Broadview
www.broadview.com.cn

“这可能是最通俗易懂的一本数据挖掘书籍”

互动通邓广梼 PPTV陶闯 联合力荐

New Internet
大数据挖掘

谭磊 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

献给Austin,Benjamin,Cody和他们的爷爷奶奶

安徽大学图书馆
藏书
www.lib.ahu.edu.cn

内 容 简 介

本书全面地介绍了如何使用数据挖掘技术从各种结构的（数据库）或非结构（Web）的海量数据中提取和产生业务知识。作者梳理了各种数据挖掘常用算法和信息采集技术，系统地描述了实际应用时如何在互联网日志分析、电子邮件营销、互联网广告和电子商务上进行数据挖掘，着重介绍了数据挖掘的原理和算法在互联网海量数据挖掘中的应用。

本书主要特点：全面介绍了数据挖掘和大数据的基本概念和技术；大量采用了实际案例，实用性强；详细介绍了大数据挖掘领域最新的商业应用。

本书是从事数据挖掘研究和开发，或者是互联网相关行业从事数据运营的专业人员理想的参考书，同时也可作为了解数据挖掘应用的入门指南。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目（CIP）数据

New Internet：大数据挖掘 / 谭磊著. —北京：电子工业出版社，2013.3
ISBN 978-7-121-19670-6

Ⅰ. ①N… Ⅱ. ①谭… Ⅲ. ①数据采集－基本知识Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2013)第 036703 号

责任编辑：徐津平
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　1/16　印张：23.5　字数：370 千字
印　　次：2013 年 3 月第 1 次印刷
印　　数：4000 册　　定价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

书评

本书是一本可读性极佳的教材。它从互联网广告的角度全面系统地介绍了数据挖掘的基本概念、方法和技术以及数据挖掘对互联网广告的实际意义，重点关注其可行性、有用性、有效性和可伸缩性问题。本书不仅适合作为数据挖掘和知识发现课程的教材，也非常适合作为电子商务、数据挖掘相关领域从业人员的参考资料。

——复旦大学计算机学院教授，博导　@黄萱菁

随着大数据时代的到来，数据科学家这一专业职位变得炙手可热。在2012年10月，《哈佛商业评论》甚至宣布“数据科学家是21世纪最性感的职业”。在本书中，作者基于大量实际项目开发和培训经验，借助最新的互联网应用案例，深入浅出地介绍了数据挖掘领域的基本技术和常用工具。本书是数据科学家完美的入门读物。

——微软亚洲研究院主管研究员，博导　@谢幸 Xing

大家都知道自己现在身处在一个信息化的时代，我们每天从传统的媒体（报纸、杂志、电视，等等）以及新媒体（互联网、网络论坛、微博，等等）获取到大量信息。在每天面对扑面而来的海量信息的同时，常常又有很多人在感叹对自己有用的或者能够让自己感兴趣的东西似乎越来越少。本书也许会为你解开这种困惑。此书深入浅出的描述了时下炙手可热的IT业界的几个词汇。

作为一般的读者可以把此书作为茶余饭后的读物，当你在同事朋友面前侃侃而谈“大数据”、“物联网”、“数据挖掘”等词汇时，相信定能吸引周围人的目光。当你明白数据是如何变成信息，信息是如何变成有用的信息时，或许你的生活也会变得更加多姿

多彩。此书也能帮助企业的经营人员更加深刻的理解如何运用IT（信息技术）提升企业的经营，让IT更好的帮助企业决策千里。当然此书更能帮助我们这些IT从业人员深入的考虑如何运用大数据挖掘技术开发出更好的产品或者解决方案，服务于各个企业，服务于我们的社会。

——富士通（中国）公司 战略规划部总经理 黄邦瑜

随着云时代的来临，大数据也吸引了越来越多的关注。之前我对大数据的了解还停留在概念上，读谭磊的新书让我有了豁然开朗的感觉，明确了自己企业在大数据方向上的目标，也了解了相关的理论和方法。我相信很多关心大数据的朋友都会从书中受益良多。

——凤凰网 CTO @吴华鹏

本书很认真实际的探讨了一个说起来很容易，但是实现起来却需要一个公司从上到下无缝配合才有可能完成的任务。能成功发挥大数据挖掘能力的公司/机构/政府，得到的优势就等于在别人还在用指南针定位目标的时候，你已经装备了卫星导航系统+雷达，做的决定变得更加快、狠、准。

