一本在实践中运用数据挖掘、集体智慧及构建推荐系统的指南

写给程序员的数据挖掘实践指南

A Programmer's Guide to Data Mining
The Ancient Art of the Numerati

[美] Ron Zacharski 著
王斌 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

写给程序员的数据挖掘实践指南

[美] Ron Zacharski 著
王斌 译

人民邮电出版社
北京

图书在版编目（CIP）数据

写给程序员的数据挖掘实践指南 / （美）扎哈尔斯基(Zacharski,R.）著 ; 王斌译. -- 北京 : 人民邮电出版社, 2015.11
ISBN 978-7-115-33635-4

Ⅰ. ①写… Ⅱ. ①扎… ②王… Ⅲ. ①数据处理－指南 Ⅳ. ①TP274-62

中国版本图书馆CIP数据核字(2015)第142432号

版权声明

◆ 著　［美］Ron Zacharski
译　王　斌
责任编辑　陈冀康
责任印制　张佳莹　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　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：20.25
字数：403 千字　2015 年 11 月第 1 版
印数：1-3 000 册　2015 年 11 月北京第 1 次印刷
著作权合同登记号　图字：01-2013-0776 号

定价：59.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

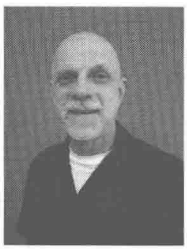
内 容 提 要

数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。大多数数据挖掘的教材都专注于介绍理论基础，因而往往难以理解和学习。

本书是写给程序员的一本数据挖掘指南，可以帮助读者动手实践数据挖掘、应用集体智慧并构建推荐系统。全书共 8 章，介绍了数据挖掘的基本知识和理论、协同过滤、内容过滤及分类、算法评估、朴素贝叶斯、非结构化文本分类以及聚类等内容。本书采用“在实践中学习”的方式，用生动的图示、大量的表格、简明的公式、实用的 Python 代码示例，阐释数据挖掘的知识和技能。每章还给出了习题和练习，帮助读者巩固所学的知识。

本书适合对数据挖掘、数据分析和推荐系统感兴趣的程序员及相关领域的从业者阅读参考；同时，本书也可以作为一本轻松有趣的数据挖掘课程教学参考书。

作者简介



Ron Zacharski 拥有软件开发和计算语言学方面的背景。他是一位计算机科学副教授，并且为从事机器学习和信息提炼的几家创业公司担任过咨询顾问。此前，他在 New Mexico 的计算研究实验室工作，从事机器翻译、特别是人们较少学习的语言方面的研究工作。他曾获得明尼苏达大学计算机科学博士学位，爱丁堡大学的语言学博士后学位，并且拥有音乐艺术学士学位。他还是一位松冈操雄曹洞宗的禅师。

译者简介

王斌　博士，中国科学院信息工程研究所研究员，博士生导师，中国科学院大学兼职教授，研究方向为信息检索、自然语言处理与数据挖掘。主持国家 973、863、国家自然科学基金、国际合作基金、部委及企业合作等课题近 30 项，发表学术论文 130 余篇，领导研制的多个系统上线使用，曾获国家科技进步二等奖和北京市科学技术二等奖各一项。现为中国中文信息学会理事、信息检索、社会媒体处理、语言与知识计算等多个专业委员会委员、《中文信息学报》编委、中国计算机学会高级会员及中文信息处理专业委员会委员。多次担任 SIGIR、ACL、CIKM 等会议的程序委员会委员。《信息检索导论》、《大数据：互联网大规模数据挖掘与分布式处理》、《机器学习实战》、《Mahout 实战》的译者。2006 年起在中国科学院大学讲授《现代信息检索》研究生课程，该课程曾获全校优秀课程称号，累计选课人数已超过 1500 人。迄今培养博士、硕士研究生近 40 名。

译 者 序

这些年来，朋友见面老问我的一句话就是：王斌，你又翻译什么书了？确实，从2008年翻译第一本书《信息检索导论》开始，我就有点一发不可收拾，先后独自或合作翻译了《大数据：互联网大规模数据挖掘与分布式处理》（包括第一版和第二版）、《机器学习实战》、《Mahout实战》、《驯服文本》（待出版）5本书6个版本。“翻译”已经成为我的标签之一。应该说，翻译带来的最大乐趣来自和大家共享好书的喜悦，这种喜悦会传递到我的工作上、生活中，带来满满的正能量。我选择翻译的书的内容都不会超出信息检索、数据挖掘、机器学习、自然语言处理这些范围，这也是我相对比较了解的研究领域。在选择书籍时我并不限定到底是学术著作还是实用手册，只要能对很多读者有较大帮助就行。

