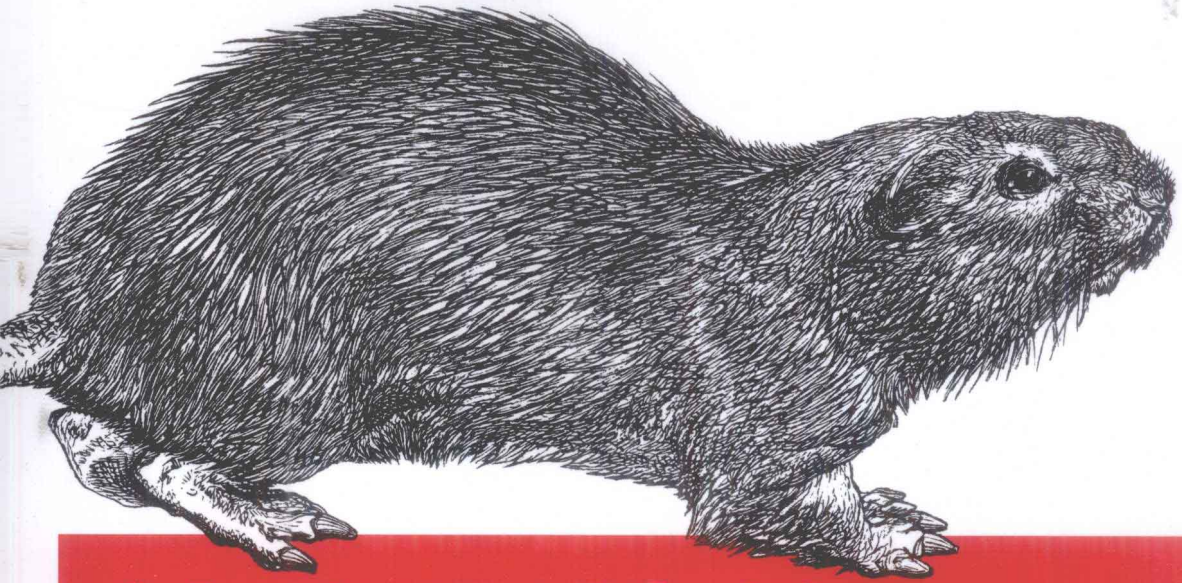
Mining the Social Web

Jolt生产效率大奖获奖图书



社交网站的数据挖掘与分析

REILLY®

机械工业出版社
China Machine Press

HZ BOOKS

Matthew A. Russell 著

师蓉 译

社交网站的数据挖掘与分析

Matthew A. Russell 著

师蓉 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/（美）罗塞尔（Russell, M. A.）著；师蓉译.
—北京：机械工业出版社，2012.1
(O'Reilly精品图书系列)

书名原文：Mining the Social Web
ISBN 978-7-111-36960-8

I. 社… II. ①罗… ②师… III. 数据采集 IV. TP274
中国版本图书馆CIP数据核字（2011）第280179号

北京市版权局著作权合同登记
图字：01-2011-1505号

书　　名/　社交网站的数据挖掘与分析
书　　号/　ISBN 978-7-111-36960-8
责任编辑/　谢晓芳
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/　北京京师印务有限公司
开　　本/　178毫米×233毫米　16开本　19.75印张
版　　次/　2012年2月第1版　2012年5月第3次印刷
定　　价/　59.00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

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

本书赞誉

“这是一本非读不可的书，因为现在的数据都散落在各处，令人眼花缭乱。[Matthew] Russell这位API达人、社交媒体专家（当然他更像是数据方面的科学家）在社交媒体数据挖掘方面巧妙地开创了先河。”

——Nick Ducoff，Infochimps公司CEO

“这是一本能让你了解新一代在线数据资源挖掘技术的重要指南。Russel做了件很了不起事，他为社交网站的从业人员准备了一本通俗易读的操作手册，涵盖了如何存取数据以及如何从原始数据中提取有价值的信息的简单方法。”