这会是一个大家都努力尝试做大数据挖掘的时代，关键在于，谁能够更疯狂的热爱数据，更理性的尊重数据。

——小米科技联合创始人，副总裁黄江吉 @小米 KKWong

大数据时代的到来让世界变得越来越透明，自由民主是信息社会的生态，无论是生活领域还是行政领域，大众对透明的可视化数据呈现都有迫切的需求，在企业决策、营销决策、医疗、教育等各个领域都需要大数据。大数据流行伊始，技术行业和学术界都非常需要优质的学习书籍，本书作者把自己的互联网数据工作经验与大数据行业发展结合，深入浅出，对行业发展有重大意义，是国内少见的互联网前沿研究的精品之作。

——Web 2.0 研究者，西瓜世界创始人 @柳华芳

此为试读，需要完整PDF请访问：www.ertongbook.com

有人甚至说，“数据是新的石油”，大数据将彻底改变人类文明的发展脉络，重塑我们对于世界、对于生活的认知。谭磊这本书很及时，很深刻的阐述大数据挖掘的各种方法，对于从事数据挖掘的同行来说，是一本不可多得的好书。

——盛大游戏技术保障中心高级总监　@陈桂新

认识 Raymond 很多年，知道他技术很强，这次倒是第一次知道他的文笔也是如此好。大数据的重要性早已不言而喻，我们对此的关注度也是非常高。Raymond 的这本书深浅适中，既符合技术人员的需求，对于非技术的电商从业人员帮助也是很大的。

——阿里巴巴集团资深总监　陈宜

本书是目前国内大数据挖掘类书籍中不可多得的，有理论有实战，非常值得大数据时代的相关研究者阅读。

——腾讯开发高级总监　宋永柱

本书以一位有丰富实践经验的数据工程师的独特视角，以详实的数据和深入浅出的论述揭示了大数据概念下的实际问题，专注于大数据的实用价值和方法，使之不再是虚幻时髦的炒作概念。不同于很多注重解释算法的数据挖掘方面的书籍，本书从“为什么”入手，以通俗易懂的案例展示了大数据领域的全貌，并很好地同时把握了在大数据领域的基本概念和前沿技术。这本书不仅为初学者揭开了大数据这一日趋重要领域的神秘面纱，也为专业人士提供了进一步深入研究的入口。

——微软研究院首席研究员　周礼栋博士

谭磊在这本书中展示了数据挖掘的基本理念和应用场景，能让你在几个小时内读懂数据挖掘，是进入大数据时代的一个敲门砖。

——前腾讯产品总监，现火花无线 CEO 吴国鸿
@火花无线吴国鸿

一场长跑竞赛，并不是一开始冲在最前的人就可以获得最后的冠军，而是取决于战术和耐力。对于互联网产品而言也是如此。随着海量数据的堆砌，其在商业上的价值已经成为企业对未来发展的巨大依托。未来的互联网不再是速度的对决，而是深度的较量！如何正确且深度挖掘数据背后蕴藏的宝藏，这本书将会给出大家希望得到的答案。

——车邻会、卡内网络科技创始人兼 CEO　@吕笋

几年来大数据的运用，给商业世界带来巨大影响。《纽约时报》报道过一个案例，美国超市 Target 通过分析购买数据居然比她父亲还要预先猜测出女孩怀孕的消息！而 Target 正是运用数据挖掘技术，有效提高了细分顾客群体的推广营销效果。本书涵盖该领域相关的技术理论基础概论，并且也提供以互联网为主的各种商业大数据运用前沿的实例，具有很强的实际操作指导意义。对大数据趋势感兴趣的读者，不管是技术人员，或者是管理人员，都能从这本书里获益。

——前 24 券团购网 CTO，互联网创业者　@Bruce 黄海旻

数据就是一座巨大而未知的矿藏，是所有公司最值钱的财富之一，也是当下所有公司都想挖掘的秘密；数据是会说话的，关键是我们如何读懂和理解他，本书能引导我们大家如何读懂他，如何用他指导我们的产品运营和产品设计，如何做精准营销，是非常值得推荐的一本数据分析类书籍。

——著名互联网数据库架构师　金官丁　@mysqlops

本书循序渐进地剖析了大数据挖掘算法在搜索和广告等方面的应用，理论描述深入浅出，应用案例非常精彩，互联网专业知识丰富。本书适合作为搜索广告等相关领域研发的参考手册，也适合作为数据挖掘及 Web 应用的学习教材。

