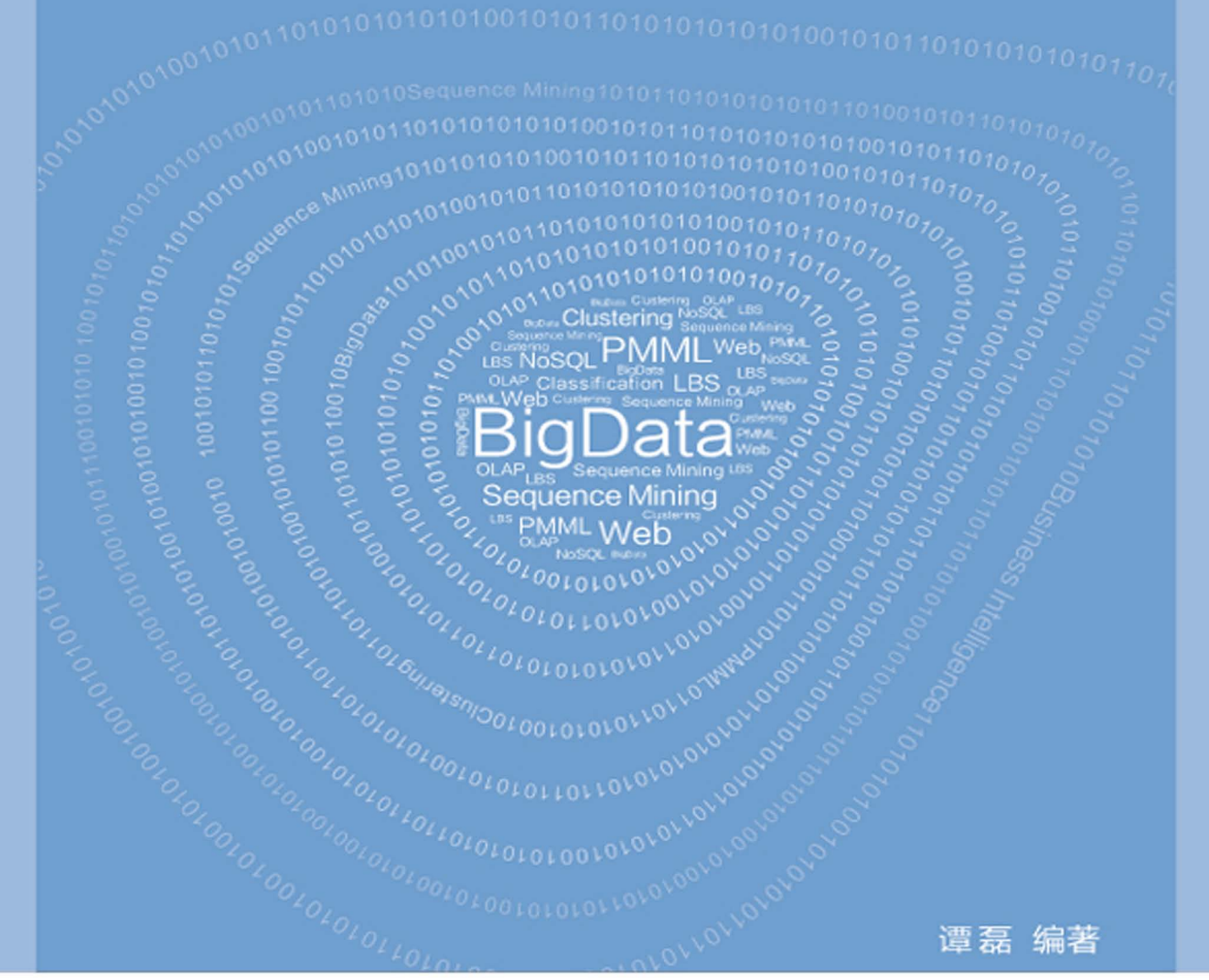
Broadview
www.broadview.com.cn

# New Internat
# 大数据挖掘




谭磊 编著

这是一种革命，我们确实正在进行这场革命。庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来。没有哪个领域不会受到影响，迎来“大数据时代”(Age of Big Data)。

——纽约时报

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

# *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

## 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。



**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

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

# 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！

邓广梼
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 Lea rning* 期刊上发表的 *PAC Lear ning 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

# 第 1 章
# 绪论——从淘金客到矿山主

“大数据”（Big Data）是自 2011 年以来最时髦的几个 IT 词汇之一。随着互联网的高速发展，如今我们拥有的数据量已经可以用“大”来称呼了。

不过现如今，我们同样正在面临着一个尴尬的境地——数据丰富，信息匮乏（Data Rich But Information Poor）。快速增长的海量数据，已经远远地超过了人们的理解能力，如果不借助强有力的工具，很难弄清大堆数据中所蕴含的信息。

很多时候，海量的数据只是数据，并未成为我们可以应用的信息。而我们做数据挖掘和数据分析的主要目的就是为了实现数据的价值，这恰恰也是本作品的目的。但凡读者能从书中获取到一丝对他们在数据应用方面能有所提高的信息，于愿足矣。

2012 年 10 月哈佛商业评论有一篇专题文章《数据科学家：21 世纪最性感的职业》（*Data Scientist: The Sexiest Job of the 21st Century*）。文章中指出“数据科学家”是在企业中新出现的一个职业，而其主要工作就是在大数据上找出有用的信息。虽然过去有些企业也能意识到数据挖掘的重要性，但是最近几年一些新技术的出现使得充分利用大数据的价值成为可能。作者认为，数据科学家能够称为一个性感职业的原因是在于数据本身。因为数据是性感的，所以数据科学家才能被称为性感的职业。

和其他讲述数据挖掘理论和算法的书籍不同，本书的目的主要是在普及数据挖掘基本知识的基础上，让大家通过一些数据挖掘在互联网应用领域的实例来理解什么是基于大数据的数据挖掘。换句话说，这本关于大数据挖掘的书的重点在于“大”和“如

何挖掘”的应用实例。本书尽量采用浅显易懂的语言组织，希望没有太多计算机专业背景的同学也能读懂数据挖掘。

## 1.1 大数据时代的“四 V”

自互联网诞生以来，数据一直以惊人的速度增长。门户网站、搜索引擎、社交网络的先后问世引领着传统互联网数据不断膨胀。而从 2008 年开始，以 iPhone 和 iPad 为代表的智能手机和平板电脑的快速普及又推动了移动互联网数据的迅猛增长。移动互联网能更准确、更快地收集用户信息，比如位置、活动轨迹、生活信息等数据。除此之外，由能够测量物体运动、震动、温度、湿度等特性的各种传感器与计算机相互连接形成的一个更大的网——物联网（The Internet Of Things）也正在逐步成型。可以预计，未来数据的膨胀速度定会继续加剧。

目前企业级的数据仓库和应用都建立在传统的关系型数据库上，然而面对动辄上亿至万亿条数据的查询分析，传统方式显得越来越力不从心。而且传统的数据处理方式对于非常规化的数据处理也没有很好的解决方法。同时，大量化、多样化、快速变化的数据逐渐超出了现有企业的 IT 架构和基础设施的承载能力，从而导致许多企业目前的网络环境、存储、架构越来越不能适应新的数据格局。大量的数据“被沉睡”在那里，价值被埋没，等着我们去发现。

Gartner（高德纳）公司研究认为，新产生的数据量每年正以至少 50%的速度递增，而这个速度使得每年新增的数据量不到两年就会翻一番。Cisco（思科）公司在一份报告中推测 2015 年仅移动数据量将会突破每月 6EB，等于 60 亿 GB 字节；而 IDC 最新的数字宇宙（Digital Universe）预计，到 2020 年世界上的数据存储总量将达到 35ZB，等于 35 万亿 GB 字节。而这个数字还是受到了存储能力的限制。

“大数据”时代已经来临！“Big Data”也成为最近两年的Buzzword（时髦词汇）之一。

IBM 提出的“三 V”概念，即大量化（Volume）、多样化（Variety）和快速化（Velocity），是“大数据”时代的显著特征，这些特征正在给现在的 IT 企业带来巨大挑战。所谓“三 V”，因为这三个英文词 Volume、Variety 和 Velocity 的开始首字母都是“V”。而最近这两年，着眼数据应用的专家们提出了大数据的“四 V”概念。“四 V”概念是在原有的“三 V”基础上增加了第四个首字母为 V 的词——Value（价值），即企业要实现的是大数据的价值。在作者看来，第四个“V”才是关键。如果我们不能够实现数据的价值，那么再海量的数据也是没有价值的。如图 1-1 所示。

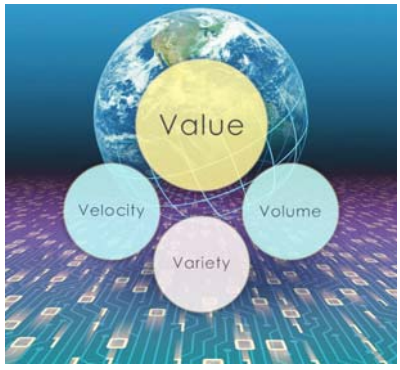



图 1-1　大数据“四 V”示意图

- 在大数据的四“V”中，大量化（Volume）是显而易见的。如果没有大量的数据，我们就无法称其为“大数据”。如今，各家企业的数据量正在从 GB、TB 级向着 PB、EB 级大踏步迈进。1PB=1,024TB=1,048,576GB=1,125,899,906,842,624 Byte（字节），而 1EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Byte。
- 多样化（Variety）是指半结构化、非结构化数据的量和结构化数据一样在飞速增长。全世界 40 亿手机用户已经将

他们自己变成了数据流的提供者，同时手机制造商在他们的产品中嵌入了 3 千万个传感器，而且这一装机量正以每年 30%的速度增长。各个企业采集的数据并不限于传统的数据格式，非结构化数据的增长速率超过了结构化数据的增长速率。所谓半结构化，是指数据有一定结构，但又没有固定的模型描述。结构化和半结构化数据通常能够用普通的 XML 模式来描述，但是非结构化数据就需要特殊处理了。

- 快速化（Velocity）主要是指商业和各种相关领域处理的交易以及数据在以越来越高的速度和频率产生。每一分钟都有大量的数据在商业环境和互联网环境中产生。
- 四“V”中的价值（Value），则是指数据运营和应用的重要性。如果没有数据分析和数据挖掘，数据还只是数据。只有通过处理和分析过的数据才能转化成信息，归纳成知识。本书所强调的就是在“大数据”时代数据的价值。

毋庸置疑，新增的价值（Value）是这四个“V”中最值得我们关注的一个“V”。我们做数据挖掘和数据分析的主要目的也就是为了实现数据的价值。在“大数据”时代，数据将是企业的核心资产，如何充分利用历史的和每天产生的海量数据，如何从海量数据中提取有价值（Value）的信息，如何把信息转化成商业智能的知识和规则，对企业生成竞争力乃至企业成败起到至关重要的作用。如今企业可以轻易收集到客户的多种行为数据，从中提取到的信息一旦善加利用，便可以为企业带来巨大的回报，如沃尔玛通过数据挖掘发现消费者行为习惯中啤酒与尿不湿的神奇关联；美国联邦调查局通过对银行信用卡记录的数据挖掘发现恐怖分子踪迹；银行、电信和保险业通过建立用户信息和交易记录的数据仓库和分析模型来提高利润、降低风险等。

除了这四个“V”之外，业内也有学者和从业者提出不少其他关于大数据的“V”，值得我们关注。在这之前我们还真没有意识到有这么多有趣的英文词是以“V”为首字母的：数据的可

验证性（Verification）、可变性（Variability）、真实性（Veracity）和邻近性（Vicinity）。可验证性（Verification）指的是数据需要经过验证，因为数据量大了之后，带来的一个后果必然是数据质量的良莠不齐以及因不同级别的用户介入而产生的数据安全问题。可变性（Variability）指的主要是数据格式的可变性，着重于非关系型数据。真实性（Veracity）指的是因为数据来自不同的源头，而有些数据的来源（比如 Facebook 上的评论和 Twitter 上的跟帖）的可信度是需要考虑在内的。邻近性（Vicinity）和大数据的存储相关，处理数据的程序和服务器需要能够就近获取资源，否则会造成大量的浪费和效率的降低。

公众对于“大数据”的关注从谷歌趋势分析中可见一斑。图 1-2 是在 Google 趋势分析的网站上做的“Big Data”查询得到的结果，详情可查看网址：http://www.google.com/trends/?q=big+data。

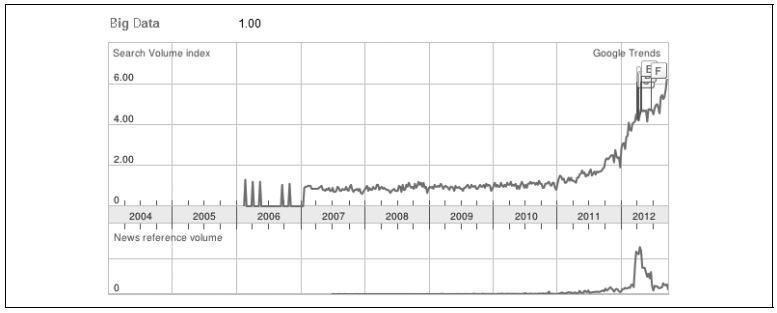



图 1-2　表示“大数据”关注度的谷歌趋势图

从图 1-2 中我们可以看到“大数据”（Big Data）从 2007 年起就开始受到关注，而 2012 年的关注度更是直线上升。

## 1.2 什么是大数据挖掘

什么是数据挖掘呢？古人云：“物以类聚，人以群分”。这句话正是描述了数据挖掘中的一类算法——聚类算法。

要看一个人是怎样的，只需要看他周围都有什么样的朋友，

而从数据挖掘的角度来说，用聚类算法预测一个对象的特征，只需要看它周围对象的特征。

大数据挖掘在本书中的定义是指在大数据上做“数据挖掘”的过程。这里的“数据挖掘”是指对数据进行处理和研究，并从数据中提取有用信息和发现知识的过程。

### 1.2.1 从数据分析到数据挖掘

对于分析和处理数据，我们另外一个常用的词是数据分析，那么数据分析和数据挖掘之间有什么区别呢？

从本质上来说，数据分析和数据挖掘都是为了从收集来的数据中提取有用信息，发现知识，而对数据加以详细研究和概括总结的过程。在不少场景中，数据分析和数据挖掘这两个概念是可以互换的。而它们之间最大的区别是数据本身的不同，这主要表现在以下两个方面：

- 数据量的不同，数据分析通常是存储在数据库或者文件中，一个应用的数据数量级在 MB 或是 GB，而数据挖掘的应用数据动辄 TB，甚至 PB。
- 数据类型的不同，数据挖掘的对象不仅仅是文本，还有音频、视频和图片数据，并且不仅是规范化数据，而且还有半规范化和不规范数据。

从某种意义上来讲，它们之间的区别就像淘金客和矿山主，不同点在于淘金客只在一条小溪上工作，甚至几十个人共享一条小溪，通常只能通过手工作业用沙漏从沙里淘金。而矿山主则占有整座巨大的矿山，由于矿山拥有成分复杂的矿石和数量繁多的伴生矿物，这时矿山主就不能仅仅依靠手工作业，而需要建立一个以机器为劳动力的现代化工业企业才能做到最大程度的效率的产出。

数据挖掘与传统的数据分析（如查询、报表、联机应用分析）的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得出的信息通常具有先前未知性、有效性

和可实用性三个特征。而数据分析主要是一个假设检验的过程，是一个严重依赖于数据分析师手工作业的过程。如果有高水平的淘金客，我们就能淘出金子。

数据分析主要采用的是统计学的技术。在统计学领域，我们可以将数据分析划分为两大类：探索性数据分析（Exploratory Data Analysis，EDA）和验证性数据分析（Confirmatory Data Analysis，CDA）。