本书的宗旨是为程序员提供快速的数据挖掘入门指南。整本书通过真实数据和实例来阐述数据挖掘中的基本技术。书中实例的Python代码和相应数据都可以从网站免费下载获得，读者可以利用这些代码和数据进行实际操作，从而快速掌握数据挖掘的基本概念和技术。书中的实例都特别贴近读者的生活，包括音乐推荐、运动员分类、糖尿病判定等例子都和我们的生活息息相关。

值得一提的是，本书实例中用到的运动员都是真名实姓，好多运动员的大名都如雷贯耳，其中也不乏中国运动员。即使有些运动员我之前并不熟悉，但是网上搜索之后都可以看到一段段运动明星的介绍。对于特别喜欢体育运动的我来说，见到这些名字，看到这一段段介绍，都让我兴奋不已。与这些体育明星相关的实例是我最喜欢翻译的内容之一。和其他很多技术类书籍不同的是，本书引入了很多生动活泼的插图和文字。这些插图中的人物或欢喜、或悲伤、或激动、或愤怒、或思考、或俏皮、或悠闲、或忙碌，这些插图在体现人生百态的同时，也大大缩短了技术和读者之间的距离。本书的另一个特点是十分简洁，作为入门指南，简洁确实是生命线。

本书作者Ron Zacharski的经历颇具传奇色彩：他初学音乐，做了十年的音乐理疗师。后获得计算机科学博士学位，专攻自然语言处理。现在是一名软件开发工程师，同时也是一

名禅宗信奉者。这也是作者一开始就引入日本禅宗大师铃木俊隆（Shunryu Suzuki）的名作《禅者的初心》的原因。对于禅宗我并不了解，查阅一番之后也是懵懵懂懂，只知道禅宗对大名鼎鼎的苹果公司 CEO 乔布斯产生过巨大的影响。或许禅宗的思想体现在整本书的写作当中，等待有心的读者去发现、去领略。

感谢出版社和编辑部的辛勤工作，感谢译者所在的中国科学院信息工程研究所第二研究室的领导、同事以及译者家人对翻译本书的大力支持。

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。

王　斌

2015 年 4 月 29 日　于闵庄路

序

这种简单的练习如果持之以恒，就会获得某种神奇的力量。在获得之前，它很神奇，但是获得之后，却也平淡。

铃木俊隆（Shunryu Suzuki）

《禅者的初心》（Zen Mind, Beginner's Mind[①]）

在阅读本书之前，你可能认为 Pandora[②]、Amazon 推荐系统、恐怖分子自动数据挖掘系统这样的系统十分复杂，其算法背后的数学一定复杂到只有博士才能理解。你可能会认为这些系统的开发人员都像火箭研制专家一样厉害。本书的写作目的之一就是掀开上述复杂性的“面纱”，展示其背后的一些最基本的东西。我们得承认 Google、美国国家安全局以及其他一些地方有很多超级聪明的天才，他们能够开发出令人炫目的复杂算法，但是大多数情况下，数据挖掘只依赖于一些通俗易懂的原理。在阅读本书之前，你可能认为数据挖掘是一个相当惊艳的技术，而读完本书之后，我希望你会说数据挖掘其实平谈无奇。

上面的日文字符表示的是“初心”（Beginner's Mind）这个概念，即一种热切探索各种可能的开放心态。我们当中大部分人都听说过下面这个故事的某个版本（也许来自李小龙的《猛龙过江》）：

---

① 日本禅师铃木俊隆用英文所著的“Zen Mind，Beginner's Mind”（中文译名《禅者的初心》）是畅销英语世界 30 年的禅学著作，据说对乔布斯一生影响很大。其中文版于 2004 年出版。——译者注

② Pandora 电台在美国、澳大利亚和新西兰提供自动音乐推荐系统服务，地址为 http://www.pandora.com。——译者注

某位教授在寻求心灵的启迪，他拜访一位大师以求精神上的指引。在大部分时间内教授滔滔不绝，包括列举他在生活中学到的所有东西以及撰写的所有论文。大师问："喝茶吗？"然后就开始将茶倒入教授的茶杯，一直倒呀倒呀，直到茶溢出茶杯，流到桌子上、地板上……教授大叫："你在干什么？"大师说："倒茶。"然后接着说："你的心就像这个茶杯，它被各种思想所占据，已经没法再听进任何东西了。在我们开始探讨之前，你必须要清空你的心灵。"

