新基金（资助号：HITNSRIF2015083）对本书的翻译提供了部分资助。最后，黑龙江省教育厅高等教育教学改革项目（资助号：JG2013010224）、哈尔滨工业大学研究生教育教学改革研究项目（资助号：JCJS-201309）也对本书的翻译提供了大力支持。

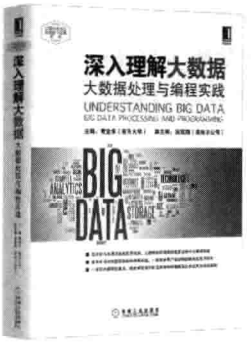
本书涉及的技术较广，鉴于译者水平有限，译文中难免存在一些问题，真诚地希望读者朋友将你的意见发到译者邮箱 tonghuasu@gmail.com。

苏统华
哈尔滨工业大学软件学院
2014年10月20日

译者简介

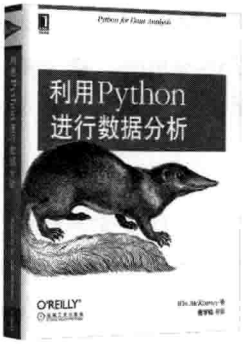
苏统华 博士，硕士生导师，CUDA研究中心以及教学中心负责人。主要研究方向包括：物联网大数据智能信息处理、大规模并行计算、模式识别、智能媒体交互与计算等。作为自然手写中文文本识别的开拓者，四年内代表工作被同行大篇幅他引约300次；他所建立的HIT-MW库为全世界100多家科研院所采用；目前负责国家自然科学基金项目2项。2013年，他领导的研究组在文档分析和识别国际会议（ICDAR’2013）上获得手写汉字识别竞赛的双料冠军；2014年，两项手写文字识别核心技术授权给某高新技术公司，正在为超过200万终端用户提供技术服务。著有英文专著《Chinese Handwriting Recognition: An Algorithmic Perspective》（德国施普林格出版社），出版5本大数据分析方面的译作（机械工业出版社）。

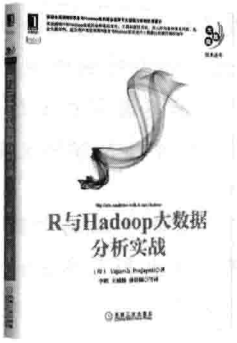
推荐阅读


深入理解大数据
大数据处理与编程实践
UNDERSTANDING BIG DATA
BIG DATA
ANALYTICS
STORAGE

实用数据分析

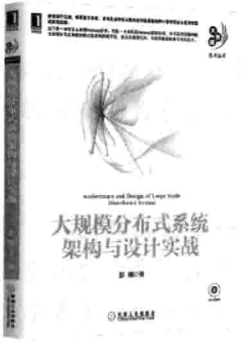
Clojure数据分析秘笈

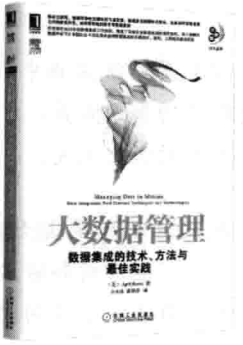

Python for Data Analysis
利用Python
进行数据分析
O'REILLY


R与Hadoop大数据
分析实战

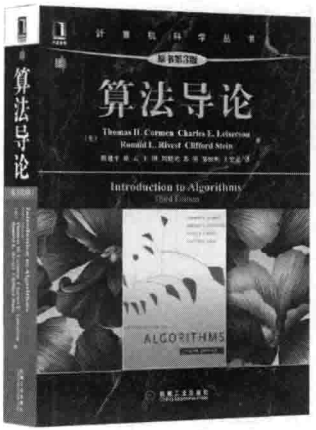
Hadoop集群与安全

机器学习实践指南
案例应用解析


大规模分布式系统
架构与设计实战


大数据管理
数据集成的技术、方法与
最佳实践

推荐阅读

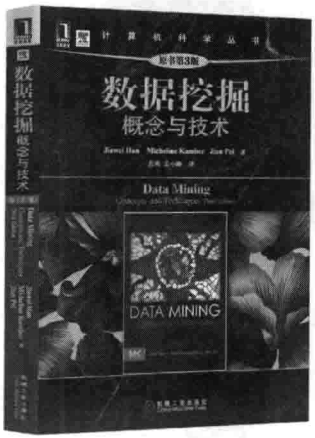


算法导论（原书第3版）

作者：Thomas H. Cormen 等 译者：殷建平 等 ISBN：978-7-111-40701-0 定价：128.00元

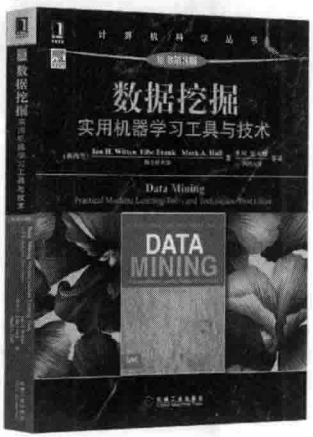
机器学习

作者：Tom Mitchell 译者：曾华军 等 ISBN：978-7-111-10993-7 定价：35.00元



数据挖掘：概念与技术（原书第3版）

作者：Jiawei Han 等 译者：范明 等 ISBN：978-7-111-39140-1 定价：79.00元



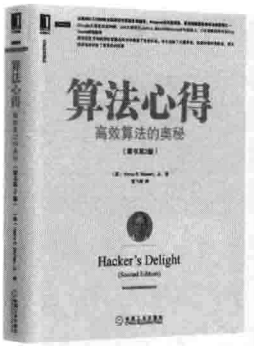
数据挖掘：实用机器学习工具与技术（原书第3版）

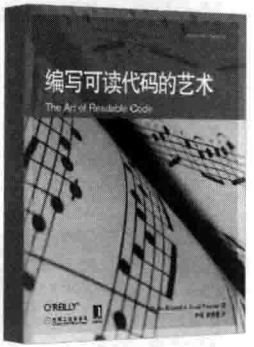
作者：Ian H. Witten 等 译者：李川 等 ISBN：978-7-111-45381-9 定价：79.00元