——Pete Warden，OpenHeatMap.com创始人

“每当我参与的项目涉及社交数据分析时，本书肯定是我的必备参考书。书中有很多有用的示例，如果你正开发与数据挖掘相关的项目，那么我郑重向你推荐本书。本书不仅对初学者有用，对数据挖掘的资深人士也极具参考价值。”

——Abe Music，Zaffra公司总裁

“作者编写本书时肯定乐在其中。他很巧妙地将传统的文本、图形挖掘库与时下流行的社交媒体应用联系到了一起。其示例既言之有物又不失简洁性，不但为读者提供了有价值的建议，也为读者进行更深入的开发和探讨提供了帮助。不论是对于刚涉足社交网络数据挖掘的新手，还是需要了解最新社交媒体API的资深研究人员来说，本书都是一本很好的指南。”

——Chris Augeri，Nebraska大学高级研究员

“对任何想涉足社交数据挖掘的人来说，这都一本非常好的书。作者进行了深入研究，从第1章开始就提供了丰富的示例。它不但易懂而且很有趣。如果你对社交网络数据的挖掘、分析和可视化感兴趣的话，那么本书是你的首选。”

——Jeffrey Humphries博士；计算机科学家

“在未来几年中，几乎没有任何事情可以阻碍我们通过软件自动了解人际交往方式的脚步。这个话题广博而精深。它一直是众多学术论文和学位论文的研究主题。Matthew真的是将一些容易忽略的东西联系到了一起：他对一系列庞杂而深奥的技术以及埋没在社交网站内的人际沟通方面的知识进行了很实用的讲解。本书出自一位技术高人之手——他对新工具的编程技术了如指掌。本书将为你打开软件未来十年走向的大门。”

——Tim Estes，Digital Reasoning Systems公司创始人、CEO

“本书告诉你如何最有效地利用Twitter API。”

——Raffi Krikorian，Twitter公司平台服务组

“Matthew精心挑选的这些数据源、分析技术、数据管理工具以及可视化方面的话题都非常有趣，全面体现了“如何从社交网络获取有价值的信息”领域的最新思想。他举的例子很生动，是进行深入研究很好的起点。Matthew非常在乎读者能否理解这些材料；本书随处都给出了适时、适用且真正有帮助的提示和建议。本书能激发我深入研究数据进而分析这一领域的兴趣。”

——Roger Magoulas，O’Reilly Media公司市场研究总监

译者序

作为译者，按图书原貌准确转述作者意图即是对图书最有力的推介。借写译者序的机会，表达一下本人对社交媒体的看法。

近年来，Twitter、Facebook、LinkedIn这些社交媒体开始蓬勃发展，且爆发出令人炫目的能量。无论是对消费者、企业或销售商，社交媒体都是一个很热门的话题，人们可以在社交媒体中看到很多最新的报道，了解世界各地的局势，推广新产品，告诉好友自己的最新状况，知道很久未曾谋面的好友的状况。人们花在社交媒体中的时间越来越多，它逐渐改变了人们的生活、交流方式，它传播的信息已经成为人们浏览互联网的重要内容，它制造了人们在社交生活中争相讨论的一个又一个热门话题，如因在微博中炫富而被广泛关注的“郭美美”事件，虽然我无意于纠缠该事件的来龙去脉，但是也要感叹一下社交媒体的无穷力量。

当我们的生活因为社交媒体变得便捷时，也遇到了很多问题。因为随之而来的是互联网中庞杂的数据和一些毫无价值的信息。在互联网中花费了大把时间仍然无法找到有用信息的情况时有发生，所以，如何有效地利用社交媒体，如何读取数据，如何从原始数据中提取出有价值的信息便成为亟待解决的问题。

这时，Matthew编写的这本书便成为你应该拥有的一本书。对于任何想要涉足社交数据挖掘分析的人来说，本书都是你的必备参考书。本书不仅介绍了挖掘数据、分析数据的技术，还提供了对数据管理工具以及可视化方面的话题的讨论。此外，为了能让读者充分了解这些技术，Matthew还提供了简单易懂的示例和非常有价值的建议。