对我来说，最优秀的程序员就是空茶杯，他们能够以开放的心态不断地探索新技术（noSQL、noge-js 等）。而普通程序员的心被各种错觉杂念所缠绕，比如 C++很好、Java 很差、PHP 是 Web 开发的唯一方式、MySQL 是唯一考虑的数据库，等等。我希望你从本书中找到一些有价值的思想，并且希望读者在阅读时保持初心。正如铃木俊隆所说的那样：

初学者的思维饱含可能，久习者的思维则饱受羁绊。

前　言

在你面前是一个学习基本的数据挖掘技术的工具。绝大多数数据挖掘教材关注数据挖掘的基础理论知识，因此众所周知给读者带来理解上的困难。当然，不要误解我的意思，那些书中的知识相当重要。但是，如果你是一名想学习一点数据挖掘知识的程序员，你可能会对入门者实用手册感兴趣。而这正是本书的宗旨所在。

本书内容采用“做中学”的思路来组织。我希望读者不是被动地阅读本书，而是通过课后习题和本书提供的 Python 代码进行实践。我也希望读者积极参与到数据挖掘技术的编程当中。本书由一系列互为基础的小的知识点堆积而成，学完本书以后，你就对理解数据挖掘的各种技术打下了基础。

本书各章内容简介

第 1 章　数据挖掘简介及本书使用方法

介绍数据挖掘的概念以及处理的问题，并给出本书学习结束后读者的预期收获。

第 2 章　协同过滤——爱你所爱

介绍社会过滤，给出了多个基本距离的定义，包括曼哈顿距离、欧氏距离以及明式距离等。介绍了皮尔逊相关系数的概念。给出了一个基本过滤算法的 Python 实现。

第 3 章　协同过滤——隐式评级及基于物品的过滤

讨论可用的用户评级类型。用户可以显式给出评级（点赞/点差、5 星或者其他评级方式），也可以隐式给出评级，比如如果用户从亚马逊网站购买了一款 MP3 播放器，那么就可以认为这种购买行为代表了“喜欢”。

第 4 章　内容过滤及分类——基于物品属性的过滤

前面章节中使用了用户对商品的评级信息来进行推荐。本章利用商品本身的属性来进行

推荐。包括 Pandora 在内的一些公司中采用了这种做法。

第 5 章 分类的进一步探讨——算法评估及 kNN

介绍分类器的评估方法，包括 10 折交叉测试、留一法和 Kappa 统计量，此外还介绍了 kNN 算法。

第 6 章 概率及朴素贝叶斯——朴素贝叶斯

探讨朴素贝叶斯分类方法，利用概率密度函数来处理数值型数据。

第 7 章 朴素贝叶斯及文本——非结构化文本分类

介绍如何利用朴素贝叶斯对非结构化文本分类。我们能否对谈论某个电影的推文进行分类，以确定它们的情感倾向性到底是正向还是反向的？

第 8 章 聚类——群组发现

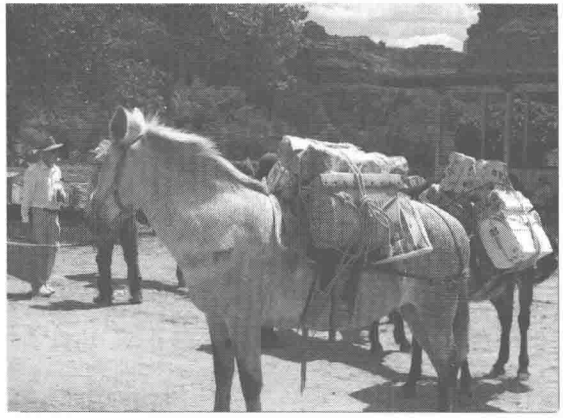
聚类，包括层次聚类和 k-means 聚类。

目　录

第 1 章
Chapter 1

数据挖掘简介及本书使用方法

假想 150 年前一个美国小镇的生活情形：大家都互相认识；百货店某天进了一批布料，店员注意到这批布料中某个特定毛边的样式很可能会引起 Clancey 夫人的高度兴趣，因为他知道 Clancey 夫人喜欢亮花纹样；于是他在心里记着等 Clancey 夫人下次光顾时将该布料拿给她看看；Chow Winkler 告诉酒吧老板 Wilson 先生，他考虑将多余的雷明顿（Renmington）[①]来福枪出售；Wilson 先生将这则消息告诉 Bud Barclay，因为他知道 Bud 正在寻求一把好枪；Valquez 警长及其下属知道 Lee Pye 是需要重点留意的对象，因为 Lee Pye 喜欢喝酒，并且性格暴躁、身体强壮。100 年前的小镇生活都与人和人之间的联系有关。