- EDA（探索性数据分析）是描述性统计分析，指为了形成值得检验的假设而对数据进行分析的一种方法，是对传统统计学假设检验手段的补充。EDA 方法由美国著名统计学家约翰·图基（John Tukey）命名，侧重于在数据之中发现新的特征。EDA 方法通常比较灵活，讲究让数据自己说话，但是通常需要依靠数据分析师的专业知识来做判断。
- CDA（验证性数据分析）则侧重于对已有假设的证实或证伪，是定性数据分析，又可称为定性资料分析或者定性研究。在做验证性数据分析时，我们往往已经有了一个假设，需要数据分析来帮助确认。CDA 在进行分析之前已经有预设的概率模型，只需把现有的数据套入到模型中。

EDA 和 CDA 在商业环境中的作用都比较普遍，和它们的名称相对应，EDA 主要用于对商业数据的探索，而 CDA 是在某一个模型之上把商业数据加入来做验证。统计学的分析方式主要有以下几种：

- 各种数学运算（Simple Math）。
- 快速傅里叶变换（FFT）。
- 平滑和滤波（Smoothing and Filtering）。
- 基线和峰值分析（Baseline and Peak Analysis）。

由此可见，一般的数据分析主要用来对数值进行处理，通常无法对诸如词语、照片、观察结果之类的非数值型数据进行分析，

但是如果需要，我们可以把这些数据以量化的方式转化或组织起来形成能够分析的数据形式。

不同于数据分析有明确目标的特点，数据挖掘是一个知识发现的过程，是一个人驱使机器（机器学习算法）在矿山中挖掘宝藏的过程。数据挖掘强调对大量观测到的数据做处理，它是涉及数据管理、人工智能、机器学习、模式识别及数据可视化等学科的边缘学科。

从数据组织上来讲，数据分析相对比较简单，数据一般以文件的形式或是单个数据库的方式组织。因为可以处理的数据量有限，如果不能把全部数据集都放入到数据分析计划中，我们必须选择所需要的数据。抽样调查是常用的一种数据选择方法。如果不能从抽样调查中选择正确的数据集，会造成分析不奏效或者产出有偏差的结果。事实上，如果企业习惯于采用数据抽样的方法，虽然好像处理了所有数据，但数据的科学性在本质上是被削弱的。

数据挖掘则可以基本保证数据的科学性。不过，海量数据不是普通数据库能够存储和处理的。所以数据挖掘必须建立在数据仓库或是分布式存储的基础之上。而且数据除了量之外，对于挖掘的实时性的要求也有所提高。在每天都有 100TB 数据产生的场景下，数据挖掘的处理速度每天必须至少是在 100TB 之上。我们在第 3 章的“数据仓库”中将会介绍如何存储和处理大规模数据。

从手段上说，数据分析的主要算法以统计学为基础。分类与预测是两种数据分析形式，它们可以用来抽取能够描述重要数据集合或预测未来数据趋势的模型中的样本。分类算法（Classification）用于预测数据对象的离散类别（Categorical Label）。预测方法（Prediction）则用于预测数据对象的连续取值。

数据挖掘不仅仅需要统计学，还大量使用了机器学习的算法。关于数据挖掘的算法，我们会在后续的第 4 章“数据挖掘算法及原理”中做概略性的介绍。

总结一下，即大数据挖掘是传统手工业式的数据分析的现代大工业形式。数据挖掘建立在拥有大量数据，并且能够让机器方便读取的数据仓库之上，采用机器学习的算法，是自动发掘知识的过程。然而这并不意味着数据分析完全被取代。就像现代大工业只是取代了手工生产的生产组织形式，而手工生产中的方法、技能等都被现代大工业吸收进来，重新赋予了新的意义。同样的，大数据挖掘也需要数据分析的算法和思路，只是用新的方法重新组织施行。如今这一过程才刚刚开始。

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、人工智能、数据库等诸多方面的研究成果的边缘学科，同时与专家系统、知识管理等研究方向不同的是，数据挖掘更侧重于企业应用。

### 1.2.2　Web 挖掘

基于互联网的挖掘（Web 挖掘）是利用数据挖掘技术从互联网上的文档中及互联网服务上自动发现并提取人们感兴趣的信息。在本书的其他部分我们都简称其为 Web 挖掘。Web 挖掘是一项综合技术，除了数据挖掘领域的所有技术之外，还涉及互联网技术、信息学、自然语言处理等多个领域。在本书中提到的大数据挖掘主要是指 Web 挖掘的应用。Web 挖掘的目的就是从大量互联网上看似杂乱无章的数据中，发现其中的规则和知识以供决策支持。尽管 Web 挖掘不同于信息检索，但它们在实现技术上却有不少相似之处，所以 Web 挖掘技术可以借鉴信息搜索技术。

一般来讲 Web 挖掘可分为三类：内容挖掘（Content Mining）、结构挖掘（Structure Mining）和用户访问模式挖掘（Web Usage Mining），而内容挖掘和用户访问模式挖掘是我们在本书中主要讲述的两个主要方面，但是内容挖掘中关于信息检索和组织方式这些与搜索相关的技术不在本书的讨论范围之内。

Web 挖掘中数据收集的信息具有海量性和实时性的特点。

所谓在互联网上“凡走过必留下痕迹”，就是指当访客从进入某网站的那一刻起，他的一切浏览行为与历程都是可以立即被记录的。在 Web 2.0 时代，我们注重的是用户体验，所以 Web 挖掘是以交互式个性化服务为终极目标的，比如我们可以应不同访客的访问习惯呈现出为其专属设计的网页，或者给不同的访客提供不同的商品和服务体验。

以“脸谱”网（Facebook）为标杆的社交网络兴起后，大量的 UGC（用户产生的内容，User Generated Content）内容包括音频、文本信息、视频、图片等非结构化数据出现了。有数据表明今天互联网上的数据总量有 80%是非数字化的。如何从非结构化数据中找出有用的信息，不仅是 Web 挖掘，还是 Web 信息检索研究的主要方向，不过 Web 信息检索研究不是本书讨论的方向。

Web 挖掘是对现代电子商务战略的一个重要技术支持，尤其是 Web 挖掘中的用户访问模式。这主要用于对客户在网上行为的分析以及潜在的顾客信息的发现。我们在第 10 章中将会讲述 Web 挖掘如何通过数据预处理、模式发现和模式分析，从数据中发现有用信息和规则以帮助电子商务企业。Web 挖掘通常的实现方法是对服务器日志（Server Log）、错误信息日志（Error Log）和本地终端数据日志（Cookie Log）等日志文件进行分析，从而挖掘出用户的访问行为、访问频次和浏览内容等信息，从中找出一定的模式和规则。

### 1.2.3 大数据挖掘之“大”

虽然从数据量级上来说，我们已经走进了大数据时代，但如今的数据分析和数据使用量却已经明显跟不上数据发展的脚步。全世界的整体信息量每两年以超过翻番的速度增长，据估计 2011 年在全世界共产生和复制约 1.8ZB 的数据。视频、图片、音频等非结构化媒体数据的应用越来越频繁，社交网络不断增长和壮大，而同时相对传统的结构化数据的个体容量和个体数量也

在迅速飙升。有数据表明，在 2012 年，数据挖掘行业中有约 3% 的业者已经在最大量级为 100PB 或者 100PB 以上规模的数据上进行数据挖掘工作。

美国市场研究公司 IDC 发布的报告表示企业中大数据的出现在部分程度上要归功于以下三个硬件方面的原因：

- 计算机硬件成本的降低。
- 与此同时，随着计算机内部存储设备的成本降低，企业比以往任何时候都更适合在“内存”中同时保有和处理更多的数据。
- 更重要的是，将服务器连接成集群拼接超级计算机（系统）的方式比以往简单得多。

以上三点硬件方面的因素促成了大数据的产生。

在对大数据的众多说法中，其中的一个提法是形容某个企业或群体创造的大量非结构化和半结构化数据。但实际上只是数据量大并不等同于大数据。在此笔者不想咬文嚼字，纠结于学术理论上“Big Data”到底指的是什么，也不想讨论什么样的数据只算是普通数据，什么样的数据可以称海量数据，或者是否海量数据就等同于“Big Data”。在本书中，大数据的特征就是在 1.1 节中提到的大量化、多样化、快速化及价值，即大数据的“四 V”特征。

笔者认为在不久的未来，数据可能成为最大的一类交易商品。在互联网上，继“入口为王”、“流量为王”和“应用为王”之后，下一个概念应当是“数据为王”。未来的大数据将会像今天的公用设施一样，有数据提供方、管理方、数据运营商、第三方服务商和监管方。数据的定量供应和处理将会形成一个新的大产业链。美国数据存储公司 EMC 首席市场官 Jeremy Burton 曾经表示：“大量杂乱无章的信息无休止地增加，带来了无穷无尽的机会，将促使社会、技术、科学和经济发生根本性改变。信息是企业最重要的资产，大数据正在促使企业改变信息管理方式，并从信息中挖掘出更大的价值。”

2012 年，Gartner 公司做的统计数字表明，全球企业 2012 年在大数据上投资总额约 290 亿美金，而这一数字在今后的 4 年，每年会以 17%的速度增长，到 2016 年会上升至 550 亿美金。

## 1.3 大数据挖掘的国内外发展

数据挖掘是一门边缘应用学科，它的蓬勃发展正是由于它在各个领域的广泛应用。而正因为它对于新算法和新技术的渴求，如今数据挖掘在学术上也已经逐渐自立门户，发展成为一项拥有多种研究方向的独立学科。Gartner 公司的报告中也指出，数据挖掘技术将是未来 10 年内最重要的技术之一。

当数据只是停留在数据存储时，它们是数据。而把数据经过加工处理，它们就成了有用的信息。如果信息组合能够合理地产生价值，特别是商业价值，我们可以称其为知识。数据挖掘的过程就是把数据加工处理变成信息，最后转化为知识的过程。如何做好数据加工的过程是数据挖掘研究的方向，而商业价值则是数据挖掘应用发展的方向。我们在本节中将从这两个角度来看数据挖掘最近的发展。

### 1.3.1 数据挖掘的应用发展

在国外，数据挖掘的应用已经非常普遍。一般较常见的应用案例发生在营销领域的零售业、直效行销界、制造业、财务金融保险、通信业、医疗服务业以及各种政府机关等。

#### 1.3.1.1 直效行销

在众多应用案例中，数据挖掘在营销领域的应用应该是最为广泛的。数据挖掘可以从销售的各项数据中发掘消费者的消费习性，即通过交易纪录找出顾客偏好的产品组合，以进行交叉销售（Cross-selling）、向上销售（Up-selling）。找出流失顾客的特征与

推出新产品的时机点等也都是数据挖掘在零售业中常见的应用。数据挖掘在直效行销商家中的应用是做得最好的。直效行销（Direct Marketing）又名零阶通路，是指制造商直接将产品出售给消费者，使通路阶层降至零阶，减少中间费用，为消费者取得较低价格的销售方式。直效行销（Direct Marketing）中强调的人群分流概念与数据库的行销方式在导入数据挖掘的技术后，可以使直效行销的发展性更为强大，例如利用数据挖掘来分析消费者的消费行为与交易纪录，结合基本数据，并依其对品牌价值的高低等级来区隔顾客，进而达到差异化行销的目的。其实我们可以把厂家直接运营电子商务的公司当作一类直效行销公司，因为他们也是去除所有中间环节，直接面对消费者的。

#### 1.3.1.2　金融业

数据挖掘在金融业中也有着充分的应用。例如，股票交易商可以利用数据挖掘来分析市场动向，并预测个别公司的营运状况以及股价走向等；又例如，采用数据挖掘中的关联规则挖掘技术，我们可以成功预测银行中不同客户的需求，一旦获得了这些信息，银行就可以改善对不同客户的服务项目。其实现在银行天天都在开发新的沟通客户的方法，而这些新方法的依据很多就来自于数据挖掘后产生的信息和规则。

我们来看几个简单的案例以便对金融业如何运用数据做直观的了解：

- 很多银行都在自己的 ATM 机上捆绑了顾客可能感兴趣的本行产品的信息，供使用本行 ATM 机的用户了解，而客户不同，等待屏所显示的内容是不同的。
- 如果在商业银行的数据库中显示，某个高信用限额的客户更换了地址，这个客户很有可能新近购买了一栋住宅，因此会有可能需要更高信用限额或者更高端的新信用卡，也可能需要一个住房改善贷款，而这些产品都可以通过信用卡账单邮寄给客户。
- 当客户打电话咨询银行客服时，数据可以有力地帮助电

话销售代表，因为客服代表的计算机屏幕上会显示出客户的特点，同时也显示出顾客会对什么样的金融产品感兴趣。

#### 1.3.1.3 制造业和医疗

与营销和金融业不同的是，制造业对数据挖掘的需求主要是在提高产品质量和降低生产成本上。比如在品质控管方面，在制造过程中找出影响产品品质最重要的因素，以提高作业流程的工作效率。能预测汽车或游艇中哪个部分容易出故障，找到改善发动机的维修周期的方法，降低返修率。或是在供应链上找出可以节省的环节来降低整体产品的生产成本。

作为制造业领域的制药公司则除了以上两个方面之外，还可以借鉴营销商的经验，通过分析销售记录找出高价值的医院或者医生来决定如何获取最高回报。

在医疗业中，数据挖掘技术可以被用来预测手术的成功率、用药的效果、诊断或是流程控制的效率等。使用频率最高的可能是医用 DSS（决策支持系统，Decision Support System），根据病人的不同症状，病人的特殊属性和各种疾病的相关性来做诊断。

#### 1.3.1.4 欺诈行为预测

对于欺诈行为的侦测（Fraud Detection），数据挖掘技术在电话公司、信用卡公司和保险公司都有着广泛的应用，这是因为在这些行业中，每年因为欺诈行为而造成的损失都非常可观，而数据挖掘可以从一些信用不良的客户数据或者历史欺诈交易中找出相似的特征并预测可能发生的欺诈交易，以达到减少损失的目的。

#### 1.3.1.5 政府机关

除了企业，在政府机关，数据挖掘也获得大量的应用实践机会。美国各级交通部门以及高速公路分管机构都已经利用数据挖掘来预测路面的生命周期。IRS（美国税务局）通过税表中的数

据来找出可能的报税欺诈嫌疑人。FBI（美国联邦调查局）和 CIA（美国中央情报局）通过搜寻异常的信用卡、银行转账记录和电话记录来查找恐怖分子的痕迹。

表 1-1 是 KDNuggets 做的 2011 年全球数据挖掘应用行业调查。前 10 大数据挖掘应用领域是用户关系管理（CRM，25%）、银行业（18.9%）、健康行业（16.7%）、教育行业（16.2%）、欺诈预防（14%）、科学研究（13.6%）、社会化网络（13.2%）、信用评级（12.7%）、直效行销（12.3%）、保险业（12.3%）。其中值得一提的是社会化网络（Social Networks）这一领域，一年之中对于 SNS 的关注度翻了一番，增加了 6.6%。

表 1-1　2011 年数据挖掘应用领域调查表

| Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters]　2011 % of voters　2010 % of voters | |
|---|---|
| CRM/ consumer analytics（57） | 25.0%<br>26.8% |
| Banking（43） | 18.9%<br>19.2% |
| Health care/ HR（38） | 16.7%<br>13.1% |
| Education（37） | 16.2%<br>9.9% |
| Fraud Detection（32） | 14.0%<br>12.7% |
| Science（31） | 13.6%<br>10.3% |
| Social Networks（30） | 13.2%<br>6.6% |

续表

| Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters] 2011 % of voters 2010 % of voters | |
|---|---|
| Credit Scoring（29） | 12.7%<br>8.0% |
| Direct Marketing/ Fundraising（28） | 12.3%<br>11.3% |
| Insurance（28） | 12.3%<br>10.3% |
| Finance（26） | 11.4%<br>11.3% |
| Telecom / Cable（25） | 11.0%<br>10.8% |
| Retail（24） | 10.5%<br>8.0% |
| Medical/ Pharma（22） | 9.6%<br>8.0% |
| Biotech/Genomics（21） | 9.2%<br>5.6% |
| Government/Military（17） | 7.5%<br>6.1% |
| Travel / Hospitality（17） | 7.5%<br>1.4% |
| Advertising（16） | 7.0%<br>9.9% |

