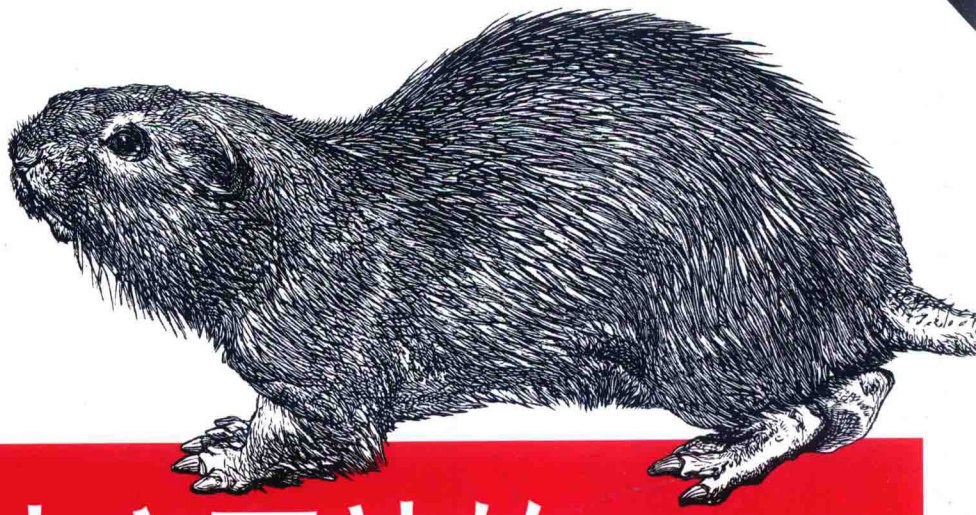
O'REILLY®

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

械工业出版社
ina Machine Press

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

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

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）。

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

前言

与其说网络是一项技术创新，不如说它是一项社交创举。
我设计它意在延伸社交性（帮助大家一起工作），
而不是为了制造一种高科技玩具。网络的终极目标是
支持并改进现实世界的网络化生存。现实世界中，
我们会组成家庭、组织协会、组建公司。现实世界中，
我们会跨越空间的樊篱建立信任，亦会近在咫尺却心生芥蒂。
——Tim Berners-Lee（万维网之父），
《Weaving the Web》（Harper）

读者必读

本书经过精心设计，为特定的目标受众提供一段难以忘怀的学习体验。那些影响心情的电子邮件、糟糕的书评或者其他误导，可能让你对本书的范围和目的产生不必要的误解。为了避免这些混乱，本前言的余下内容试图帮助你确定你是否是该书的目标受众。作为一位非常繁忙的职场人士，我认为时间是我最宝贵的财富，并且我认为对你也是这样的。尽管我经常遭遇失败，但是当我走出困境时，我真的尽力向我的邻居致敬。本前言是我试图向你（读者）致敬，我的致敬方式是清楚阐释本书能否满足你的期望。

管理你的期望

首先，本书假设你希望学习如何挖掘来自流行社交网络资源中的数据，避免在运行示例代码时遇到技术问题并且在过程中获得很多乐趣。尽管你读这本书仅仅可能是为了了解社交网络挖掘可能做什么事情，但你应该知道本书的写作风格。本书组织成让你可以跟随本书尝试许多练习，并且一旦完成了一些安装开发环境的简单步骤就能进入数据挖掘

者的行列。如果你以前编写过一些程序，应该会发现可以轻松运行这些示例代码。即使你以前从未编过程但认为自己的技术领悟能力还可以，我敢说你可以将这本书作为一次难忘旅程的出发点，它将以你从未想象的方式扩展你的思想。

为了充分享受本书及其所提供的内容，你需要对挖掘流行社交网络（如Twitter、Facebook、LinkedIn和Google+）存储数据的广阔可能性很感兴趣，需要主动下载一个虚拟机并且在IPython Notebook上重现本书的示例代码。IPython Notebook是一个奇妙的基于网络的工具，每一章的示例代码都基于它。执行这些代码通常和在键盘上按一些键一样容易，因为所有的代码都是以友好的用户接口呈现的。本书将会教你一些乐于学习的事物，并且在你的工具箱中加入了一些独立的工具，但是可能更加重要的是，它将会告诉你一个故事并且在途中给你带来快乐。这是一个关于与社交网站相关的数据科学的故事，它向你展示这些网站堆积的数据以及一些你能够使用这些数据做到的诱人潜力。

