

## <<社交网站的数据挖掘与分析>>

### 图书基本信息

书名：<<社交网站的数据挖掘与分析>>

13位ISBN编号：9787111369608

10位ISBN编号：7111369602

出版时间：2012-2

出版时间：机械工业出版社

作者：Matthew A. Russell

页数：316

译者：师蓉

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<社交网站的数据挖掘与分析>>

### 内容概要

Facebook、Twitter和LinkedIn产生了大量宝贵的社交数据，但是怎样才能找出谁通过社交媒介正在进行联系？他们在讨论些什么？或者他们在哪儿？本书简洁而且具有可操作性，它将揭示如何回答这些问题甚至更多的问题。

你将学到如何组合社交网络数据、分析技术，如何通过可视化帮助你找到你一直在社交世界中寻找的内容，以及你闻所未闻的有用信息。

本书每章都介绍了在社交网络的不同领域挖掘数据的技术，这些领域包括博客和电子邮件。你所需要具备的就是一定的编程经验和学习基本的Python工具的意愿。

通过本书，你将

- 获得对社交网络世界的直观认识

使用GitHub上灵活的脚本来获取从诸如Twitter、Facebook和LinkedIn等社交网络API中的数据

- 学习如何应用便捷的Python工具来交叉分析你所收集的数据

- 通过XFN探讨基于微格式的社交联系

- 应用诸如TF-IDF、余弦相似性、搭配分析、文档摘要、派系检测之类的先进挖掘技术

- 通过基于HTML 5和JavaScript工具包的网络技术建立交互式可视化

## <<社交网站的数据挖掘与分析>>

### 作者简介

马修·罗塞尔（Matthew A.Russell），Digital Reasoning Systems公司的技术副总裁和Zaffra公司的负责人，是热爱数据挖掘、开源和Web应用技术的计算机科学家。

他也是《Dojo: The Definitive

Guide》（O'Reilly出版社）的作者。

在LinkedIn上联系他或在Twitter上关注@ptwobrussell，可随时关注他的最新动态。

## <<社交网站的数据挖掘与分析>>

### 书籍目录

#### 前言

#### 第1章 绪论：Twitter 数据的处理

- Python 开发工具的安装

- Twitter 数据的收集和处理

- 小结

#### 第2章 微格式：语义标记和常识碰撞

- XFN 和朋友

- 使用XFN 来探讨社交关系

- 地理坐标：兴趣爱好的共同主线

- (以健康的名义)对菜谱进行交叉分析

- 对餐厅评论的搜集

- 小结

#### 第3章 邮箱：虽然老套却很好用

- mbox：Unix 的入门级邮箱

- mbox+CouchDB= 随意的Email 分析

- 将对话线程化到一起

- 使用SIMILE Timeline 将邮件“事件”可视化

- 分析你自己的邮件数据

- 小结

#### 第4章 Twitter：朋友、关注者和Setwise 操作

- REST 风格的和OAuth-Cladded API

- 干练而中肯的数据采集器

- 友谊图的构建

- 小结

#### 第5章 Twitter：tweet，所有的tweet，只有tweet

- 笔PK 剑：和tweet PK 机枪(?!?)

- 对tweet 的分析(每次一个实体)

- 并置潜在的社交网站(或#JustinBieber VS #TeaParty)

- 对大量tweet 的可视化

- 小结

#### 第6章 LinkedIn：为了乐趣(和利润?)

- 将职业网络聚类

- 聚类的动机

- 按职位将联系人聚类

- 获取补充个人信息

- 从地理上聚类网络

- 小结

#### 第7章 Google Buzz：TF-IDF、余弦相似性和搭配

- Buzz=Twitter+ 博客(???)

- 使用NLTK 处理数据

- 文本挖掘的基本原则

- 查找相似文档

- 在二元语法中发Buzz

- 利用Gmail

- 在中断之前试着创建一个搜索引擎.....

## <<社交网站的数据挖掘与分析>>

小结

第8章 博客及其他：自然语言处理（等）

NLP：帕累托式介绍

使用NLTK的典型NLP管线

使用NLTK检测博客中的句子

对文件的总结

以实体为中心的分析：对数据的深层了解

小结

第9章 Facebook：一体化的奇迹

利用社交网络数据

对Facebook数据的可视化

小结

第10章 语义网：简短的讨论

发展中的变革

人不可能只靠事实生活

期望

## <<社交网站的数据挖掘与分析>>

### 章节摘录

版权页：插图：这幅图虽然很简单，却非常有趣。

它连接了8个人，其中，DionAlmaer是共同的主线。

然而，请注意，抓取一层或多层可能会引入图中“与其他所有人都连接”的节点。

单看图的话，我们无法根据“同事”和“朋友”之间的关系，判别Dion与BenGalbraith的关系是否更为密切，但是如果他在“被他的超链接标识的目标”中提供了任何信息的话，我们就可以抓取Ben的XFN信息，搜索其他同事标签来构建“谁与谁共事”的社交网络。

更多挖掘数据的知识请查看第6章，因为它与同事和工作搭档相关。

对广度优先技术的简单分析一般我们不会停顿这么长时间来分析该方法，但是由于这个示例是我们编写的第一个真正的算法，而且我们会在本书中多次见到它，因此值得更仔细地对它进行分析。

一般来说，当你检查算法时，必须考虑两个标准：效率和有效性。

换一种说法就是：性能和质量。

任何算法的标准性能分析通常都包括分析它在最坏情况下的时间复杂度和空间复杂度，即对于一个大型数据集，执行程序所花的时间和需要的内存。

我们采用的广度优先方法实质上是广度优先搜索，只是我们并没有真正执行搜索，因为结束条件并没有把图扩展到最大深度或直到我们遍历完所有节点。

如果搜索了一些具体的东西，而不只是无限地抓取链接，它就可以被视为真正的广度优先搜索了。

## <<社交网站的数据挖掘与分析>>

### 媒体关注与评论

“ 本书是《Programming Collective Intelligence》一书的深入篇，它介绍通过Python从社交网站中采集数据的一种实践方法。

” ——Jeff Hammerbacher.Cloudera首席科学家 “ 对于探索结构化和非结构化数据的一系列工具、技术和理论，本书给出了丰富、紧凑并实用的介绍。

——Alex Martelli.Google高级主管工程师，《Python in a Nutshell》的作者

## <<社交网站的数据挖掘与分析>>

### 编辑推荐

《社交网站的数据挖掘与分析》为Jolt生产效率大奖获奖图书。

## <<社交网站的数据挖掘与分析>>

### 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>