如前文所述，数据挖掘在营销方面的应用是最广泛的。表 1-1 显示的 CRM（用户关系管理，Customer Relationship Management）指的是公司对客户和潜在客户的管理模式。如何通过数据挖掘找

到并留住客户是所有公司一直在研究的问题。

如图 1-3 所示是 Gartner 公司在 2012 年 10 月按照地区划分的企业对于大数据的查询。

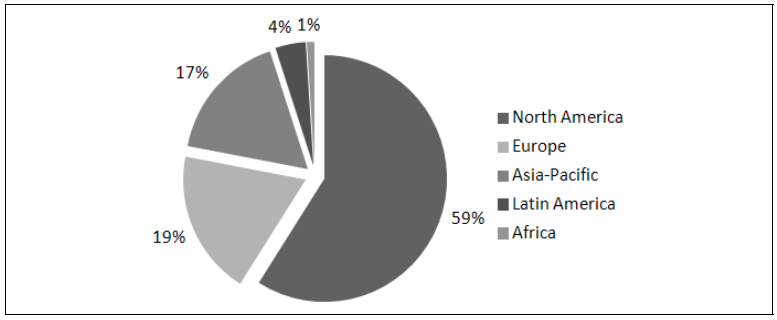



图 1-3　按地区划分的对大数据的查询

从图 1-3 中我们可以看到对大数据感兴趣的企业主要集中在北美和欧洲，占据了所有大数据查询的 78%。可想而知，在今后的若干年，大数据的发展在亚洲，包括中国在内，会有更大幅度的增长。

和 SaaS 软件即服务（Software as a Service）一样，数据挖掘应用领域在不久的将来可能会出现 DaaS 数据即服务（Data as a Service）或者 DMaaS 数据挖掘即服务（Data Mining as a Service）。换句话说，我们可以在给研究者或是企业提供在庞大的数据源基础之上，做专业的半自动分析服务。有了这样的服务，专业的数据挖掘专家可以合作采用算法、人工智能或是统计分析工具来更加充分地用好数据。

### 1.3.2　数据挖掘研究发展

由于现今数据量级的猛增以及人们对数据中潜在信息的重视，数据挖掘的研究受到越来越多的关注，在最近几年更是有了长足的进步，如以下几个方面：

- 对于大规模数据的存储、管理和使用，包括在分布式环境上建立数据仓库的方式方法。

- 知识发现语言的形式化描述和算法，即研究专门用于知识发现的数据挖掘语言。
- 数据挖掘过程中的可视化方法，使知识发现的过程能够更容易被用户理解，也便于在知识发现的过程中进行人机交互。
- 生物信息或基因（Bioinformatics/Genomics）的数据挖掘。
- Web 数据挖掘的各个方面。

另外，随着计算机计算能力的发展和业务复杂性的提高，数据的类型会越来越多、越来越复杂，使得数据挖掘发挥出越来越大的作用。从 2010 年开始，数据挖掘的研究者们加强了对于非常规数据处理的研究。非常规数据的类型或者比较复杂，或者是结构比较独特。对各种非结构化数据的开采，如文本数据、图形数据、图表、视频图像数据、音频数据（Data Mining for Audio &Video）乃至综合性多媒体数据，需要一些新的和更好的分析以及建立模型的方法，以便高质的处理这些复杂的数据，同时还会涉及为处理这些复杂或独特数据所做的一些工具和软件。

从实用性角度来讲，人们很关心的一个话题是文本数据挖掘。举个例子，在客户服务中心，把同客户的谈话转化为文本，再对这些文本数据做自然语言处理（Natural Language Processing，NLP）和情感分析（Sentiment Analysis）等，进而了解客户对服务的满意程度，客户的需求以及客户之间的相互关系等信息。无论是在数据结构还是在分析处理方法方面，文本数据挖掘和前面谈到的结构化数据挖掘都相差很大。文本数据挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的课题。目前市场上有一些类似文本数据挖掘的软件，但大部分方法只是把文本移来移去，或简单地计算某些词汇的出现频率，并没有真正达到我们在文本数据分析上真正想要的效果。自然语言处理是综合语言学和人工智能研究的一门专门的学科，不在本书讨论范围之内。

## 1.4 本书内容

与其他讲数据挖掘的书不同，本书的重点不在算法和理论本身，而是这些算法的应用，特别是在互联网上的应用。

本书引用了大量的实际案例，有时我们用了企业的原名，而还有一些，为了保护商业隐私，把真实的名字隐去了。另外，案例中的数据基本都是经过一些处理，和实际数字可能会有一定差别。而书中关于一些知名公司内部系统的运作流程和数据，我们通过互联网和各类文献上的信息，做了一些 Educated Guess（科学猜测），如果书中所讲和实际的情况略有出入，也请读者见谅。

在本书的第 2 章，主要介绍了数据挖掘的概念。先通过几个案例简单介绍数据挖掘是什么，然后通过人脸识别系统解释数据挖掘中的一个重要概念：分类。之后，会介绍分类算法的应用场景、相关技术和如何对分类效果做评估。接下来讲述的是商业智能，因为数据挖掘的最终目的就是要实现数据的价值，而商业智能是实现数据价值的一种重要方式。在第 2 章的最后介绍数据挖掘的一般流程，数据挖掘过程会有信息收集、数据集成、数据规约、数据清理、数据变换、数据挖掘实施过程、模式评估和知识表示等步骤。

第 3 章讲述的是数据仓库。对于面向海量数据的数据挖掘过程，数据仓库是不可或缺的基础。在介绍了数据仓库的基本概念和传统结构之后，以 Google 和 Facebook 的架构为例，讲述以云存储为基础的数据仓库。近两年人们把 Big Data 和 Hadoop 划上了等号，我们在这一章中简单介绍了 Hadoop，一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序。除了 Hadoop 之外，第 3 章还引进了非关系型的数据库的概念，介绍了 Hive 和 NoSQL。

第 4 章讲述数据挖掘的基本算法和原理。如果是一般的数据

挖掘理论书籍，这一章通常是重中之重，而因为本书的目的主要是为了讲述大数据挖掘的应用，算法的起源和原理不是我们的重点，所以并没有写得太详尽。在介绍数据挖掘十大经典算法之后，会着重讲述在本书中有较多应用的数据挖掘常用算法，包括分类算法、聚类算法、关联算法和序列挖掘算法。数据挖掘建模语言也是近几年研究的方向，在第 4 章最后的 4.7 节中，我们对建模语言 PMML 做了介绍。

第 5 章中讲述的是在大数据挖掘流程中，正式开始数据挖掘过程实施之前的一些准备工作，在仔细描述了为什么要做数据预处理之后，介绍了 8 个步骤中的第 2 步到第 5 步，也就是数据集成、数据规约、数据清理和数据变换。其中数据规约、数据清理和数据变换又称为数据预处理。

第 6 章以 R 语言为重点讲述数据挖掘中可以用到的各种工具。我们从 R 语言的历史说起，介绍了 R 语言的语法结构、程序包、接口和它强大的数据可视化功能，随后简单介绍 R 语言如何可以应用在数据挖掘之上。在第 6 章的第 2 部分，我们还介绍了其他的数据挖掘工具，商用的 MATLAB、IBM Intelligent Miner、SAS Enterprise Miner、SPSS Clementine 和开源工具 Weka 等。

前面 6 章做的是都是铺垫，从第 7 章开始我们以各种主要来自于互联网的案例来讲述数据挖掘的实际应用。

第 7 章讲述互联网上的日志分析。通过网站的日志分析，我们可以清楚的得知用户在什么 IP、什么时间点、用什么操作系统和什么浏览器下访问了您网站的哪个页面，是否访问成功，在您的网站上面点击了哪个链接进入了网站的另一个什么页面，以及离开您的网站去到哪里等。可以说日志分析是互联网数据挖掘的基础，包括后文中的电子邮件营销、互联网广告和电子商务都广泛使用了日志分析。

第 8 章主要讲述了两个专题：如何用数据挖掘手段通过海量数据的分析做好电子邮件营销是讨论的第一个专题，我们通过三

个实际案例讨论如何利用数据挖掘做好电子邮件营销；第二个专题是从企业邮箱的角度，举例为证，讨论如何通过数据挖掘分析辨别垃圾邮件。

第 9 章讲述的是数据挖掘在互联网广告上的应用。互联网广告目前是互联网公司最主要的收入来源。我们以网站联盟为例，一方面讲述互联网广告如何利用数据挖掘提升广告效果，另一方面介绍如何用数据挖掘方法抓出广告作弊行为。

在第 10 章中先阐述电子商务企业的需求，总结出电子商务面临的一些具体问题，然后再通过实际案例讲解怎样通过数据挖掘解决这些问题。本章我们讨论并试图解决的问题是如何通过数据充分了解客户并抓住客户。最后在 10.4 节中会通过案例来具体阐述数据挖掘在电子商务中的应用。我们会以两个电子商务企业为案例介绍数据分析和数据挖掘，包括各种相关算法的应用和实现等。

第 11 章主要讲述社会化媒体 SNS 上的 Web 挖掘，在介绍了现今比较时髦的“Like”概念之后，还讲述了互联网上的个性化，以及个人隐私的问题。

第 12 章主要讲述在移动互联网范畴中的数据挖掘问题。我们先讨论了移动互联网的特殊性。有了移动互联网的数据，我们可以真正实现用户行为定向，通过用户使用各种应用的习惯与场景，还原用户属性，了解用户兴趣和喜好，预测用户消费习惯和消费意图，实现真正的精准定向。在 12.2 节中我们以两个移动互联网应用的实际案例介绍了在 LBS 上数据挖掘的应用。最后作为结尾，在 12.3 节中讲述了我们在移动互联网上做数据挖掘所遇到的困难和问题。

本书在写作过程中参考了大量的中英文文献，这些文献除了在附录二和附录三中列举出来之外，还在每一章用到这些文献的篇章最后做了列表。其实还有许多和 Web 数据挖掘紧密相关的精彩的论文和书籍，由于作者时间有限，没有能够拜读并做引用，深感歉意。

数据挖掘是一个很大的课题，我们不可能通过简单的一本书覆盖它的全部，而且数据挖掘的应用和研究每天都有新的变化和发展。令人欣喜的是，很多数据挖掘学者和研究者在互联网上为大家提供了很多信息。在附录中为大家整理了一些相关资源，附录四列举的新浪微博，每天都有关于数据挖掘应用和研究的内容分享。笔者的新浪微博是@CarelessWhisper，欢迎关注。不过笔者平时微博发的不算多，基本也是有感随性而发，这和笔者的微博名是相一致的，因为微博本来就应该是“无心低语”（Careless Whisper）。

在附录 E 中列出了一些网址，很遗憾的是其中一些国外学者的博客在中国不能直接访问。

# 第 2 章
# 一小时了解数据挖掘

简而言之，数据挖掘（Data Mining）是有组织有目的地收集数据，通过分析数据使之成为信息，从而在大量数据中寻找潜在规律以形成规则或知识的技术。

在本章中，我们从数据挖掘的实例出发，并以数据挖掘中比较经典的分类算法入手，给读者介绍我们怎样利用数据挖掘的技术解决现实中出现的问题。

## 2.1 数据挖掘是如何解决问题的

本节通过几个数据挖掘实际案例来诠释如何通过数据挖掘解决商业中遇到的问题。2.1.1 节中关于“啤酒和尿不湿”的故事是数据挖掘中最经典的案例。而 Target 公司通过“怀孕预测指数”来预测女顾客是否怀孕的案例也是近来为数据挖掘学者最津津乐道的一个话题。

### 2.1.1 尿不湿和啤酒

很多人会问，究竟数据挖掘能够为企业做些什么？下面我们通过一个在数据挖掘中最经典的案例来解释这个问题——一个关于尿不湿与啤酒的故事。

超级商业零售连锁巨无霸沃尔玛公司（Wal Mart）拥有世界上最大的数据仓库系统之一。为了能够准确了解顾客在其门店的购买习惯，沃尔玛对其顾客的购物行为进行了购物篮关联规则分

析，从而知道顾客经常一起购买的商品有哪些。在沃尔玛庞大的数据仓库里集合了其所有门店的详细原始交易数据，在这些原始交易数据的基础上，沃尔玛利用数据挖掘工具对这些数据进行分析和挖掘。一个令人惊奇和意外的结果出现了：“跟尿不湿一起购买最多的商品竟是啤酒”！这是数据挖掘技术对历史数据进行分析的结果，反映的是数据的内在规律。那么这个结果符合现实情况吗？是否是一个有用的知识？是否有利用价值？

为了验证这一结果，沃尔玛派出市场调查人员和分析师对这一结果进行调查分析。经过大量实际调查和分析，他们揭示了一个隐藏在“尿不湿与啤酒”背后的美国消费者的一种行为模式：在美国，到超市去买婴儿尿不湿是一些年轻的父亲下班后的日常工作，而他们中有 30%～40%的人同时也会为自己买一些啤酒。产生这一现象的原因是：美国的太太们常叮嘱她们的丈夫不要忘了下班后为小孩买尿不湿，而丈夫们在买尿不湿后又随手带回了他们喜欢的啤酒。另一种情况是丈夫们在买啤酒时突然记起他们的责任，又去买了尿不湿。既然尿不湿与啤酒一起被购买的机会很多，那么沃尔玛就在他们所有的门店里将尿不湿与啤酒并排摆放在一起，结果是得到了尿不湿与啤酒的销售量双双增长。

按常规思维，尿不湿与啤酒风马牛不相及，若不是借助数据挖掘技术对大量交易数据进行挖掘分析，沃尔玛是不可能发现数据内这一有价值的规律的。

### 2.1.2 Target 和怀孕预测指数

关于数据挖掘的应用，最近还有这样一个真实案例在数据挖掘和营销挖掘领域广为流传。

美国一名男子闯入他家附近的一家美国零售连锁超市 Target 店铺（美国第三大零售商塔吉特）进行抗议：“你们竟然给我 17 岁的女儿发婴儿尿片和童车的优惠券。”店铺经理立刻向来者承认错误，但是其实该经理并不知道这一行为是总公司运行数据挖掘的结果。如图 2-1 所示。一个月后，这位父亲来道歉，

因为这时他才知道他的女儿的确怀孕了。Target 比这位父亲知道他女儿怀孕的时间足足早了一个月。


图 2-1　Target 怀孕预测指数示意图

Target 能够通过分析女性客户购买记录，“猜出”哪些是孕妇。他们从 Target 的数据仓库中挖掘出 25 项与怀孕高度相关的商品，制作“怀孕预测”指数。比如他们发现女性会在怀孕四个月左右，大量购买无香味乳液。以此为依据推算出预产期后，就抢先一步将孕妇装、婴儿床等折扣券寄给客户来吸引客户购买。

如果不是在拥有海量的用户交易数据基础上实施数据挖掘，Target 不可能做到如此精准的营销。我们将会在第 10 章具体分析 Target 的精准营销案例。

### 2.1.3　电子商务网站流量分析

网站流量分析，是指在获得网站访问量基本数据的情况下对有关数据进行的统计和分析，其常用手段就是 Web 挖掘。Web 挖掘可以通过对流量的分析，帮助我们了解 Web 上的用户访问模式。那么了解用户访问模式有哪些好处呢？

- 在技术架构上，我们可以合理修改网站结构及适度分配资源，构建后台服务器群组，比如辅助改进网络的拓扑设计，提高性能，在有高度相关性的节点之间安排快速

有效的访问路径等。

- 帮助企业更好地设计网站主页和安排网页内容。
- 帮助企业改善市场营销决策，如把广告放在适当的 Web 页面上。
- 帮助企业更好地根据客户的兴趣来安排内容。
- 帮助企业对客户群进行细分，针对不同客户制定个性化的促销策略等。