——阿里巴巴资深技术专家　林锋博士　@Frank-林峰

资讯时代里，数据对人类生活的影响和社会的掌控力在不断被放大，理解和运用庞大规模的数据成为了一项雄心勃勃的计划。本书探讨了大数据时代前沿的热点问题，描绘了大规模数据挖掘在当前环境下的典型应用。有概念分析，也有操作实例，既是一本优秀的入门读物，又适合业内人士随时翻阅参考。

——优酷资深工程师　章岑

New Internet
大数据挖掘

谭磊 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

序一

读毕谭磊（Raymond）贤弟的《New Internet：大数据挖掘》原稿后，意犹未尽，又继续读了一遍，皆因内容实在太充实，笨拙的吾一次阅览未能完全消化。

自从懵懵懂懂进入广告传播这个行业后，便与数据这位“性感”魔鬼形影不离，每次执行项目如果没有数据便如同得了爱情单思病，茶饭不思、坐立不安、辗转难眠。

本书内容安排得井井有条，艰深的理论下笔深入浅出，令吾不知不觉坠入黄金屋，整个周末“狠狠”地消化完 Raymond 的杰作。

数据不单只是性感，数据更是神圣的，神圣的数据能够提供充分的信息给各行各业，使这些企业能有所依据地及时优化其产品、服务、渠道、传播、研发等。

数据不是深不可测的，可以这样来简单理解——如同我们日常使用信用卡的数据，当我们将一个时段的数据归纳后，便可以了解自己的消费规律。将各式不同规律的消费者数据归纳后，企业便能洞察自己的产品、服务，以及用户的年龄、性别、国籍、地理位置等的规律。如何发现和运用这些性感数据的规律，便是各门各派的夺宝妙方。

这本书做了大量的资料研究，参考过丰富的素材，选纳众多案例并加以仔细分析，令吾读来得心应手，实乃学习或研究大数据的优秀参考资料，感谢 Raymond 的贡献！

邓广梅
互动通控股集团总裁
北京大学客座教授

序二

首悉数据之说，还是1997年在星传时。领导说，要注意收集数据，包括消费者接触的目的、习惯、联想等。现在想来，显示这些数据的采集来源更值得推敲，有些可能不符合数据来源的真实性。

1999年在电通，为了数据，启用市调公司，做调查，看报告。之后想来，当时设计的大多问题已经提供了供选择的答案，而答案的指向又是我们的主观认识，所以获取的数据可能不符合客观事实性要求。

之后在奥美，强调活动时的数据收集。于是用Word制作了大量的数据收集卡，现场填或发礼品换，在多个地方用了多种方法。现在想来，可能不符合数据的全面性。

再之后在宝洁，基础数据自然很多，要用数个只有几兆容量的U盘储存。但有时多了也很苦恼。因为，有需要索引时，怎么分析呢？有时免不了一个个地查，搜索关键字。现在想来，自己真的没学到一个好的数据检索方法。

2005年去了一家网游公司。作为当时国内最大的几个游戏公司之一，数据已经多到要用几个移动硬盘储存了。网游公司又历来强调数据的挖录，比如登录、消费频次、道具购买力、喜好度，等等。但总觉得挖掘得不够深。现在想来是因为数据在收集开始时，就已经是被填写后的才被收集，跟踪也是滞后的，所以缺乏主动性。

以后，因为投资了家互联网广告公司，所以知道数据该如何收集，如何分析，如何跟踪……但似乎还缺乏些什么。问自己，到底是什么，窃以为是缺乏对数据的甄选方法，白白浪费了很多与眼前无关，但实则有用的数据。这个算是缺乏数据收集的全面性吧。

此为试读，需要完整PDF请访问：www.ertongbook.com

此次有幸看了谭磊兄的《New Internet：大数据挖掘》一书，此书非纯理论之书，且立意颇高，并有许多案例，更是见解独到。

想真正了解何为数据，如何对其进行采集、分析、挖掘与应用，请看此书。

火山 Volcano
天使投资人

序三

认识作者 Raymond 已经很多年了。与 Raymond 认识、熟悉，再深入的交流，他给我的印象是思维敏锐，执行力强。自在微软工作开始，与 Raymond 便有很多交流。之后我们先后离开了微软回国创业。

自在微软时，我们就经常讨论国内互联网的发展方向，其实当初我们对于国内互联网企业的核心竞争力的意见并不一致，但有一点我们是达成共识的，就是未来互联网企业的竞争力不仅是“争夺”用户的能力，而且是“挖掘”用户价值的能力。我们都认为，挖掘用户价值的实质就是以大数据挖掘为核心的技术和运用。在这点上，中国互联网公司需要更加注重手里的数据资源，深挖出更大的信息价值，才能进一步提升用户价值或者是单用户的平均产出值（ARPU 值）。