人们知道你的喜好、健康和婚姻状况。不管是好是坏，大家得到的都是个性化的体验。那时，这种高度个性化的社区生活占据了当时世界上的大部分角落。

时间走过 100 年之后来到了 20 世纪 60 年代。个性化交互的可能性虽然有所下降但仍然存在。本地书店的店员可能会告诉某个常客“书店里上架了 James Michener[②]的新书”，这是

① 雷明顿（Renmington），著名的枪械厂商。——译者注
② James Michener（詹姆斯·麦切纳，1907-1997），美国著名的历史小说家。——译者注

因为他知道该顾客喜欢 James Michener 的作品。或者，店员可能向顾客推荐 Barry Goldwater[①] 写的 *The Conscience of a Conservative*，这是因为他知道该顾客是个坚定的保守派。某个常客去餐馆就餐，服务员可能会问“是不是像以往一样点餐？”

即使到今天，个性化仍然大量存在。我去 Mesilla 的一个本地咖啡店，咖啡店员会问我：“来一大杯加强的浓缩拿铁咖啡？”这是因为他知道这是我每天必点的品种。我将贵妇犬交给宠物美容师，她也不需要问我要修剪的样式。她知道我喜欢无修饰运动型及德式耳型。

但是从 100 年前的小镇开始，情况就有所改变。大型百货店和商场代替了街坊的百货店和其他商店。这种改变刚开始时，人们的选择还十分有限。Henry Ford 曾经说过“只要这车是黑的，顾客就可以把车漆成任何他想要的颜色”[②]。唱片店出售的唱片数目是有限的，而书店出售的书也有限。想要冰激凌？只有香草味、巧克力味或者是草莓味几种。想要洗衣机？1950 年时本地 Sears 商店[③]只有两种型号：一种是售价 55 美元的标准型，另一种是售价 95 美元的豪华型。

欢迎来到 21 世纪

进入 21 世纪，有限的选择已经成为历史。如果想购买音乐，iTunes 提供了 1100 万首歌曲供你选择。这可是 1100 万！截止到 2011 年 10 月，iTunes 已经出售了 160 亿首歌曲。如果需要更多的选择，那么可以访问 Spotify[④]，它上面有超过 1500 万首的歌曲可供选择。

想买书？亚马逊上有超过 200 万的书名可供选择。

---

① Barry Goldwater（巴里·戈德华特，1909-1998），美国政治家，共和党人，曾任亚利桑那州参议员，是 1964 年美国总统选举共和党的总统候选人。*The Conscience of a Conservative* 是其 1960 年出版的一本书。——译者注

② Henry Ford（亨利·福特，1863-1947），美国福特汽车公司的建立者。有人指出，他讲这句话是为当时制造的车只能是黑色而找借口。——译者注

③ Sears，著名零售公司。——译者注

④ Spotify，一个起源于瑞典的音乐平台，提供包括四大唱片公司和众多独立厂牌在内的约 1500 万首歌曲的流媒体服务。——译者注

想看视频？可以有如下多种选择。

NETFLIX
超过10万部视频

hulu
近5万部视频

AmazonPrime™
超过10万部视频

想买一台笔记本电脑？当在亚马逊网站的搜索框中输入 laptop 时，会返回 3811 条结果。

而如果输入 rice cooker（电饭锅），则可以得到超过 1000 条结果。

在不久的将来，我们的选择还会更多：数十亿首在线音乐、大量视频节目以及可以通过 3D 打印定制的产品，等等。

寻找相关对象

面对这么多选择，问题在于寻找相关对象。在 iTunes 的所有 1100 万首歌曲中，我非常喜欢的可能有不少，但是问题在于如何找到这部分歌曲。今晚我想从 Netflix 上观看一部流媒体视频，那么到底应该看哪一部？我想使用 P2P 下载一部视频，但是到底应该下载哪一部？并且，上述问题正变得更加糟糕：每分钟都有数 T 字节的媒体加入到网络中，每分钟 Usenet 上就有 100 个新文件，每分钟都有 24 小时时长的视频上传到 YouTube，每小时都有 180 种新书出版发行。实际上，每天真实世界中都有越来越多的物品可供购买。在所有可选对象组成的“海洋”中，寻找相关对象变得越来越困难。

如果你是媒体制作人，比如马来西亚的季小薇（Zee Avi），那么风险并不在于有人非法下载你的音乐，而在于你自己默默无闻。