人们在访问某网站的同时，便提供了个人对网站内容的反馈信息：点击了哪一个链接，在哪个网页停留时间最多，采用了哪个搜索项、总体浏览时间等。而所有这些信息都被保存在网站日志中。从保存的信息来看，网站虽然拥有了大量的网站访客及其访问内容的信息，但拥有了这些信息却不等于能够充分利用这些信息。

那么如果将这些数据转换到数据仓库中呢？这些带有大量信息的数据借助数据仓库报告系统（一般称作在线分析处理系统），虽然能给出可直接观察到的和相对简单直接的信息，却也不能告诉网站其信息模式及怎样对其进行处理，而且它一般不能分析复杂信息。所以对于这些相对复杂的信息或是不那么直观的问题，我们就只能通过数据挖掘技术来解决，即通过机器学习算法，找到数据库中的隐含模式，报告结果或按照结果执行。

为了让电子商务网站能够充分应用数据挖掘技术，我们需要采集更加全面的数据，采集的数据越全面，分析就能越精准。在实际操作中，有以下几个方面的数据可以被采集：

- 访客的系统属性特征。比如所采用的操作系统、浏览器、域名和访问速度等。
- 访问特征。包括停留时间、点击的 URL 等。
- 条款特征。包括网络内容信息类型、内容分类和来访 URL 等。
- 产品特征。包括所访问的产品编号、产品目录、产品颜色、产品价格、产品利润、产品数量和特价等级等。

当访客访问该网站时，以上有关此访客的数据信息便会逐渐被积累起来，那么我们就可以通过这些积累而成的数据信息整理出与这个访客有关的信息以供网站使用。可以整理成型的信息大致可以分为以下几个方面：

- 访客的购买历史以及广告点击历史。
- 访客点击的超链接的历史信息。
- 访客的总链接机会（提供给访客的超级链接）。
- 访客总的访问时间。
- 访客所浏览的全部网页。
- 访客每次会话的产出利润。
- 访客每个月的访问次数及上一次的访问时间等。
- 访客对于商标总体正面或负面的评价。

在本书的第 7 章我们会具体讲述如何做互联网日志分析，在第 9 章的互联网广告应用中我们会讲述如何利用这些访客信息来提升广告效果，在第 10 章中我们会以实际的电子商务网站为例来介绍如何通过数据挖掘有效地为电子商务做好服务。

## 2.2 分类：从人脸识别系统说起

美国电视剧《反恐 24 小时》中有一集，当一个恐怖分子用手机拨打了一个电话，从 CTU（反恐部队）的计算机系统中便立刻发出恐怖分子出现的预警。很多好莱坞的大片中此类智能系统的应用也比比皆是，它能从茫茫人群中实时找出正在苦苦追踪的恐怖分子或间谍。而在 2008 年北京奥运会上，最引人注意的 IT 热点莫过于“实时人脸识别技术”在奥运会安检系统中的应用，这种技术通过对人脸关键部位的数据采集，让系统能够精确地识别出所有进出奥运场馆的观众身份。

目前人脸识别技术正广泛的应用于各种安检系统中，警方只需将犯罪分子的脸部数据采集到安检数据库，那么只要犯罪分子

一出现，系统就能精确地将其识别出来。现如今人脸识别技术已经相对成熟，谷歌在 Picasa 照片分享软件的工具中就已经加入了人脸识别功能。当然，人脸识别技术牵涉到隐私，是把双刃剑，谷歌在谷歌街景地图中故意将人脸模糊化，变得无法识别就是这个原因。如图 2-2 所示为人脸识别示意图。

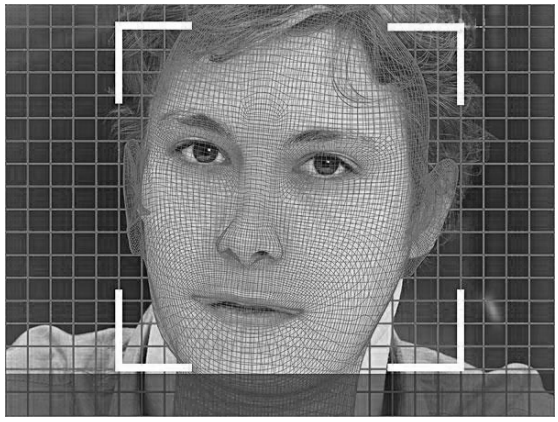


图 2-2　人脸识别示意图

虽然需要借力于其他技术，但是人脸识别中的主要技术还是来自于数据挖掘中的分类算法（Classification）。

让我们从一个最简单的事实来解释分类的思想。设想一下，一天中午，你第一次到三里屯，站在几家以前从未去过的餐厅门前，现在的问题是该选择哪家餐厅用餐。应该怎样选择呢？假设您没有带手机，无法上网查询，那么可能会出现如下两种情况：

- 第一种，你记起某位朋友去过其中一家，并且好像他对这家的评价还不错，这时，你很有可能就直接去这家了。
- 第二种，没有类似朋友推荐这类先验知识，你就只能从自己以往的用餐经历中来选择了，例如你可能会比较餐厅的品牌和用餐环境，因为似乎以前的经历告诉自己，品牌响、用餐环境好的餐厅可能味道也会好。

不管是否意识得到，在最终决定去哪家吃的时候，我们已经根据自己的判断标准把候选的这几家餐厅分类了，可能分成好、

中、差三类或者值得去、不值得去两类。而最终去了自己选择的那家餐厅，吃完过后我们自然也会根据自己的真实体验来判定我们的判断准则是否正确，同时根据这次的体验来修正或改进自己的判断准则，决定下次是否还会来这家餐厅或者是否把它推荐给朋友。

选择餐厅的过程其实就是一个分类的过程，此类分类例子是屡见不鲜的。在古时，司天监会依赖长时间积累的信息，通过观察天象对是否会有天灾做出分类预测。古人则通过对四季气候雨水的常年观察，总结出农作物最佳播种时间。在伯乐的《相马经》中，就通过简单分类区分出赢马的三条标准："大头小颈，弱脊大腹，小颈大蹄"。

其实在数据挖掘领域，有大量基于海量数据的分类问题。通常，我们先把数据分成训练集（Training Set）和测试集（Testing Set），通过对历史训练集的训练，生成一个或多个分类器（Classifier），将这些分类器应用到测试集中，就可以对分类器的性能和准确性做出评判。如果效果不佳，那么我们或者重新选择训练集，或者调整训练模式，直到分类器的性能和准确性达到要求为止。最后将选出的分类器应用到未经分类的新数据中，就可以对新数据的类别做出预测了。

### 2.2.1　分类算法的应用

本节将为大家介绍数据挖掘中的分类算法在一些行业中的代表性应用。我们将算法应用分为表述问题和解决过程两个阶段，表述问题即需要运用数据挖掘能够理解和处理的语言来阐述业务问题，最重要的是能够用正确且符合实际的方式把业务问题转化成数据挖掘问题，这往往决定了后续工作是否能有效的展开，尝试解决一个不符合实际的业务问题往往会使得数据挖掘的工作陷入数据的海洋中，既费时费力又得不到想要的结果。而解决过程，顾名思义就是将表述清楚的问题通过数据挖掘的方法加以解决的过程。在我们把业务领域的问题很清晰地转化为数据挖

掘领域的问题之后，解决问题也就变得相对直截了当。

分类算法的应用非常广泛，只要是牵涉到把客户、人群、地区、商品等按照不同属性区分开的场景都可以使用分类算法。例如我们可以通过客户分类构造一个分类模型来对银行贷款进行风险评估，通过人群分类来评估酒店或饭店如何定价，通过商品分类来考虑市场整体营销策略等。

在当前的市场营销行为中很重要的一个特点是强调目标客户细分。无论是银行对贷款风险的评估还是营销中的目标客户（或市场）细分，其实都属于分类算法中客户类别分析的范畴。而客户类别分析的功能也正在于此：采用数据挖掘中的分类技术，将客户分成不同的类别，以便于提高企业的决策效率和准确度。例如呼叫中心设计时可以分为呼叫频繁的客户、偶然大量呼叫的客户、稳定呼叫的客户和其他客户，以帮助呼叫中心寻找出这些不同种类客户的特征。这样的分类模型可以让呼叫中心了解不同行为类别客户的分布特征。

下面是几个做得比较成熟的具体分类应用描述和解决过程。

#### 2.2.1.1 直邮营销（Direct Mail）

直邮营销是直效行销的一种，是把传统邮件直接发送给消费者的营销方式，而且很多传统行业把直邮营销作为整个营销体系中一个重要的组成部分，涉及的行业主要是大型商场、大卖场、商业连锁店铺、专卖店等。当然由于直邮营销的应用很广，所以这种方式也同样适用于其他行业。

**案例阐述：**A 公司是一家汽车 4S 店，公司拥有完备的客户历史消费数据库，现公司准备举办一次高端品牌汽车的促销活动，为配合这次促销活动，公司计划为潜在客户（主要是新客户）寄去一份精美的汽车销售材料并附带一份小礼品。由于资源有限，公司仅有 1000 份材料和礼品的预算额度。

**表述问题：**这里新客户是指在店中留下过详细资料但又没有消费记录的客户。这次促销活动的要求是转化收到这 1000 份材料和礼品的新客户，让尽量多的新客户能够最终成为 4S 店的消

费客户。

**解决问题**：公司首先找出与这次促销活动类似的已经举办过的促销活动的历史消费数据，再将这个历史数据集中，把促销结果分成正反两类，正类用来表示可以最终消费的客户。通过历史数据的训练我们可以得出一个分类器，如果用的是决策树，我们还能够得出一个类似 If-Then（如果-就）的规则，而这个规则能够揭示参加促销活动并最终消费的客户的主要特征。由于分类结果最后可以表示成概率形式，如此，用经过测试集测试过的分类器对新客户进行分类，将得到的正类客户的概率由大到小排序，这样就可以生成一个客户列表，营销人员按着这个表由上至下数出前 1000 个客户并向他们寄出材料和礼品即可。

#### 2.2.1.2　客户流失模型

这一模型的应用出现在我国的移动通信行业，其目的主要是为了降低客户流失率。

**案例阐述**：我国的移动通信行业经过了前几年的高速发展，近一段时间的发展速度逐渐缓慢下来。注册用户常常处于一种动态变化的状态，即不断有老客户离网，又不断有新客户入网。大量的低消费客户和大量老客户的离网使得移动通信公司无法快速向前发展。

**表述问题**：当务之急在于降低客户流失率，这里需要解决的问题是如何找出这些将要流失的客户，如何采取适当的挽留措施减少客户的流失。

**解决问题**：我们需要建设客户流失模型。和直邮营销一样，其目的也是为了对新客户进行分类。只不过客户流失模型是为了找出那些不稳定易流失的客户。整个建模过程与直邮营销类似。移动通信企业的最大优势在于这类公司的规模往往很大，数据收集和存储的能力也比一般企业强很多，所以它们会拥有较详细的客户消费数据，这对于数据挖掘的最终成功有着非常重要的作用。

#### 2.2.1.3 垃圾邮件处理

**案例阐述**：对于企业和个人，如何处理垃圾邮件都是很头疼的一件事情。在盘石公司开发的磐邮系统中，每个客户可以有300G 的邮件储存容量，虽然有足够的容量容纳垃圾邮件，但是没有过滤掉的垃圾邮件仍然会造成糟糕的用户体验。

**表述问题**：如何对每个邮箱中收到的每封邮件进行处理，将有用邮件保留而过滤掉垃圾邮件是用户关心的一大问题。

**解决问题**：目前的垃圾邮件过滤方法主要是采用文本挖掘技术（Text Mining）。作为数据挖掘的重要分支，文本挖掘在数据挖掘传统方法的基础上引入了语义处理等其他学科知识。在垃圾邮件过滤的分类技术中最常见的是贝叶斯分类法。贝叶斯分类法主要是通过对邮件的信封标题、主题和内容进行扫描和判别。近来，因为垃圾邮件发送方式随着各家企业邮箱开发商的反垃圾技术的提升而变化，通过附件（PDF、图像等）方式发送垃圾邮件的专业户也越来越多，所以扫描的内容又增加了一项检查附件的工作。

我们会在第 8 章“数据挖掘和邮件营销”中再次讨论垃圾邮件的判别问题。

#### 2.2.1.4 信用卡分级

**案例阐述**：现如今金融行业的竞争异常激烈。在美国，出现在每一家邮箱里最多的信件恐怕就是信用卡邀请信。如何吸引合适的用户来使用信用卡，以及准确分析申请人的信用风险，是每个商业银行最关注也是最头痛的事情。银行要不惜一切代价吸引低风险高价值的客户，但是对于高风险的信用卡申请者要尽量避免。

**表述问题**：如何把信用卡申请者分类为低、中、高风险。

**解决问题**：我们需要建设客户风险模型对客户的风险进行分类。整个建模过程与直邮营销类似。不过因为行业的特殊性，申请表中包含了大量关于用户的个人信息，再加上通常会做的

客户信用查询，可以用来参考的数据维度比前面的三个案例都要多一些，所以相对来说建模的精准度也会高很多。

除了上面列出的四种典型问题之外，分类数据挖掘还有很多不同类型的应用，例如文献检索和搜索引擎中的自动文本分类技术，安全领域的入侵检测等。

不过，不是所有分类的场景使用分类数据挖掘都有实际操作性。美国政府曾在“9・11”发生后提出一项全面信息识别计划（Total Information Awareness Project），这项计划的目的是建立系统，利用数据挖掘技术对全美居民的通话记录和信用卡支付记录等海量数据信息进行分析，并利用这个系统来识别隐藏在美国的全部恐怖分子。除去涉及的个人隐私问题和海量数据如何获取和处理的问题之外，单纯从数据挖掘问题本身来说，这个计划的可行性就要打个大问号。假设通过数据挖掘技术建立了一个 99% 的分类器来识别恐怖分子，虽然这个分类器的精度已经是相当好了，但是整个美国一天之中可产生的相关数据保守估计就会有约十亿条，在产生如此庞大的增量情况下，这个 99%的分类器每天至少也要忽略掉近千万条可疑数据，那么就可以说这种分类器几乎毫无用处。可能是基于这个原因，2003 年这个计划被终止，虽然之后还是有若干个类似的计划被提出并尝试，但其效果都很有限。正如前所述，除非另辟捷径，否则这项计划能够成功实施的可能性很小。

### 2.2.2　数据挖掘分类技术

从分类问题的提出至今，已经衍生出了很多具体的分类技术。下面主要简单介绍四种最常用的分类技术，不过因为原理和具体的算法实现及优化不是本书的重点，所以我们尽量用应用人员能够理解的语言来表述这些技术。而且我们会在第 4 章再次给读者讲述分类算法和相关原理。

在我们学习这些算法之前必须要清楚一点，分类算法不会百

分百准确。每个算法在测试集上的运行都会有一个准确率的指标。用不同的算法做成的分类器（Classifier）在不同的数据集上也会有不同的表现。

#### 2.2.2.1 KNN，K 最近邻算法

K 最近邻（K-Nearest Neighbor，KNN）分类算法可以说是整个数据挖掘分类技术中最简单的方法。所谓 K 最近邻，就是 K 个最近的邻居，说的是每个样本都可以用它最接近的 K 个邻居来代表。

我们用一个简单的例子来说明 KNN 算法的概念。如果您住在一个市中心的住宅内，周围若干个小区的同类大小房子售价都在 280 万到 300 万之间，那么我们可以把你的房子和它的近邻们归类到一起，估计也可以售 280 万到 300 万之间。同样，您的朋友住在郊区，他周围的同类房子售价都在 110 万到 120 万之间，那么他的房子和近邻的同类房子归类之后，售价也在 110 万到 120 万之间。

KNN 算法的核心思想是如果一个样本在特征空间中的 K 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。KNN 方法在类别决策时，只与极少量的相邻样本有关。由于 KNN 方法主要靠周围有限的邻近样本，而不是靠判别类域的方法来确定所属类别，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。