Big Data 作为业界在 2012 年讨论得最多的话题，受到的重视程度很高，也因而有了不少相关的文章和书籍。在此之前，讲述大数据和数据挖掘的书虽然很多，但是大多比较偏理论，给实际应用者的帮助并不大。而 Raymond 的这本《New Internet：大数据挖掘》则从一个全新的角度讲述了在数据挖掘领域的大数据，给予数据挖掘和运营人员很好的实战指导。

大数据挖掘这个课题涉及的学科很多，要写好关于数据挖掘的书既要有丰富的实践经验做基础，还需要有扎实的理论知识。我很高兴地看到，Raymond 在这本新书中把他之前的实践和理论知识有机地结合起来了。

陶闯 Vincent Tao

PPTV CEO，Ph.D.

作者的话

从接到侠少的约稿到现在已经四个月了，但对大数据挖掘的关注是远不止四个月的。很感谢侠少给我这个机会，在写书的过程中我对于大数据挖掘的理解也上升了一个台阶，因为当你试图给第二个人解释你自以为很了解的概念时会发现自己了解的深度还远远不够。第一次写完之后自己再读又发现新的需要修改的内容，如此反复多次，终于大致成稿。现在的版本中一定还有用词不恰当的地方，请各位读者海涵。

数据对于人们到底意味着什么？我在写书的过程中一直在思考这个问题。数据挖掘并不是一门崭新的学科，而是综合了统计分析、机器学习、数据库等多方面研究成果的应用学科。而近年来的大数据又使得数据挖掘有了革命性的发展。

诸行无常，诸法无我。在大数据的环境中唯一不变的是变化，我们在本书中讲述的理论和概念很可能过了两年甚至一年就会发生变化，这也是互联网时代的本质特征。

窃认为，写一本书，即便是教科书，也不能停留在理论层面。如果一本书写成阳春白雪那是非常失败的。自有计算机这个专业以来，做计算机理论研究和做计算机应用之间就有一道鸿沟。比如笔者读书时在 *Machine Learning* 期刊上发表的 *PAC Learning Axis-aligned Rectangles with Respect to Product Distributions from Multiple-Instance Examples* 一文，虽然提出了一个很美丽的 PAC 学习算法，但是这个算法的实现性仅仅停留在理论层面。本书的初衷就是把“大数据挖掘”写成“最炫民族风”，所以书中所举的实例基本都是切实可行的实际案例，限于商业原因，我们不能详细描述全部的具体实施过程，如果读者有疑问，欢迎随时和我交流。

而一本书也一定不能只是信息资料和概念的堆砌。本书在陈

述大数据的事实和概念的同时，也尽量揭示在这些事实和概念背后的原理和实际运用。

这本书不是一个人的战斗。在这本书的写作过程中，我得到了很多人的帮助。首先要感谢的是互动通 HdtMedia 的 Michael 和 Clarence 两位前辈对我的大力支持和鼓励，让我有力量可以写完这本书。我要感谢 Microsoft 总部云平台的首席开发经理陈众同学、Microsoft 亚洲研究院的周礼栋博士和微软搜索技术部首席开发经理刘欣同学给本书的结构提出的修改意见。感谢复旦大学的黄萱菁博导和微软亚洲研究院的谢幸博导，他们除了在百忙之中给本书写了书评之外，还提出了宝贵的修改建议。

还要感谢江峰、韩冬、曹晓波、王海、荷铁勇、楼建强、李嘉骅、吴浩苗等同学帮我查找数据挖掘相关资料，鲍佳、刘晓鹏、俞舒、李悌开、戴霖和匙枘等同学帮我校验一些章节。特别要感谢董雅楠同学多次通读全书，挑出的错别字和语法问题令我汗颜，让我觉得全国普通话考试还是有必要的。

思美传媒的江山同学、淘宝开放平台的冯光同学、UTC 的于振伟同学、车邻网的吕笋同学、火花无线的吴国鸿同学、聚流电商的周为同学和首正信息的罗俊峰同学为本书提供了大量精彩的案例和数据，在此一并表示特别的谢意。

Raymond @CarelessWhisper

2013 年 1 月 28 日

目 录

此为试读，需要完整PDF请访问：www.ertongbook.com