本书前两章介绍了进行数据挖掘所需要的基本工具和背景知识，第3~9章讨论了对流行社交媒体数据的挖掘、分析和可视化，第10章对语义网进行了简单讨论。

很荣幸能翻译Matthew大师的作品，虽然我知道我的语言可能无法完美地展现本书的精彩之处，但我还是要庆祝经过自己的努力，不够完美但勤恳认真地完成了本书的翻译工作。

在本书的翻译过程中，特别感谢樊旺斌提供的帮助，感谢陈钢、田思源审校了部分译稿，感谢他们的鼓励和严谨。

师蓉

目录

前言 1

第1章 绪论：Twitter数据的处理 9

Python开发工具的安装 9

Twitter数据的收集和处理 11

小结 24

第2章 微格式：语义标记和常识碰撞 26

XFN和朋友 27

使用XFN来探讨社交关系 29

地理坐标：兴趣爱好的共同主线 37

（以健康的名义）对菜谱进行交叉分析 41

对餐厅评论的搜集 43

小结 45

第3章 邮箱：虽然老套却很好用 47

mbox：Unix的入门级邮箱 48

mbox+CouchDB=随意的Email分析 54

将对话线程化到一起 70

使用SIMILE Timeline将邮件“事件”可视化 79

分析你自己的邮件数据 82

小结 84

第4章 Twitter：朋友、关注者和Setwise操作.....85
REST风格的和OAuth-Cladded API.....86
干练而中肯的数据采集器.....90
友谊图的构建.....108
小结.....116
第5章 Twitter：tweet，所有的tweet，只有tweet.....118
笔PK剑：和tweet PK机枪（?!?）.....118
对tweet的分析（每次一个实体）.....121
并置潜在的社交网站（或#JustinBieber VS #TeaParty）.....144
对大量tweet的可视化.....155
小结.....163
第6章 LinkedIn：为了乐趣（和利润？）将职业网络聚类.....164
聚类的动机.....165
按职位将联系人聚类.....167
获取补充个人信息.....183
从地理上聚类网络.....188
小结.....192
第7章 Google Buzz：TF-IDF、余弦相似性和搭配.....194
Buzz=Twitter+博客（???）.....195
使用NLTK处理数据.....198
文本挖掘的基本原则.....201
查找相似文档.....208
在二元语法中发Buzz.....215
利用Gmail.....221
在中断之前试着创建一个搜索引擎…….....225
小结.....226

第8章 博客及其他：自然语言处理（等）.....228
NLP：帕累托式介绍.....228
使用NLTK的典型NLP管线.....231
使用NLTK检测博客中的句子.....234
对文件的总结.....237
以实体为中心的分析：对数据的深层了解.....245
小结.....256
第9章 Facebook：一体化的奇迹.....257
利用社交网络数据.....258
对Facebook数据的可视化.....274
小结.....294
第10章 语义网：简短的讨论.....296
发展中的变革.....296
人不可能只靠事实生活.....297
期望.....301

前言

与其说网络是一项技术创新，不如说它是一项社会创新。我设计的是一种社会效应（帮助人们一起工作），而不是一种高科技玩具。网络的终极目标是支持并改进现实世界中存在的各种各样的“网”，我们有家庭“网”、协会“网”和公司“网”，我们会亲远疏近。

——Tim Berners-Lee，《Weaving the Web》(Harper)

是否要阅读本书

如果你有编程基础，并且对通过社交网络进行数据挖掘和分析来洞察身边的机会感兴趣的话，你就选对书了。在短短几页介绍之后，我们就会开始动手编写代码。然而，坦率地说，你对本书最主要的抱怨可能就是所有的章节都太短了。遗憾的是，当你试图抓住每天变化并且有大量机会的空间时，情况总是这样。也就是说，我非常赞同“80-20法则”（*http://en.wikipedia.org/wiki/Pareto_principle*），而且我坚信本书在践行“用80%的可用时间来探讨最有趣的20%的内容”方面，是个合理的尝试。