#### 2.2.2.2 决策树（Decision Tree）

如果说 KNN 是最简单的方法，那决策树应该是最直观最容易理解的分类算法。最简单的决策树的形式是 If-Then（如果-就）式的决策方式的树形分叉。

比如下面这样一棵决策树，根据样本的相貌和财富两个属性把所有样本分成“高富帅”、“帅哥”、“高富”和“屌丝”四类。

```
If（obj.相貌=="帅"）then
    {
        If (obj.财富>=1000000000) then
            {
            print（obj.Name + "高富帅"）;
            }
        else
            {
            print（obj.Name + "是帅哥"）;
            }
    }
else
    {
        If (obj.财富>=1000000000) then
            {
            print（obj.Name + "是高富"）;
            }
        else
            {
            print（obj.Name + "是屌丝"）;
            }
    }
```

决策树上的每个节点要么是一个新的决策节点，要么就是一个代表分类的叶子，而每一个分支则代表一个测试的输出。决策节点上做的是对属性的判断，而所有的叶子节点就是一个类别。决策树要解决的问题就是用哪些属性充当这棵树的各个节点的问题，而其中最关键的是根节点（Root Node），在它的上面没有其他节点，其他所有的属性都是它的后续节点。在上面的例子中，（obj.相貌==“帅”）就是根节点，两个（obj.财富>=1000000000）是根节点下一层的两个决策节点，四个 print 标志着四个叶子节点，各自对应一个类别。

所有的对象在进入决策树之后根据各自的“相貌”和“财富”属性都会被归到四个分类中的某一类。

大多数分类算法（如下面要提的神经网络、支持向量机等）都是一种类似于黑盒子式的输出结果，你无法搞清楚具体的分类方式，而决策树让人一目了然，十分方便。决策树按分裂准则的不同可分为基于信息论的方法和最小 GINI 指标（Gini Index）方法等。

#### 2.2.2.3 神经网络（Neural Net）

在 KNN 算法和决策树算法之后，我们来看一下神经网络。

神经网络就像是一个爱学习的孩子，你教他的知识他不会忘记，而且会学以致用。我们把学习集（Learning Set）中的每个输入加到神经网络中，并告诉神经网络输出应该是什么分类。在全部学习集都运行完成之后，神经网络就根据这些例子总结出他自己的想法，到底他是怎么归纳的就是一个黑盒了。之后我们就可以把测试集（Testing Set）中的测试例子用神经网络来分别作测试，如果测试通过（比如 80%或 90%的正确率），那么神经网络就构建成功了。我们之后就可以用这个神经网络来判断事务的分类。

神经网络是通过对人脑的基本单元——神经元的建模和连接，探索模拟人脑神经系统功能的模型，并研制一种具有学习、联想、记忆和模式识别等智能信息处理功能的人工系统。神经网络的一个重要特性是它能够从环境中学习，并把学习的结果分别存储于网络的突触连接中。神经网络的学习是一个过程，在其所处环境的激励下，相继给网络输入一些样本模式，并按照一定的规则（学习算法）调整网络各层的权值矩阵，待网络各层权值都收敛到一定值，学习过程结束。然后我们就可以用生成的神经网络来对真实数据做分类。

#### 2.2.2.4 支持向量机 SVM（Support Vector Machine）

和上面三种算法相比，支持向量机的说法可能会有一些抽象。我们可以这样理解，尽量把样本中的从更高的维度看起来在一起的样本合在一起，比如在一维（直线）空间里的样本从二维平面上可以把它们分成不同类别，而在二维平面上分散的样本如果我们从第三维空间上来看就可以对它们做分类。

支持向量机算法的目的是找到一个最优超平面，使分类间隔最大。最优超平面就是要求分类面不但能将两类正确分开，而且使分类间隔最大。在两类样本中离分类面最近且位于平行于最优超平面的超平面上的点就是支持向量，为找到最优超平面，只要

找到所有的支持向量即可。对于非线性支持向量机，通常做法是把线性不可分转化成线性可分，通过一个非线性映射将低维输入空间中的数据特征映射到高维线性特征空间中，在高维空间中求线性最优分类超平面。

支持向量机算法是我们在做数据挖掘应用时很看重的一个算法，而原因是该算法自问世以来就被认为是效果最好的分类算法之一。

### 2.2.3　分类算法的评估

在整个分类数据挖掘工作的最后阶段，分类器（Classifier）的效果评价所占据的地位不容小视，正如前文所述，没有任何分类器能够百分百的正确，任何分类算法都会发生一定的误差，而在大数据的情况下，有些数据的分类本身就是比较模糊的。因此在实际应用之前对分类器的效果进行评估显得很重要。

对分类器的效果评价方法有很多，由于图形化的展示方式更能为大家所接受，这里介绍两种最常用的方式，ROC 曲线和 Lift 曲线来做分类器的评估。

在介绍两种曲线之前，为了方便说明，假设一个用于二分类的分类器最终得出的结果如表 2-1 所示。

表 2-1　混淆矩阵示意表

| 混淆矩阵 | 预 测 值 | | |
|---|---|---|---|
| 实际值 | | 0 | 1 |
| | 0 | A | B |
| | 1 | C | D |

这张表通常被称为混淆矩阵（Confusion Matrix）。在实际应用中，常常把二分类中的具体类别用 0 和 1 表示，其中 1 又常常代表我们关注的类别，比如直邮营销中的最终消费客户可以设定为 1，没有转化成功的客户设为 0。通信行业客户流失模型中的流失客户可设置为 1，没有流失的客户设置为 0。矩阵中的各个

数字的具体含义为，A 表示实际是 0 预测也是 0 的个数，B 表示实际是 0 却预测成 1 的个数，C 表示实际是 1 预测是 0 的个数，D 表示实际是 1 预测也是 1 的个数。

图 2-3 是一张 ROC 曲线图，ROC 曲线（Receiver Operating Characteristic Curve）是受试者工作特征曲线的缩写，该曲线常用于医疗临床诊断，数据挖掘兴起后也被用于分类器的效果评价。

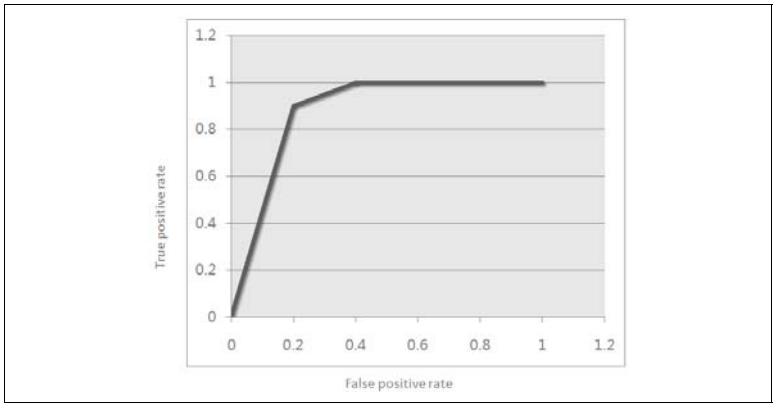



图 2-3　ROC 曲线图

如图 2-3 所示为一张很典型的 ROC 曲线图，从图中可以看出该曲线的横轴是 FPR（False Positive Rate），纵轴是 TPR（True Positive Rate）。首先解释一下这两个指标的含义：TPR 指的是实际为 1 预测也是 1 的概率，也就是混淆矩阵的 D/(C+D)，即正类（1）的查全率。FPR 指的是实际为 0 预测为 1 的概率即 B/(A+B)。

前面说过，分类中比较关心的都是正类的预测情况，而且分类结果常常是以概率的形式出现的，设定一个阈值，如果概率大于这个阈值那么结果就会是 1。而 ROC 曲线的绘制过程就是根据这个阈值的变化而来的，当阈值为 0 时，所有的分类结果都是 1，此时混淆矩阵中的 C 和 A 是 0，那么 TPR=1，而 FPR 也是 1，这样曲线达到终点。随着阈值的不断增大，被预测为 1 的个数会减少，TPR 和 FPR 同时减少，当阈值增大到 1 时，没有样本被预测为 1，此时 TPR 和 FPR 都为 0。由此可知，TPR 和 FPR 是同方向变化的，这点在上图中可以得到体现。

由于我们常常要求一个分类器的 TPR 尽量高，FPR 尽量小，表现在图中就是曲线离纵轴越近，预测效果就越好。为了更具体化，人们也通过计算 AUC（ROC 曲线下方的面积）来评判分类器效果，一般 AUC 超过 0.7 就说明分类器有一定效果。在图 2-3 中的 ROC 曲线中，曲线下方的面积 AUC 数值超过了 0.7，所以分类器是有一定效果的。

下面我们再来看 Lift 曲线的绘制。Lift 曲线的绘制方法与 ROC 曲线是一样的，不同的是 Lift 曲线考虑的是分类器的准确性，也就是使用分类器获得的正类数量和不使用分类器随机获取正类数量的比例。以直邮营销为例，分类器的好坏就在于与直接随机抽取邮寄相比，采用分类器的结果会给公司带来多少响应客户（即产生多少最终消费），所以 Lift 分类器在直邮营销领域的应用是相对比较广泛的。

由图 2-4 可以发现，Lift 曲线的纵轴是 Lift 值，它的计算公式是 $Lift = pv/k$，其中 $pv = D/(B+D)$，这个参数的含义是如果采用了分类器，正类的识别比例；而 $k = (C+D)/(A+B+C+D)$，表示如果不用分类器，用随机的方式抽取出正类的比例。这二者相比自然就解决了如果使用者用分类器分类会使得正类产生的比例会增加多少的问题。Lift 曲线的横轴 RPP（正类预测比例，Rate of Positive Predictions）的计算公式是 $RPP = (B+D)/(A+B+C+D)$。

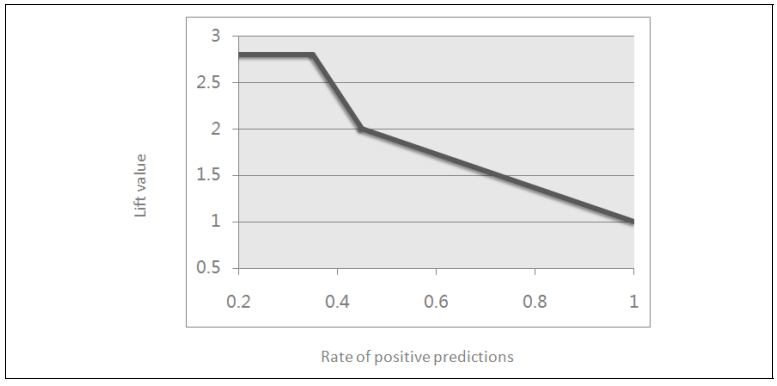



图 2-4　Lift 曲线图

Lift 曲线的绘制过程与 ROC 曲线类似，不同的是 Lift 值和 RPP 是反方向变化的，这才形成 Lift 曲线与 ROC 曲线相反的形式。

## 2.3 一切为了商业

马云在 2012 年网商大会上的演讲中说过：“假如我们有了一个数据预报台，就像为企业装上了一个 GPS 和雷达，企业的出海将会更有把握。”。这里的数据预报台就是下文所述的商业智能。

### 2.3.1 什么是商业智能（Business Intelligence）

数据挖掘的最终目的是要实现数据的价值，而商业智能是在企业中实现数据价值的最佳方式之一。商业智能（Business Intelligence，简称 BI）的概念最早是 Gartner 公司于 1996 年提出来的。当时将商业智能定义为一类由数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的，以帮助企业决策为目的技术及其应用。Gartner 公司的 Howard Dressner 把商业智能定义成为把数据转化成信息，并通过迭代发现（Iterative Discoveries）把信息转化成商业上可用的知识。

在我们看来，商业智能就是能够从（海量）业务和相关数据中提取有用的信息，把信息转化成知识，然后根据这些知识采用正确的商务行为的工具。在本书的范畴内，我们提到的 BI（商业智能）工具都是指在数据挖掘基础上的工具。如图 2-5 所示。


图 2-5　商业智能示意图

现在数据挖掘技术在商业应用中已经相当广泛，因为对数据

挖掘技术进行支持的三种基础技术已经发展成熟，这三种基础技术是：

- 海量数据收集和存储技术。
- 强大的计算机集群和分布式计算技术。
- 数据挖掘算法。

商业数据库现在正在以一个空前的速度增长，并且数据仓库正在广泛地应用于各种行业。对计算机硬件性能越来越高的要求，也可以用现在已经成熟的并行多处理机的技术来满足。另外数据挖掘算法经过了这 10 多年的发展也已经成为一种成熟、稳定，且易于理解和操作的技术。

现在面临的尴尬的境地是数据丰富，信息匮乏（Data Rich But Information Poor）。快速增长的海量数据，已经远远地超过了人们的理解能力，如果不借助强有力的工具，很难弄清大堆数据中所蕴含的知识。结果，重要决策只是基于制定决策者的个人经验，而不是基于信息丰富的数据。数据挖掘就这样应运而生，数据挖掘填补了数据和信息之间的鸿沟。Erik Brynjolfsson 曾经说过：有数据支持的（商业）决定总是更好的决定。

数据在商业运营上要能起到作用，我们必须要做到：

- 理解数据的上下文，明白数据到底支持商业运营的什么过程。
- 简化过程，使得数据更加便于管理。
- 在不同的渠道、应用和设备上整合数据。
- 丰富、匹配和清理数据，提高数据质量。
- 充分利用数据，比如整合关于消费者、市场和机会的数据。
- 选择合适的存储介质，比如私有云、公有云还是专门设计的云存储。
- 获取最终结果数据并在各种终端上用可视化方式展示（包括移动终端）。

在最开始制定商业智能数据战略时，考虑的不应该是技术，

而是从商业角度出发，看到底需要完成怎样的商业目标，再来制定数据挖掘过程。

比如在商业银行信用卡部门，我们需要做信用卡欺诈监测。商业目的很明确，就是要以最快的速度发现 90%以上的欺诈交易，而可以提供的数据就是之前所有的交易记录。那么如何判别某一个交易可能是欺诈行为呢？常用的数据挖掘方式是通过神经网络。我们通过正面和负面的实例训练这个神经网络，然后给每个交易打分，如果低于某个数值，那么就判定这条交易是正常的，否则就判定它为欺诈交易。

商业智能还有一个重要的原因是竞争。现在的企业竞争对象不一定来自身边，甚至不一定来自于同一个国家，商业竞争的全球化导致了中国企业必须提高对商业智能的重视，因为商业智能在欧美的企业中正相当普及。

当我们已经建立了一套完整的商业智能系统之后，可以通过如图 2-6 所示的流程来定期做数据分析。


图 2-6　商业智能分析示意图

下面我们对图 2-6 的商业智能分析中的各个阶段做个简单的解释。

- 看趋势：即观察关键考核指标 KPI 数据的日、周、月、季度、年的图表曲线趋势。KPI 数据是上升了还是下降了。关联的其他相关 KPI 曲线，是否呈现了应该有的关联性。环比同比的百分比如何等。
- 寻找变异：即找到单一 KPI 数据中的异常值，或者关联数据中非关联的异常部分。
- 分析原因：当我们找到了异常值，就需要分析造成这一异常的原因。看异常发生的时间节点，看内部和外部的关联活动，看问题发生原因的构成，并把原因分解成独立的元素一一列出，标出权重，哪些是相对影响较大的，

哪些又是可能的原因等。

- 制定对策：在正确的分析了相关原因后，就需要给出解决方法和策略。一般来说，一个原因对应一个解决策略。当然也可能有多个解决策略对应于同一个原因。我们选择最切合实际，最可执行的对策和行动策略。