如果从头到尾阅读本书，你会注意到这个故事是按照章节顺序展开的。尽管每一章会大体遵循一个容易理解的模式来介绍一种社交网站、教你如何使用它的API获取数据，并且介绍一些分析数据的技术，该书讲述的内容会越来越广泛，同时也会越来越复杂。本书的前几章花一些时间介绍基本概念，然而后面的章节将会系统地建立在前几章的基础上并且逐渐介绍一系列挖掘社交网络的工具和技术，你可以将其应用到你生活的其他方面。

一些最流行的社交网站最近几年已经从流行转变到主流再转变到家喻户晓，它们改变着我们线上和线下的生活方式，它能够让技术给我们呈现出最好的（有时是最坏的）一面。综合来说，本书的每一章都将社交网站与数据挖掘、分析和可视化技术的内容组织在一起，以探索数据并且回答以下典型的问题：

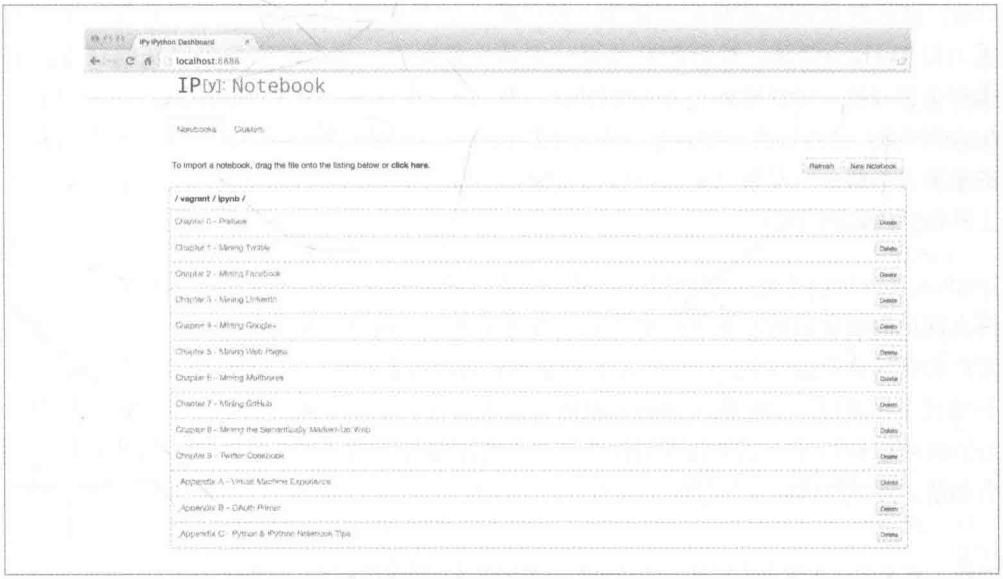
- 谁与谁相识，哪些人是他们社交网络中共有的？
- 某人与其他人交流有多频繁？
- 哪一个社交网络关系为特定的领域产生了最大的价值？
- 在网络世界里，地理位置是如何影响你的社会关系的？
- 谁是某个社交网络里最有影响力的人（最流行的人）？
- 人们在谈论些什么（这个有价值吗）？
- 基于人们在数字世界使用的人类语言，人们感兴趣的是什么？

这些基本问题的答案经常会产生一些有价值的见解，并且为企业家、社会科学家以及其他急于理解一个问题空间并且寻找解决方案的实践者展现盈利的机会。从零开始构建一个一站式的杀手级应用程序（killer app）来回答这些问题，探索远远超出经典可视化