虽然本书篇幅有限，但是它涵盖了很多内容。总的来说，本书堪称广博而非精深，虽然当前的时机非常适合就此主题进行更深入的探讨，但是本书对这些有趣的挖掘和分析技术的深入程度比较有限。本书的编排风格，既适合你通读全书来全面了解社交网络数据方面的入门知识，也适合你根据自己的兴趣来挑选感兴趣的章节来阅读。也就是说，每一章的设计都很短小且相对独立，但是，我在每章内容材料的编排上还是精心安排了前后顺序的，这样，你在阅读全书时才会感到顺畅。

在过去的几年中，Facebook、Twitter和LinkedIn这类社交网站，已经从时尚变为主流，现在已经成为全球现象。在2010年第一季度，广受欢迎的社交网站Facebook已经超过Google，成为人们最经常访问的网站[注1]，这也证实了人们网络消费方式的明确转变。依

注1：　见第9章的第一段。

此事实断言“网络现在已经成为社会文化现象而不再是研究和信息的工具”可能为时尚早；然而，这一数据的确可以表明，社交网站正在以搜索引擎所不具备的方式，大规模地满足了人类的一些基本诉求。社交网络确实正在改变我们的网络生活[注2]，它们能够让技术给我们呈现出最好的（有时是最坏的）一面。社交网络的爆发只是现实世界和网络世界之间的差距不断缩小的一种方式。总的来说，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，来回答以下问题：

- 谁与谁相识，他们共同的朋友是谁？
- 某人与其他人多久联系一次？
- 人与人之间的交流在多大程度上是对称的？
- 网络中最沉默/最健谈的人是谁？
- 网络中最具影响力/人气最高的人是谁？
- 人们正在聊什么（而且它有趣吗）？

要回答这些问题，通常都会涉及两人或更多的人，并且需要找出他们之间存在关系的上下文。回答这些问题所涉及的工作只是开始，更复杂的分析过程还在后面，但是你必须找个地方下手，因为社交网络API和开源工具包都很容易掌握。

笼统来说，本书把社交网站[注3]看成是由人、活动、事件、概念等组成的一幅“图”。Google和Facebook这些行业领袖已经开始逐渐推广以“图”为中心（graph-centric）的理念，而不再强调以“网络”为中心的说法了，因为它们在同时推广以“图”为基础的API。事实上，Tim Berners-Lee建议，也许他应该使用巨大全球图（Giant Global Graph，GGG）（*http://dig.csail.mit.edu/breadcrumbs/node/215*）来代替万维网（World Wide Web，WWW），因为“网”和“图”可以在定义互联网的拓扑结构的上下文中任意互换。虽然Tim Berners-Lee最初的设想能否实现仍有待观察，但是我们所熟知的网络正在因社交数据而变得越来越丰富。我们回顾过去的几年时，最明显的变化就是：由一个固有的语义网创建的第二级和第三级影响是实现真正的语义网的必要条件。两者之间的差距也变得越来越小了。

注2：Facebook的创始人马克·扎克伯格，被《时代》杂志评为2010年的年度人物（*http://www.time.com/time/specials/packages/article/0,28804,2036683_2037183_2037185,00.html*）。

注3：参阅*http://journal.planetwork.net/article.php?lab=reed0704*，换一个角度看待关注数字标识的社交网络。

还是不要阅读本书

从零开始构建自己的自然语言处理器、探究可视化图库的典型用法，以及任何与构建相关的最新技术这类活动，都不属于本书的范围。如果你想了解以上任何内容而购买本书的话，你一定会很失望。然而，并不能仅仅因为“在区区几百页中进行文本分析或记录匹配既不真实，也无法体现我们的最佳技术”，就认为本书无法让你找到疑难问题的合理解决方案，无法将这些方案应用于社交网站，在此过程中并无乐趣可言。当然，这也并不妨碍你在这些诱人的研究领域中培养积极的兴趣。本书的篇幅有限，只能给你开开胃，最多也就是能培养起你在数据处理方面的兴趣，并不能赋予你开天辟地的神奇力量。