### 2.3.2　数据挖掘的九大定律

数据挖掘通用流程 CRISP-DM 的缔造者之一 Tom Khabaza 曾总结了在数据挖掘上的九大定律，如下所示。

（1）Business Goals Law：每个数据挖掘解决方案的根源都是有商业目的的。

（2）Business Knowledge Law：数据挖掘过程的每一步都需要以商业信息为中心。

（3）Data Preparation Law：数据挖掘过程前期的数据准备工作要超过整个过程的一半。

（4）NFL Law：NFL（没有免费午餐，No Free Lunch）。对于数据挖掘者来说没有免费的午餐，数据挖掘的任何一个过程都是来之不易的。

（5）Watkins' Law：此定律以此命名是因为 David Watkins 首次提出这个概念。这个定律说的是在数据的世界里，总是有模式可循的。您找不到规律不是因为规律不存在，而是因为您还没有发现它。

（6）Insight Law：数据挖掘可以把商业领域的信息放大。

（7）Prediction Law：预测可以为我们增加信息。

（8）Value Law：数据挖掘模式的精准和稳定并不决定数据挖掘过程的价值，换句话说技术手段再精妙，没有商业意义和合适的商业应用是没有价值的。

（9）Law of Change：所有的模式都会变化。

上面这九条其实归根到底就是一条，商业决定数据挖掘。数据挖掘各类技术和算法的飞速发展不能让我们偏离以商业行为

为核心的方向，只是纯粹为了追求高深的技术而忽略或损害到商业目的就本末倒置了。

## 2.4 数据挖掘很纠结

数据挖掘的世界既是地雷阵，同时又是金矿。大量的数据没能被及时处理，称得上是暴殄天物。虽然通过保存相关数据，我们可以保证以后对数据信息的方便使用，但是对于工作量日趋繁重的数据保存工作，很多企业可能还是选择荒废部分数据。大数据时代已经来临，不管有多大困难，我们从现在开始都需要考虑评估和集成数据挖掘应用。即使不能找到合适的数据挖掘方法来处理数据，至少我们需要用数据仓库把原始数据保留起来，以供将来使用。

下面列举一些我们在给企业做数据挖掘时看到的问题：

- 对于数据挖掘需要解决的问题，很少有现成的解决方案，而且于某个问题，可能有多种数据挖掘算法可以使用，但通常只有一个最好的算法。当我们选择了一个数据挖掘算法时，首先要弄清楚它是否适合想解决的问题。如果本身方法选择不合适，那么再好的执行也没有用。我们在第 6 章会介绍常用的数据挖掘算法以及它们所适合处理的问题。
- 从市场角度来看，数据挖掘依旧面临其他因素的挑战。数据挖掘非常有前景，但是市场中数据噪声太多，会导致数据价值大大降低。以无线营销为例，大量的虚假应用下载和使用以及虚假好评差评等数据严重干扰了数据的准确性，大大降低了数据的价值。
- 在中国，数据挖掘市场整体来说还不成熟。首先在意识上，一些商业领袖们对数据挖掘将信将疑，不愿意做投入；另一方面，采用了数据挖掘的公司只追求最后的结

果，而对数据挖掘过程、数据的存储、数据挖掘结果的知识积累和呈现不重视。

- 数据挖掘有时导出的结果是不完善的，每次导出的结果和应用的数据集直接相关。如果数据集发生变化，就需要重新进行挖掘。如果没有考虑数据变化而盲目采用数据变化之前的策略，那么结果是不可预料的。

这些问题都是确实存在的，其中关于市场的问题在一定时间之后会有好转，而数据挖掘过程中的这些问题就需要数据分析师和数据应用使用者提高自己的经验来解决了。

## 2.5 数据挖掘的基本流程

数据挖掘有很多不同的实施方法，如果只是把数据拉到 Excel 表格中计算一下，那只是数据分析，不是数据挖掘。本节主要讲解数据挖掘的基本规范流程。CRISP-DM 和 SEMMA 是两种常用的数据挖掘流程。

### 2.5.1　数据挖掘的一般步骤

从数据本身来考虑，数据挖掘通常需要有信息收集、数据集成、数据规约、数据清理、数据变换、数据挖掘实施过程、模式评估和知识表示 8 个步骤。

步骤（1）信息收集：根据确定的数据分析对象，抽象出在数据分析中所需要的特征信息，然后选择合适的信息收集方法，将收集到的信息存入数据库。对于海量数据，选择一个合适的数据存储和管理的数据仓库是至关重要的。

步骤（2）数据集成：把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。

步骤（3）数据规约：如果执行多数的数据挖掘算法，即使是在少量数据上也需要很长的时间，而做商业运营数据挖掘时数

据量往往非常大。数据规约技术可以用来得到数据集的规约表示，它小得多，但仍然接近于保持原数据的完整性，并且规约后执行数据挖掘结果与规约前执行结果相同或几乎相同。

步骤（4）数据清理：在数据库中的数据有一些是不完整的（有些感兴趣的属性缺少属性值）、含噪声的（包含错误的属性值），并且是不一致的（同样的信息不同的表示方式），因此需要进行数据清理，将完整、正确、一致的数据信息存入数据仓库中。不然，挖掘的结果会差强人意。

步骤（5）数据变换：通过平滑聚集、数据概化、规范化等方式将数据转换成适用于数据挖掘的形式。对于有些实数型数据，通过概念分层和数据的离散化来转换数据也是重要的一步。

步骤（6）数据挖掘过程：根据数据仓库中的数据信息，选择合适的分析工具，应用统计方法、事例推理、决策树、规则推理、模糊集，甚至神经网络、遗传算法的方法处理信息，得出有用的分析信息。

步骤（7）模式评估：从商业角度，由行业专家来验证数据挖掘结果的正确性。

步骤（8）知识表示：将数据挖掘所得到的分析信息以可视化的方式呈现给用户，或作为新的知识存放在知识库中，供其他应用程序使用。

数据挖掘过程是一个反复循环的过程，每一个步骤如果没有达到预期目标，都需要回到前面的步骤，重新调整并执行。不是每件数据挖掘的工作都需要这里列出的每一步，例如在某个工作中不存在多个数据源的时候，步骤（2）便可以省略。

步骤（3）数据规约、步骤（4）数据清理、步骤（5）数据变换又合称数据预处理。在数据挖掘中，至少 60%的费用可能要花在步骤（1）信息收集阶段，而其中至少 60%以上的精力和时间花在了数据预处理过程中。

## 2.5.2　几个数据挖掘中常用的概念

除了 2.2 节中所述的分类，还有一些概念是我们在数据挖掘中常用的，比如聚类算法、时间序列算法、估计和预测以及关联算法等。我们将在本节中介绍几个常用概念以加深读者对数据挖掘的理解。

### 2.5.2.1　聚类

所谓聚类，就是类或簇（Cluster）的聚合，而类是一个数据对象的集合。

和分类一样，聚类的目的也是把所有的对象分成不同的群组，但和分类算法的最大不同在于采用聚类算法划分之前并不知道要把数据分成几组，也不知道依赖哪些变量来划分。

聚类有时也称分段，是指将具有相同特征的人归结为一组，将特征平均，以形成一个“特征矢量”或“矢心”。聚类系统通常能够把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（Subset），这样在同一个子集中的成员对象都有相似的一些属性。聚类被一些提供商用来直接提供不同访客群组或者客户群组特征的报告。聚类算法是数据挖掘的核心技术之一，而除了本身的算法应用之外，聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。

图 2-7 是聚类算法的一种展示。图中的 Cluster1 和 Cluster2 分别代表聚类算法计算出的两类样本。打“+”号的是 Cluster1，而打“〇”标记的是 Cluster2。

在商业中，聚类可以帮助市场分析人员从消费者数据库中区分出不同的消费群体，并且概括出每一类消费者的消费模式或者消费习惯。它作为数据挖掘中的一个模块，可以作为一个单独的工具以发现数据库中分布的一些深层次的信息，或者把注意力放在某一个特定的类上以作进一步的分析并概括出每一类数据的特点。

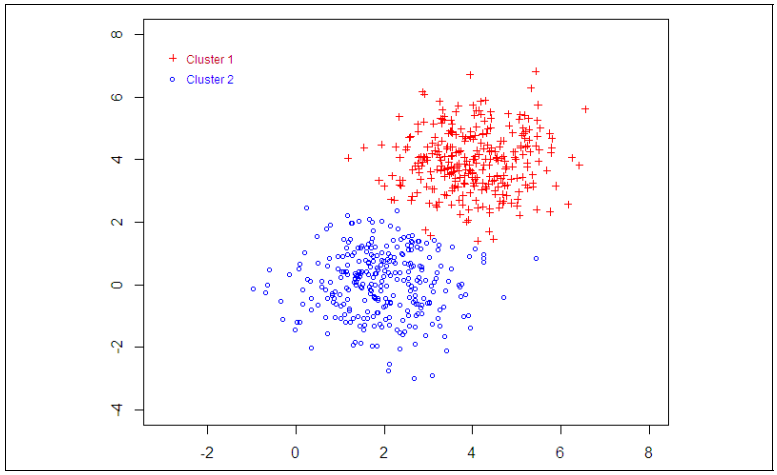



图 2-7　聚类算法示意图

聚类分析的算法可以分为划分法（Partitioning Methods）、层次法（Hierarchical Methods）、基于密度的方法（Density-Based Methods）、基于网格的方法（Grid-Based Methods）和基于模型的方法（Model-Based Methods）等。

比如，下面几个场景比较适合应用聚类算法，同时又有相应的商业应用：

- 哪些特定症状的聚集可能预示什么特定的疾病？
- 租同一类型车的是哪一类客户？
- 网络游戏上增加什么功能可以吸引哪些人来？
- 哪些客户是我们想要长期保留的客户？

聚类算法除了本身的应用之外还可以作为其他数据挖掘方法的补充，比如聚类算法可以用在数据挖掘的第一步，因为不同聚类中的个体相似度可能差别比较大。例如，哪一种类的促销对客户响应最好？对于这一类问题，首先对整个客户做聚集，将客户分组在各自的聚集里，然后对每个不同的聚集，再通过其他数据挖掘算法来分析，效果会更好。

我们会在 4.4 节详细介绍聚类算法是如何实现的。本书中多次提到的 RFM 模型也是基于聚类算法的数据挖掘模型。而在营销领域的客户关系管理中，RFM 聚类模型也是最经常被使用的

一种模型。

#### 2.5.2.2　估测和预测

估测（Estimation）和预测（Prediction）是数据挖掘中比较常用的应用。估测应用是用来猜测现在的未知值，而预测应用是预测未来的某一个未知值。估测和预测在很多时候可以使用同样的算法。估测通常用来为一个存在但是未知的数值填空，而预测的数值对象发生在将来，往往目前并不存在。

举例来说，如果我们不知道某人的收入，可以通过与收入密切相关的量来估测，然后找到具有类似特征的其他人，利用他们的收入来估测未知者的收入和信用值。还是以某人的未来收入为例来谈预测，我们可以根据历史数据来分析收入和各种变量的关系以及时间序列的变化，从而预测他在未来某个时间点的具体收入会是多少。

估测和预测在很多时候也可以连起来应用。比如我们可以根据购买模式来估测一个家庭的孩子个数和家庭人口结构。或者根据购买模式，估测一个家庭的收入，然后预测这个家庭将来最需要的产品和数量，以及需要这些产品的时间点。

对于估测和预测所做的数据分析可以称作预测分析（Predictive Analysis），而因为应用非常普遍，现在预测分析被不少商业客户和数据挖掘行业的从业人员当作数据挖掘的同义词。

我们在数据分析中经常听到的回归分析（Regression Analysis）就是经常被用来做估测和预测的分析方法。所谓回归分析，或者简称回归，指的是预测多个变量之间相互关系的技术，而这门技术在数据挖掘中的应用是非常广泛的。在第 4 章中的分类算法和序列算法都可以运用到回归的技术。

#### 2.5.2.3　决策树

在所有的数据挖掘算法中，最早在 2.2.2 节中提到的决策树可能是最容易让人理解的数据挖掘过程。决策树本质上是导致做出某项决策的问题或数据点的流程图。比如购买汽车的决策树可

以从是否需要 2012 年的新型汽车开始，接着询问所需车型，然后询问用户需要动力型车还是经济型车等，直到确定用户所最需要的车为止。决策树系统设法创建最优路径，将问题排序，这样，经过最少的步骤，便可以做出决定。

据统计，在 2012 年，被数据挖掘业者使用频率最高的三类算法是决策树、回归和聚类分析。而且因为决策树的直观性，几乎所有的数据挖掘的专业书籍都是从某一个决策树算法开始讲起的：如 ID3/C4.5/C5.0，CART，QUEST，CHAID 等。

有些决策树做得很精细，用到了数据大部分的属性，这时，我们可能闯入了一个误区，因为在决策树算法上我们需要避免的一个问题是把决策树构建得过大，过于复杂。过于复杂的决策树往往会过度拟合（Over-Fitting），不稳定，而且有时候无法诠释。这时我们可以把一棵大的决策树分解成多棵较小的决策树来解决这一问题。

我们来看一个商用的决策树实例。图 2-8 中展示的是用 IBM SPSS Modeler 数据挖掘软件构建的一棵决策树，是美国商业银行用以判断客户的信用等级的决策树模型。

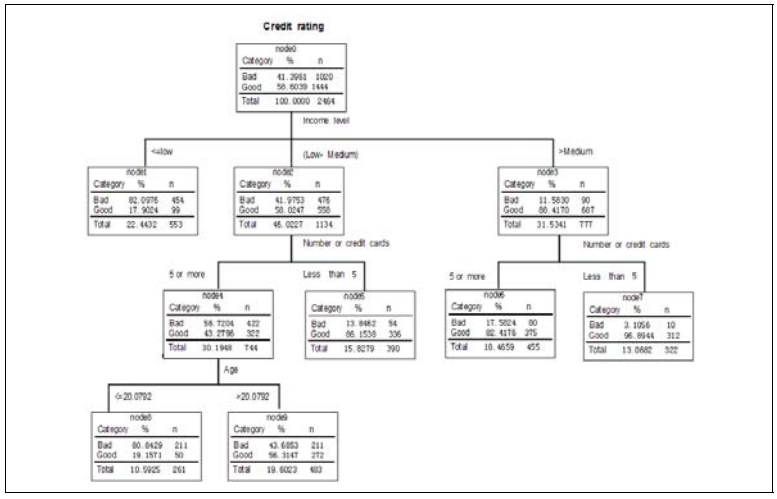



图 2-8　信用决策树示意图

图 2-8 是根据收入、信用卡数量和年龄构建的决策树，并以

80%的准确率作为划分的阈值。第一个分支查的是收入，设立了两个关键数据分隔点，按照收入把人群先划分成 3 组：低收入、中等收入和高收入。其中低收入的节点直接变成叶子节点，这组人中 82.0976%的人的信用等级是差的（Bad），而且信用卡个数或者年龄对信用等级的分类没有帮助。决策树的第二层判断是根据已经拥有的信用卡个数。以此作为判断，高收入人群可以再做划分。其中拥有卡个数在 5 个或以上的，82.4176%信用等级是优质的（Good），而拥有卡的数量在 5 张以下的，高达 96.8944%的人信用等级是优质的。因为这棵树一共有 6 个叶子节点，所以我们最终划分出 6 组人群，其中有一组信用等级为优质的人群占比 56.3147%，是无法判断的。其中在数据上表现最好的是高收入而信用卡个数在 5 张以下的人，把他们判断为优质信用等级有 96.8944%的准确率。

如果我们手里还有别的数据，比如是否有房有车，是否结婚等，那么通过测试，可以进一步提高这棵决策树的精度。

### 2.5.3　CRISP-DM

1999 年，在欧盟（European Commission）的资助下，由 SPSS、DaimlerChrysler、NCR 和 OHRA 发起的 CRISP-DM Special Interest Group 组织开发并提炼出 CRISP-DM（CRoss-Industry Standard Process for Data Mining），进行了大规模数据挖掘项目的实际试用。