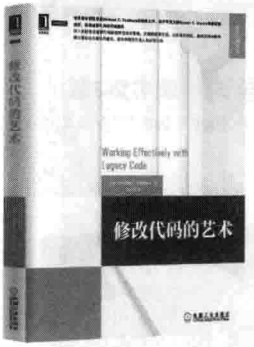
此为试读，需要完整PDF请访问：www.ertongbook.com

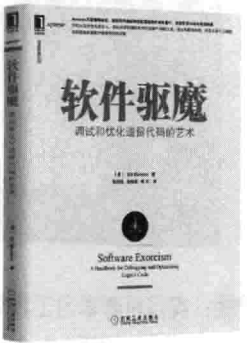
推荐阅读

Mastering Algorithms with C
算法精解
C语言描述
O'REILLY


算法心得
高效算法的奥秘
Hacker's Delight


编写可读代码的艺术
The Art of Readable Code
O'REILLY


Working Effectively with Legacy Code
修改代码的艺术


软件驱魔
调试和优化遗留代码的艺术
Software Exorcism

代码之殇
（原书第2版）

目录

前言 1

第一部分 社交网络导引

序幕 13

第1章 挖掘Twitter：探索热门话题、发现人们的谈论内容等 15

1.1 概述 15
1.2 Twitter风靡一时的原因 16
1.3 探索Twitter API 18
1.4 分析140字的推文 33
1.5 本章小结 47
1.6 推荐练习 48
1.7 在线资源 48

第2章 挖掘Facebook：分析粉丝页面、查看好友关系等 50

2.1 概述 51
2.2 探索Facebook的社交图谱API 51
2.3 分析社交图谱联系 62

2.4 本章小结 85
2.5 推荐练习 86
2.6 在线资源 86

第3章 挖掘LinkedIn：分组职位、聚类同行等 88

3.1 概述 89
3.2 探索LinkedIn API 89
3.3 数据聚类速成 94
3.4 本章小结 124
3.5 推荐练习 125
3.6 在线资源 126

第4章 挖掘Google+：计算文档相似度、提取搭配等 127

4.1 概述 128
4.2 探索Google+ API 128
4.3 TF-IDF简介 138
4.4 用TF-IDF查询人类语言数据 145
4.5 本章小结 164
4.6 推荐练习 165
4.7 在线资源 165

第5章 挖掘网页：使用自然语言处理理解人类语言、总结博客内容等 167

5.1 概述 168
5.2 抓取、解析、爬取网页 168
5.3 通过解码语法来探索语义 174
5.4 以实体为中心的分析：范式转换 192
5.5 人类语言数据处理分析的质量 200
5.6 本章小结 203
5.7 推荐练习 203
5.8 在线资源 204

第6章 挖掘邮箱：分析谁和谁说什么以及说的频率等 206
6.1 概述 207
6.2 获取和处理邮件语料库 207
6.3 分析Enron语料库 225
6.4 探索和可视化时序趋势 241
6.5 分析你自己的邮件数据 244
6.6 本章小结 250
6.7 推荐练习 251
6.8 在线资源 251
第7章 挖掘GitHub：检查软件协同习惯、构建兴趣图谱等 253
7.1 概述 254
7.2 探索GitHub的API 254
7.3 使用属性图为数据建模 260
7.4 分析GitHub兴趣图谱 264
7.5 本章小结 286
7.6 推荐练习 287
7.7 在线资源 287
第8章 挖掘带标记语义网：提取微格式、推断资源描述框架等 289
8.1 概述 290
8.2 微格式：易于实现的元数据 290
8.3 从语义标记过渡到语义网：一个小插曲 304
8.4 语义网：发展中的变革 304
8.5 本章小结 310
8.6 推荐的练习 311
8.7 在线资源 311

第二部分 Twitter实用指南

第9章 Twitter实用指南 317

9.1 访问Twitter的API（开发目的） 318
9.2 使用OAuth访问Twitter的API（产品目的） 319
9.3 探索流行话题 323
9.4 查找推文 324
9.5 构造方便的函数调用 325
9.6 使用文本文件存储JSON数据 326
9.7 使用MongoDB存储和访问JSON数据 327
9.8 使用信息流API对Twitter数据管道抽样 329
9.9 采集时序数据 330
9.10 提取推文实体 332
9.11 特定的推文范围内查找最流行的推文 333
9.12 特定的推文范围内查找最流行的推文实体 335
9.13 对频率分析制表 336
9.14 查找转推了状态的用户 337
9.15 提取转推的属性 339
9.16 创建健壮的Twitter请求 340
9.17 获取用户个人资料信息 343
9.18 从任意的文本中提取推文实体 344
9.19 获得用户所有的好友和关注者 345
9.20 分析用户的好友和关注者 347
9.21 获取用户的推文 348
9.22 爬取好友关系图 350
9.23 分析推文内容 351
9.24 提取链接目标摘要 353
9.25 分析用户收藏的推文 356
9.26 本章小结 357
9.27 推荐练习 358
9.28 在线资源 359

第三部分 附录

附录A 关于本书虚拟机体验的信息 363

附录B OAuth入门 364

附录C Python和IPython Notebook的使用技巧 368

此为试读,需要完整PDF请访问: www.ertongbook.com