也许现如今这样的提醒显得有点多此一举，但是我还是要强调一点（非常重要的一点）：本书所述内容，通常都假定你已经连接到了互联网。本书并不适合在你度假远行时随身携带，因为其中包含很多有超链接的参考内容，而且所有的代码示例都通过超链接直接与GitHub相连，GitHub（*http://github.com*）是一个非常社交化的Git（*http://git-scm.com*）存储库，它一直保持可用的最新示例代码。这种做法希望社交编码能加强志同道合的读者之间的协作，例如，有的人想一起扩展示例，也有人想一起探索有趣的问题。但愿大家能够对资源进行分叉、扩展以及改进——能结识一些志趣相投的新朋友的话更好。像API文档这种现成的资源也可以通过超链接来方便地获取，并且我们认为你更习惯使用在线内容，而不是那些注定会过期的印刷文字。

注意： 与本书配套的bug修复源代码的官方GitHub存储库是*http://github.com/ptwobrussell/Mining-the-Social-Web*。本书的官方Twitter账号是*@SocialWebMining*。

如果你需要一本能让你在sharded MySQL聚类这样的分布式计算平台或者是诸如Hadoop或Redis这样的NoSQL（*http://en.wikipedia.org/wiki/NoSQL*）技术上快速进步的参考书的话，我们并不推荐本书。我们确实使用了一些非常规的存储技术，如CouchDB（*http://couchdb.apache.org*）和Redis（*http://code.google.com/p/redis*），但它们都是在一台机器上运行的，因为这样就能很好地解决眼前的问题了。然而，如果你兴趣强烈且需要水平扩展性的话，它并不能真正连续不断地将示例移植到分布式技术中。我强烈建议你首先要掌握好基础知识，并且要保证代码首先能在一个相对简单的环境中运行，然后再将其移植到一个更复杂的分布式系统中，这样，一旦数据访问不再是在本地时，你就可以依此对算法进行较大的调整以保持其高性能。如果你想要这么做的话，Dumbo（*http://github.com/klbostee/dumbo/wiki*）是用于研究的很好选择。请继续关注本书的Twitter账号（*@SocialWebMining*）来获取有关Dumbo的扩展示例。

虽然我们会尽量遵守运营管理相关网站的条款并领会其精神，但是你可能会对从社交网

站所获得的数据进行加工，本书对于这些做法的法律后果并没有提供任何意见。有些无奈的是，许多最流行的社交网站的授权许可条款禁止在它们提供的平台之外使用这些数据，但目前来说，这种做法是意料之中的事。大部分社交网站就像是带围墙的花园，但是从他们的立场（以及他们的投资者的立场）来说，这些公司提供的很多价值目前依赖于控制平台和保护他们用户的隐私；这种平衡很难维持，而且短期内不可能有较大的改观。

最后的小提示是：本书略微倾向于*nix环境[注4]，因为有些可视化的内容可能会给Windows用户带来些小麻烦。然而，一旦出现这类问题，建议采用合理的替代方案或临时性方案，例如启动VirtualBox（*http://www.virtualbox.org*）在Linux环境中运行示例。幸运的是，并不经常发生这种情况，当出现这些问题时，你可以忽略相应的章节继续阅读，这并不会影响你阅读的乐趣。

工具和先决条件

本书唯一的先决条件就是你需要主动地学习一些Python知识，并且做好了亲自动手处理社交数据的准备。本书的任何技术或者示例都不需要太多数据分析、高性能运算、分布式系统、机器学习或者任何其他特别的背景知识。一些示例可能会涉及你以前没有使用过的结构，如线程池（*http://en.wikipedia.org/wiki/Thread_pool_pattern*），但不必担心——我们使用Python编程。Python直观的语法、优秀的数据处理生态系统软件包和核心数据结构（实际上是JSON）（*http://www.json.org*），使它成为一种优秀的教学工具，虽然功能强大却很容易使用。在其他情况下，如处理自然语言时，我们会使用一些处理高级事务的包，但是我们将会从应用程序设计者的角度来使用这些技术。由于在其他编程语言中也很可能存在非常相似的绑定，因此如果你愿意的话，这应该是移植代码示例的必备练习套路。（但愿这可以在GitHub中用得上！）除了上面的介绍之外，本书没有过多纠结于使用Python的利弊，因为它是非常适合该项工作的工具。如果你是编程新手或者从来没有见过Python语法，那么你只要保证没有跳过前几页内容即可。如果你正在寻找可靠的介绍资料，网上有很多不错的文档，Python的官方教程（*http://docs.python.org/tutorial*）就是很好的起点。