CRISP-DM 提供了一个数据挖掘生命周期的全面评述。它包括项目的相应周期，它们的各自任务和这些任务的关系。在这个描述层，识别出所有关系是不可能的。所有数据挖掘任务之间关系的存在是依赖用户的目的、背景和兴趣，最重要的还有数据。SIG 组织已经发布了 CRISP-DM Process Guide and User Manual 的电子版。CRISP-DM 的官方网址是 http://www.crisp-dm.org/。在这个组织中，除了 SPSS 是数据挖掘软件提供商，其他的几个发起者都是数据挖掘的应用方。所以 CRISP-DM 和 SPSS 自有开

发的 SPSS Modeler 契合度非常好。

一个数据挖掘项目的生命周期包含六个阶段。这六个阶段的顺序是不固定的，我们经常需要前后调整这些阶段。这依赖每个阶段或是阶段中特定任务的产出物是否是下一个阶段必须的输入，图 2-9 中箭头指出了最重要的和依赖度高的阶段关系。

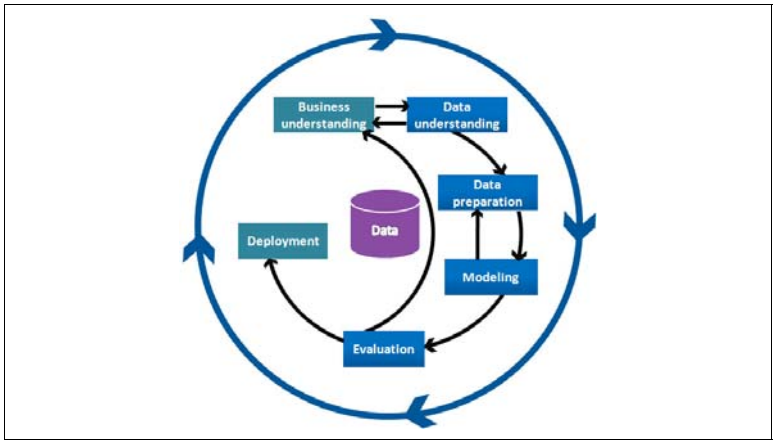



图 2-9　CRISP-DM 数据挖掘过程示意图

图 2-9 中最外面这一圈表示数据挖掘自身的循环本质，每一个解决方案发布之后代表另一个数据挖掘的过程也已经开始了。在这个过程中得到的知识可以触发新的，经常是更聚焦的商业问题。后续的过程可以从前一个过程中得到益处。

我们把 CRISP-DM 的数据挖掘生命周期中的六个阶段，也就是图 2-9 中的概念解释如下：

- 业务理解（Business Understanding）

最初的阶段集中在理解项目目标和从业务的角度理解需求，同时将这个知识转化为数据挖掘问题的定义和完成目标的初步计划。

- 数据理解（Data Understanding）

数据理解阶段从初始的数据收集开始，通过一些活动的处理，目的是熟悉数据，识别数据的质量问题，首次发现数据的内部属性，或是探测引起兴趣的子集去形成隐含信息的假设。

- 数据准备（Data Preparation）

数据准备阶段包括从未处理的数据中构造最终数据集的所有活动。这些数据将是模型工具的输入值。这个阶段的任务能执行多次，没有任何规定的顺序。任务包括表、记录和属性的选择，以及为模型工具转换和清洗数据。

- 建模（Modeling）

在这个阶段，可以选择和应用不同的模型技术，模型参数被调整到最佳的数值。一般，有些技术可以解决一类相同的数据挖掘问题。有些技术在数据形成上有特殊要求，因此需要经常跳回到数据准备阶段。

- 评估（Evaluation）

到这个阶段，你已经从数据分析的角度建立了一个高质量显示的模型。在开始最后部署模型之前，重要的事情是彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键目的是确定是否有重要业务问题没有被充分的考虑。在这个阶段结束后，一个数据挖掘结果使用的决定必须达成。

- 部署（Deployment）

通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得的知识需要便于用户使用的方式重新组织和展现。根据需求，这个阶段可以产生简单的报告，或是实现一个比较复杂的、可重复的数据挖掘过程。在很多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作。

除了 CRISP-DM 之外，还有 SEMMA 也是通用的标准数据挖掘流程。SEMMA（Sample，Explore，Modify，Model，Assess 的英文首字母缩写）的意思是抽样、检查、修改、设立模型和评估，是由 SAS 公司所倡导的。

### 2.5.4　数据挖掘的评估

评价一个数据挖掘系统主要从准确性、性能、功能性、可用性和辅助功能五个主要方面来考虑。

- 准确性

评估数据挖掘系统最关键的因素是准确性。通过在数据挖掘系统上执行算法做的预测和分类的准确率，我们可以判断系统中的算法是否合理，数据采集是否全面以及数据预处理工作是否完善。

- 性能

该系统能否在我们需要的商业平台运行；软件的架构是否能连接不同的数据源；操作大数据集时，性能变化是线性的还是指数的；运算的效率到底怎样，能否符合实际应用需求；是否基于某种开源框架；是否易于扩展；运行的稳定性等。

- 功能性

该系统是否提供足够多样的算法；能否避免挖掘过程黑箱化；软件提供的算法能否应用于多种类型的数据；用户能否调整算法和算法的参数；软件能否从数据集随机抽取数据建立预挖掘模型；能否以不同的形式表现挖掘结果等。

- 可用性

系统的用户界面是否友好；可视化效果是否好；是否易学易用；系统面对的用户是初学者，高级用户还是专家；错误报告对用户调试是否有很大帮助；应用的领域是专攻某一专业领域还是适用多个领域等。

- 辅助功能

是否允许用户更改数据集中的错误值或进行数据清洗；是否允许值的全局替代；能否将连续数据离散化；能否根据用户制定的规则从数据集中提取子集；能否将数据中的空值用某一适当均值或用户指定的值代替；能否将一次分析的结果反馈到另一次分析中，等等。

对于不同的数据挖掘算法，我们采用的评价方式是不同的。

在 2.2.3 节中我们提到了用来评估分类器的混淆矩阵（Confusion Matrix），这里的图 2-10 所示是混淆矩阵的另外一种表现方式。

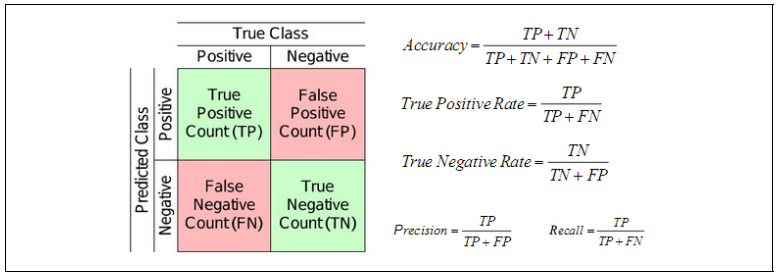



图 2-10　混淆矩阵示意图

一个数据挖掘系统最终的评价在于是否能够产生商业价值。如果没有商业价值，再完美的系统也是没有意义的。

在本书中多次讲述的关联算法，我们采用的标准是用两个概念来表示的，这两个分别为支持度和置信度。关于支持度和置信度的概念，我们会在 4.4 节中介绍。

## 2.5.5　数据挖掘结果的知识表示

数据挖掘系统最后的结果需要以一种美观和直观的方式呈现给用户。不幸的是，在中国乃至其他亚洲地区，数据可视化的工作被严重忽略。我见到国内数据挖掘的可视化展现在很多时候是用微软的 Office 来呈现的。

我们来看一下国外的数据挖掘业者是怎样用直观的图表方式展示数据的。图 2-11 是根据英国国家统计局 2012 年的统计数据整理的，是在不同行业男女平均收入差距的图表，图中显示的是人均收入为 25000 英镑的行业中男女的工资差距。在此可以很直观地看到在同一行业中，男人平均要比女人的收入高。

Google 为数据分析和数据挖掘提供了一个开放的作图工具 Google Chart，你可以输入网址 https://developers.google.com/chart/进行试用。

你可以很方便地在 Google Chart 中植入数据，例如可以直接从 Google 的网站上把程序复制粘贴到你的网页上来显示数据。图 2-12 是在 Google Chart 上用世界银行（World Bank）的数据整理出的按照地区来划分的受孕率和平均寿命的分布图。关于如何利

用 Google Chart 来编程，您可以参考 Google 提供的线上文档：

https://developers.google.com/chart/interactive/docs/quick_start

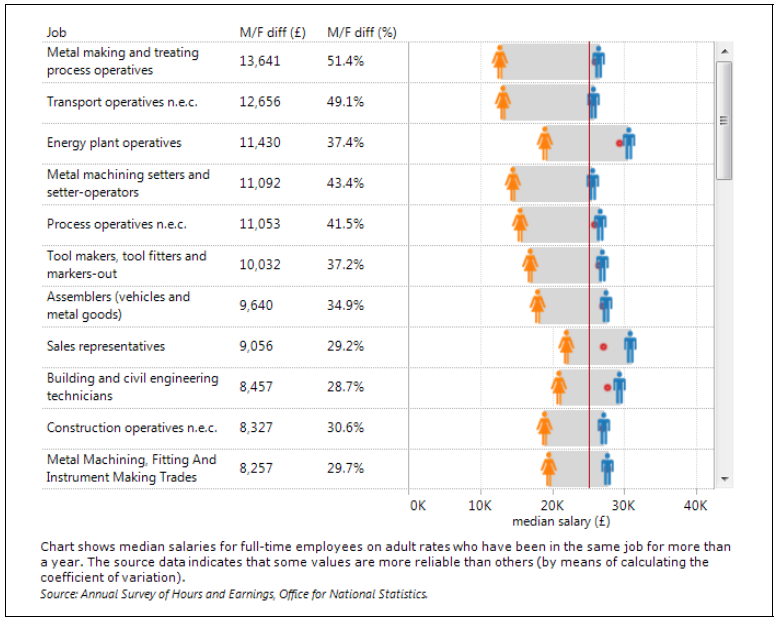



图 2-11　英国男女平均工资差距示意图

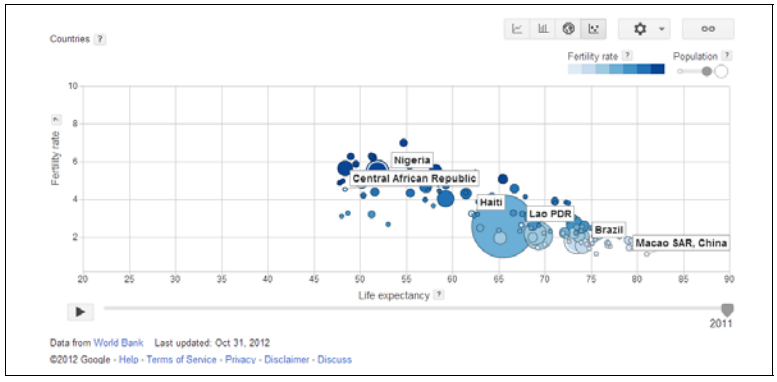



图 2-12　世界受孕率和平均寿命对比图

从图 2-12 中可以很直观地看到，一般来说，越是经济发达的地区，人们的平均寿命越长，但是受孕率就越低。图 2-12 中的中非共和国（Central African Republic），平均寿命只有 48.3 岁，

而受孕率却高达 4.55。作为对比，我们看澳门（Macao SAR，China），平均寿命达到 81 岁，而受孕率只有 1.12。

图 2-13 是根据美国健康局数据所做的糖尿病分布图，是用 Tableau Software 公司的免费软件做的，下载地址为 http://www.tableausoftw-are.com/public/gallery/geography-diabetes。在这个网页上你可以调节右下角的三个关于肥胖率、穷困率和白人比例的开关。调节之后，可以很直观地发现：肥胖率越高，糖尿病患者比例越高；穷困率越高，糖尿病患者比例越高；白人占比越低，糖尿病患者比例越高。

Tableau Software 是最近两年最火的数据可视化工具，用以显示最终数据挖掘结果是没有问题的。但是遗憾的是如果我们需要展示纯原始数据，数据量如果过大则显示效果不能保证。不过，数据可视化是数据挖掘学者们的重要研究方向之一。在不久的将来，我们一定会看到一个像 Tableau Software 一样做得如此形象的图形展示程序，而这样的程序应当会是建立在一个类似 Hadoop（见 3.5.2 节）和 NoSQL（见 3.5.5 节）的分布式数据系统之上的。

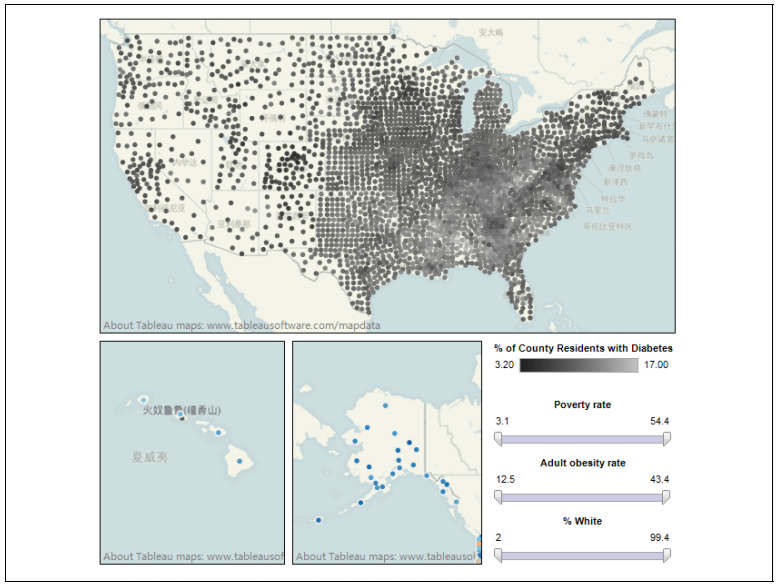



图 2-13　糖尿病占比示意图

如果追求图像展现的酷炫视觉效果，那么你必须要好好浏览网站 http://visual.ly/，它是 2012 年最火的视觉可视化社区。图 2-14 截自该网站，展示的是 Wikipedia 中有地理位置的文章标示。亮度和文章的密集度成正比。最亮的地方，比如西欧和美国加州及东北地区。

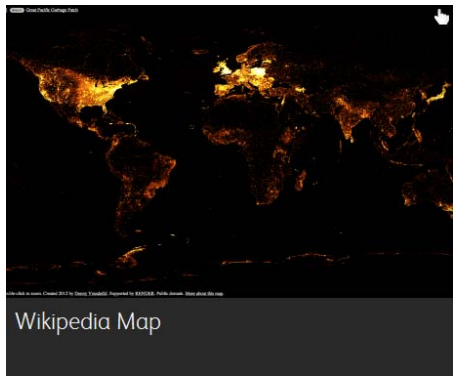



图 2-14　维基百科带地理位置文章发表示意图

图 2-15 也来自 http://visual.ly/，展示的是芬兰首都人民的年龄和负债率的对比，采用三维效果，以展示年龄和负债率对比在各个年份的变化。

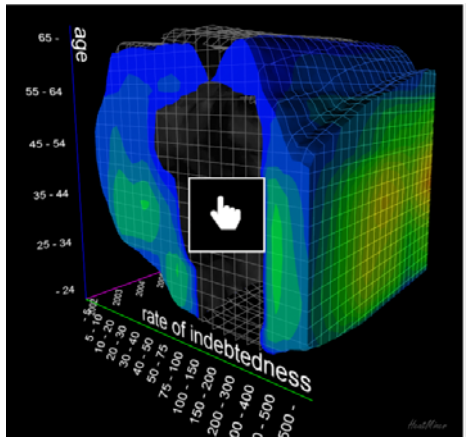



图 2-15　芬兰首都人民的年龄和负债率的对比示意图

除了刚才提到的这些互联网上的数据图形展示工具，我们在第 6 章的 R 语言介绍中会举例说明如何用 R 语言开源工具来作图。

所谓开源，指的是软件开发者把软件系统的原始代码公开，使得其他的软件开发者和爱好者可以对软件进行修改。在本书中隆重推出的 R 语言和 Hadoop 等都是开源软件。

## 2.6 本章相关资源

- 本章相关参考文献：

[1] Brynjolfsson, Erik, Hitt, Lorin M. and Kim, Heekyung Hellen: *Strength in Nu mbers: Ho w Do es Da ta-Driven Decisionmaking Affect Firm Performance*, 2011-4-22.