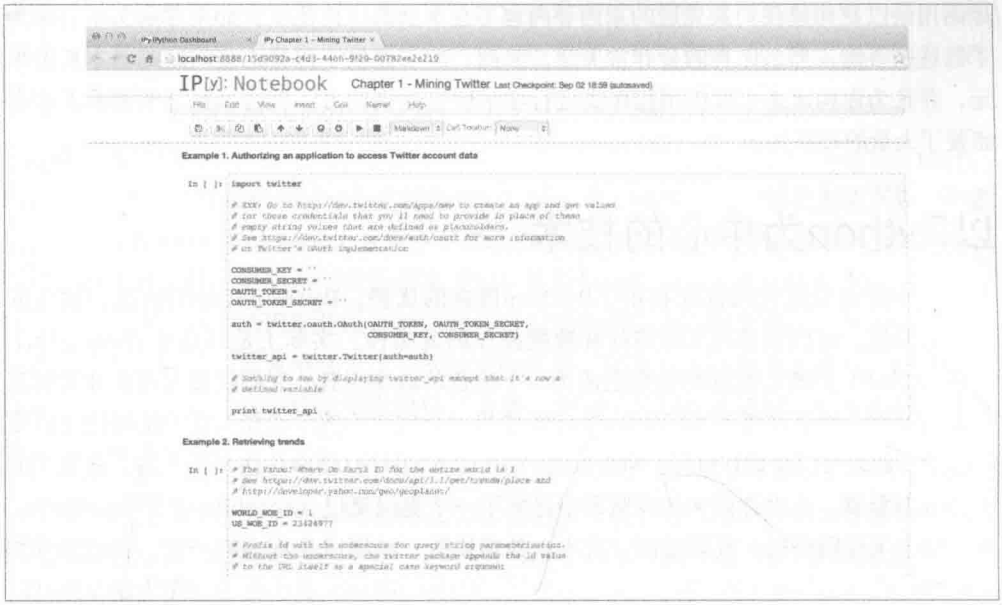
库的用法以及构建任何最先进的东西等内容不在本书的讨论范围。如果你购买本书是为了做这些事情，那么你真的会非常失望。然而，本书提供了回答这些问题的基本构造单元，并且为你构建杀手级应用程序或进行学术研究提供助力。自己浏览几章看看，本书涵盖了大量的必备知识。

以Python为中心的技术

本书中的所有示例代码特意利用了Python语言的优势。Python直观的语法、迷人的包生态系统，可以最小化API访问和数据操作的复杂性。实际上是JSON（*http://bit.ly/1a1kFaF*）的核心数据结构使它成为一个出色的教学工具，它不仅强大而且非常容易启动和运行。如果这还不足以让Python既成为一个伟大的教学选择又成为挖掘社交网络的选择，那么可以借助IPython Notebook（*http://bit.ly/1a1kFr4*）这个强大的、交互式的Python解释器，它在你的Web浏览器中提供了一个类似笔记（notebook）的用户体验，并且结合了代码执行、代码输出、文本、数学排版、绘图以及更多的功能。我们很难想到更好用户体验的学习环境，因为它将提供样例代码的麻烦最小化了，作为读者你可以跟着它一起执行代码而不会遇到任何麻烦。图P-1提供了一个IPython Notebook体验的图示，它显示了书中每一章Notebook的精简展示（dashboard）。图P-2显示了其中一个Notebook的视图。



图P-1：IPython Notebook概览，其中为Notebook的精简展示



图P-2：IPython Notebook概览，其中为“Chapter 1-Mining Twitter”（本书英文版第1章）的Notebook

书中的每一章都对应一个附带示例代码的IPython Notebook。这使得学习代码、修改bug、按照自己的目的自定义成为一种乐趣。如果你编写过一些程序但是却从来没有看到过Python语法，提前浏览几页一定是你需要的。优秀的文档可以在线获得，如果你正在寻找一个权威的Python编程语言的介绍，那么官方的Python教程（*http://bit.ly/1a1kDj8*）是很好的一个起点。本书的Python源代码使用Python 2.7编写，它是2.x系列的最新发行版。（尽管可能会碰到点问题，但我们不难想象可以使用一些自动化工具向上转换到Python 3。）

IPython Notebook无疑是很好用的，但是如果你刚接触Python编程，那么仅仅建议你跟随网上的说明配置你的开发环境可能会有些适得其反（甚至可能是无理的）。为了使你尽可能愉快地体验这本书，一个一站式的虚拟机可能会更合适，它包含IPython Notebook并预装了你重现本书示例所需的其他所有要求条件。你需要做的就是按部就班，大约15分钟就可以运行了。如果你有编程背景，你将能够配置自己的开发环境，但是我希望你会相信：虚拟机体验是更好的出发点。

注意： 更多关于本书虚拟机体验的详细信息见附录A。附录C同样也值得你注意：它提供了一些IPython Notebook的提示以及本书源代码中使用的常见Python编程惯用法。

无论你是一位Python新手还是高手，本书最新修复bug的源代码以及附带的用于构建虚