本书试图从各种可视化工具和工具包中，有选择性地介绍一系列有用的可视化工具，既有电子表格类的消费类产品，也有Graphviz（*http://www.graphviz.org*）类的工业产品，还有Protovis（*http://vis.stanford.edu/protovis*）这种尖端的HTML5（*http://en.wikipedia.org/wiki/HTML5*）技术。每一章都会适当地介绍一些新的可视化技术，但我们会尽量顺其自然，让它讲得通。你会对从这些工具中构建轻量级原型的想法感到满意的。也就是

注4：　*nix是Linux/Unix环境的术语，在此处等同于非Windows。

说，本书的大部分可视化内容只是现成的示例或者只是对API项目做了很小的改动，所以只要肯学，你就能做到。

本书约定

下面是本书关于印刷字体方面的一些约定：

斜体（Italic）
: 用于电子信箱、URL、文件名、路径名以及用于强调第一次介绍的新的术语。

等宽字体（`Constant width`）
: 用于文件内容以及命令输出，来表示模块、方法、语句以及命令。

等宽粗体（**`Constant width bold`**）
: 用于程序代码段，来显示应该由用户输入的命令或文字，有时则用于强调程序代码的一部分。

等宽斜体（*`Constant width italic`*）
: 用于程序代码段中可替换的部分以及一些注释。

<等宽字体>（`<Constant width>`）
: 表示应该以真实程序代码取代的语法单元。

注意： 表示和附近文字相关的技巧、建议或一般性注释。

警告： 表示和附近文字相关的警告和注意事项。

代码示例的使用

各章中大多数编号的示例都可以从GitHub（*https://github.com/ptwobrussell/Mining-the-social-web*）上下载——它是本书的官方代码库。我们鼓励你持续关注这个库，以便获得最新的修正了bug的代码，以及由作者和其他社交编码社区编写的更多示例。

本书旨在帮助你完成你的工作。一般来说，可以在程序和文档中使用本书的代码。如果你复制了代码的关键部分，那么你就需要联系我们获得许可。例如，利用本书的几段代码编写程序是不需要许可的。售卖或出版O'Reilly书中示例的CD-ROM确实需要我们的许可。引用本书回答问题以及引用示例代码不需要我们的许可。将本书的大量示例代码用于你的产品文档中需要许可。

如果你在参考文献中提到我们，我们会非常感激，但我们也并不强求。参考文献通常包括标题、作者、出版社和ISBN。例如：“《Mining the Social Web》by Matthew A Russell. © 2011 Matthew Russell，978-1-449-38834-8。”

如果你认为你对代码示例的使用已经超出以上的许可范围，我们很欢迎你通过*permissions@oreilly.com*联系我们。

联系我们

有关本书的任何建议和疑问，可以与下面的出版社联系：

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

我们会在本书的网页中列出勘误表、示例和其他信息。可以通过*http://oreilly.com/catalog/9781449388438/*访问该页面。

读者可以通过Get Satisfaction（*http://getsatisfaction.com/oreilly*）向作者和出版社申请常见的帮助。

读者也可能通过GitHub上的问题追踪系统（*http://github.com/ptwobrussell/Mining-the-Social-Web/issues*），对示例代码以及本书的任何其他内容归档分类。

要评论或询问本书的技术问题，请发送邮件到：

bookquestions@oreilly.com

有关我们的书籍、会议、资源中心以及O'Reilly网络，可以访问我们的网站：

http://www.oreilly.com
http://www.oreilly.com.cn