[2] Thuraisingham, B. *Web D ata Mi ning a nd A pplications in Business Intelligence and Counter-Terrorism*. Crc Press, 2003.

- 本章相关网址：

[1] http://www.crisp-dm.org/

[2] http://www.kdnuggets.com

[3] http://www.tableausoftware.com

[4] https://developers.google.com/chart/

[5] http://www.hbr.org

[6] http://visual.ly/

[7] http://blogs.hbr.org/

[8] http://www.khabaza.com/

# 第 3 章
# 数据仓库——数据挖掘的基石

在大数据的前提下，如何解决数据存储是一个大问题。数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的用于支持管理决策的数据集合。数据仓库系统是一个信息整合平台，从业务处理系统获得数据，通常以星型模型或雪花模型进行数据组织，并为用户提供从数据中获取信息和知识的各种手段。

对于面向海量数据的数据挖掘过程，数据仓库是不可或缺的基础。在本章中我们会讲述数据仓库的概念和原理，以及适合大规模互联网数据处理的数据仓库。

## 3.1 存放数据的仓库

数据仓库（Data Warehouse），顾名思义，是一个数据的仓库，而作为仓库，就一定有存放、组织、归类和货物准备的功能。所以数据仓库就是数据存放、组织归类，并准备好给顾客使用的地方。

目前企业级的数据库和应用都建立在传统的关系型数据库上，然而面对动辄上亿以至万亿条数据的查询分析，传统方式越来越力不从心。相关研究表明，2009～2020 年，全球数字信息量将实现 44 倍的增长，其中需要管理的文件数将增加 67 倍，总存储容量将增长 30 倍。企业在 PB 级甚至 EB 级的数据中寻找相关信息无异于大海捞针，制定信息驱动决策的成本和复杂性将与日俱增。

在这个数据量爆炸式增长的大数据时代，面对日益增长的非结构化和多结构化数据洪流的冲击，企业如何管理、分析数据，发掘数据价值并形成洞察力，已经成为企业提升竞争力的关键因素。

数据仓库就是在这样的背景之下得到快速发展的。

### 3.1.1　数据仓库的定义

目前，数据仓库一词尚没有一个完全统一的定义，因而很多所谓的数据仓库系统在盛名之下其实难符。著名的数据仓库专家 W.H. Inmon 曾给予如下的概念描述：数据仓库（Data Warehouse）是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，用于支持管理决策。

对于 Inmon 提出的数据仓库的概念，我们可以从两个层次予以理解。首先，数据仓库主要用于支持决策，面向分析型数据处理，它不同于普通企业现有的操作型数据库；其次，数据仓库可以对多个异构的数据源有效集成，集成后按照主题进行重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

对每个企业来说，数据仓库的建设是一个系统工程，是一个不断建设、发展、完善的过程，通常需要较长的时间。这就要求各企业对整个系统的建设提出一个全面、清晰的远景规划及技术实施蓝图，将整个项目的实施分成若干个阶段，分步实施，在每个阶段按照快速原型法予以实施，不断迭代修正，力求每一步都可见效，不仅可较快从当前投资中获得收益，而且可以在已有的基础上，结合其他的业务系统，逐步构建起完整、健壮的数据仓库系统。

传统数据库在日常的管理事务处理中获得了巨大的成功，但是对管理人员的决策分析要求却无法满足。因为，管理人员常常希望能够对组织中的大量数据进行分析，了解业务的发展趋势。而传统数据库只保留了当前的业务处理信息，缺乏决策分析所需

要的大量历史信息。为满足管理人员的决策分析需要，就需要在数据库的基础上产生适应决策分析的数据环境——数据仓库。

根据 Inmon 定义数据仓库概念的含义，数据仓库拥有以下四个特点：

- 面向主题（Subject Oriented）

操作型数据库的数据组织是面向事务处理任务的，各个业务系统之间各自分离，而数据仓库中的数据通常是按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。典型的主题有顾客、产品和账目等。

- 集成的（Integrated）

面向事务处理的操作型数据库通常与某些特定的应用相关，不同数据库之间相互独立，并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库、源文件等进行数据抽取、清理的基础上经过系统加工、汇总和整理得到的，消除了源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

- 相对稳定的（Non-Volatile）

操作型数据库中的数据通常会与应用程序交互，会被实时更新，数据根据需要及时发生变化。数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询。一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作相对较少，通常只需要定期的加载、刷新。甚至对于大部分的应用，只需要“数据访问”一个操作。

- 反映历史变化的（Time Variant）

操作型数据库主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含大量相关的历史信息，系统记录了企业从过去某一时间（如开始建立应用数据仓库的时间点）一直到目前的各个阶段的信息。因为数据挖掘过程可能需要这些历史数据做

定量分析和预测。

一般来说，数据仓库有星型模型和雪花模型两大类。所谓星型模型指的是数据仓库以一个大的中心表格为核心，而一组小的表格补充中心表格中每一维的数据。雪花模型是在星型模型的基础上把数据进一步分解到每个小的表格中。

### 3.1.2　数据仓库和数据库

数据仓库是 Data Warehouse，而数据库是 Data Base，在中文上是一字之差，而从英文字面上来讲，后者应该叫“数据基础”，Base 翻译成“库”有意译的成分。数据库本身是文件系统的革命性升级版本，在数据库出现以前，人们把信息存储在计算机系统的某一个文件中。数据库发展到数据仓库也是互联网时代的一次变革。

归根到底，数据仓库和传统数据库的不同之处在于数据的不同。

- 在数据仓库中，数据包含了过去的数据以及综合的、集成的和提炼过的信息，结构相对灵活。
- 大量数据会进入数据仓库，但是一旦加入之后，对于数据修改和更新的操作会比较少。

以 Google 和 Facebook 为首的互联网公司，存储和处理的数据量在传统的数据仓库之上又有了提升，而且增加了大量的非结构化数据，基于云存储的数据仓库已经和传统的一体式数据仓库又有所不同。我们可以称新的数据仓库为分布式数据仓库（Distributed Data Warehouse），或者用一个时髦的词汇：分布式数据云存储（Distributed Data Cloud）。图 3-1 展示了数据仓库的发展历程。

数据仓库通常存有海量的数据。TB、PB 级别的数据仓库比比皆是。在这样的数据量情况下，做到数据查询和即时灵活访问对于数据仓库的系统设计是有相当高的要求的。

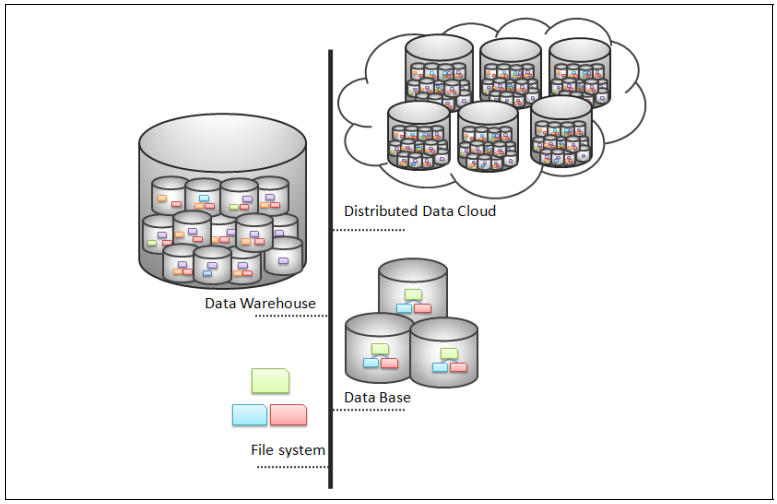



图 3-1　数据仓库发展示意图

在美国，从 20 世纪 90 年代末到 21 世纪初，采用传统的技术构建的数据仓库系统，通常软件成本是在 1000 万到 1500 万美元，需要花两年左右的时间来构建。而在目前的成熟技术之上，软件成本大幅下降，但是平均也需要 150 万美元，三个月的时间才能构筑成型。如果采用了全开源技术，软件成本本身可能是免费的，但是在开源工具之上做的定制化增值开发还是需要很多费用的。至于费用是多少，这就要看数据仓库的规模和性能需求。

在数据仓库上的数据流架构、数据管理架构、业务数据架构、数据库组织的性能以及安全性等各种细节问题，就不在本书的讨论范围之内了。

## 3.2 传统的数据仓库介绍

数据仓库是竞争非常激烈的一个软件行业分支，各大软件公司在数据仓库上投放了大量的资源，而公司并购行为也相当频繁。比如行业内最知名的 SPSS 公司在 2009 年以 12 亿美金的现金被 IBM 收购。SAP 也在 2010 年以 58 亿美元收购了 Sybase 数

据库提供商。各大厂商通过产品完善和收购集成，逐渐形成一套各自的全面数据仓库解决方案。如表 3-1 所示。

表 3-1　数据仓库知名厂商列表

| 公司名称 | 数据库产品 | 数据仓库工具 | ETL 工具 | 报　表 | OLAP | 数据挖掘工具 |
|---|---|---|---|---|---|---|
| Teradata | Teradata | Teradata RDBMS/Teradata MetaData Services | Teradata ETL Automation | BTEQ | 无 | Teradata Warehouse Miner |
| Oracle | Oracle | Oracle Warehouse Builder | Oracle Warehouse Builder | Oracle Reports | Oracle Express/ Discover | Oracle Data Miner |
| IBM | DB2 | IBM DWE Design Studio | IBM WebSpere DataStage | IBM Cognos | IBM DB2 OLAP Server | IBM Intelligent Miner/ IBM SPSS Clementine |
| Microsoft | SQL Server | SQL Server Managerment Studio | SSIS | SSRS | SSAS | SQL Server Data Mining |
| SAS | 无 | SAS Warehouse Administtrator | SAS ETL Studio | SAS Report Studio | SAS OLAP Server | SAS Enterprise Miner |
| SAP | Sybase IQ | PowerDesigner/ Warehouse Control Center | Sybase IQ InfoPrimer | BusinessObjects | Sybase IQ OLAP | 无 |

表 3-1 中的各个产品都已经经过多年的实践验证，商用化程度比较高，已经成为标准化的产品。比如对数据量要求很高的 MySpace 和 AT &T 使用的是 Teradata 平台上的数据仓库，而对数据精度要求很高的中国农业银行采用的是 SAP 平台上的数据仓库等。

我们下面来看一下两个 IT 龙头企业 IBM 和 Microsoft 的数据仓库系统。

IBM 公司的数据仓库产品称为 DB2 Data Warehouse Edition，它结合了 DB2 数据服务器的长处和 IBM 收购的 SPSS 商业智能基础产品，集成了用于仓库管理、数据转换、数据挖掘以及 OLAP 分析和报告的核心组件，提供了一套基于可视数据仓库的商业智能解决方案。如图 3-2 所示。

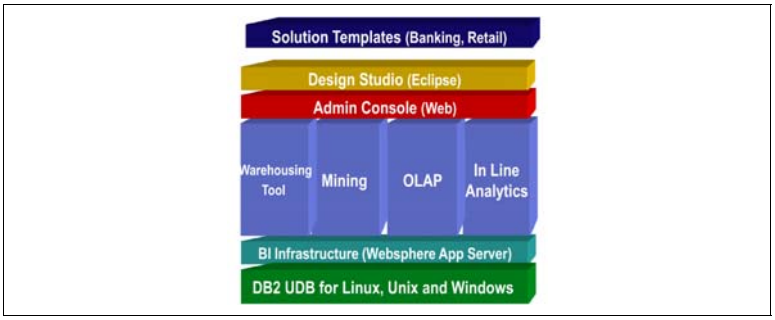



图 3-2　IBM 的 DB2 数据仓库示意图

微软的 SQL Server 提供了多种服务和工具来实现数据仓库系统的整合。主要有 SSIS（SQL Server Integration Service，整合服务工具），SSAS（SQL Server Analysis Service，分析工具）和 SSRS（SQL Server Reporting Service，报告工具）。这几个以 SS 开头的服务工具为用户提供了可用于构建分析应用程序所需的各种特殊工具和功能，可以实现数据仓库系统需要的建模、ETL、建立查询分析或图表、定制 KPI、建立报表和构造数据挖掘应用及发布等功能。如图 3-3 所示。

可以说在商用数据仓库系统中，微软的 BI 体系是把微软多个分裂的产品堆在一起的一个数据仓库系统，提供的功能最为全面，也是相对低成本的组合，不过产品之间的耦合度也相对较低。

有不少数据仓库服务提供商在近来的数据挖掘技术发展基础上，也开始走开源大数据技术路线。目前两个比较有代表性的方案就是 Cloudera 和 Hortonworks。总体来看，大数据市场的竞争还是比较激烈的，并没有出现一家或者数家独占鳌头的情况。

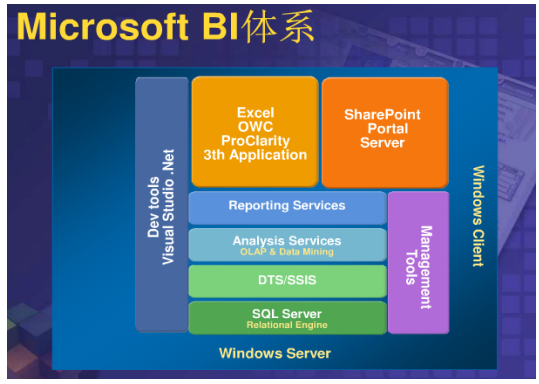



图 3-3　微软的商业智能系统示意图

## 3.3 数据仓库基本结构

企业数据仓库的建设，是以现有企业业务系统和大量业务数据的历史积累为基础的。数据仓库不是静态的概念，只有把信息及时交给需要这些信息的使用者，供他们做出改善其业务经营的决策，信息才能发挥作用，信息才有意义。而把信息加以整理归纳和重组，并及时提供给相应的管理决策人员，是数据仓库的根本任务。因此，从总体产业链的角度看，数据仓库建设需要一个过程，是一个浩大的工程。

图 3-4 是数据仓库结构的示意图，表示了在数据仓库中数据处在四个不同阶段的数据流（Data Flow）：处理前的数据来自不同的数据源；清理数据是在 ETL 过程中发生；数据存储在数据仓库中；数据仓库中用类似 OLAP 的服务来做数据的分析处理。

我们按照图 3-4 的层次对数据仓库系统的各个部分做一个简单介绍：

数据源（Data Source）是数据仓库系统的基础，是整个系统的数据来源。通常包括企业内部信息和外部信息。内部信息包括存放于关系数据库中的各种业务处理数据和各类文档数据。外部

信息包括各类法律法规、市场信息和竞争对手的信息等。

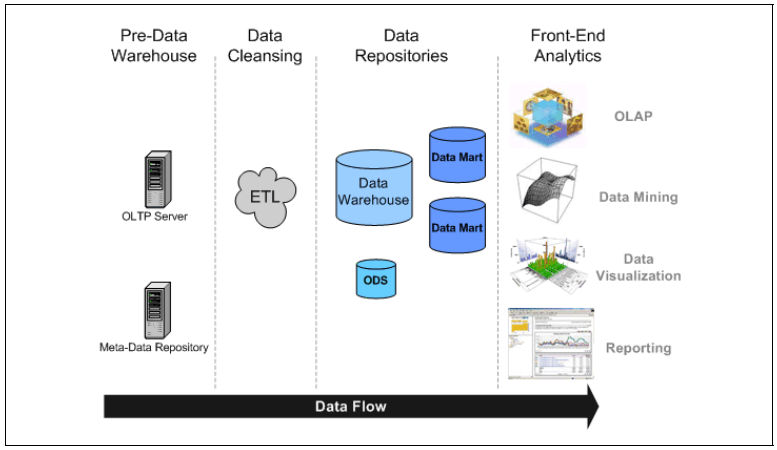



图 3-4　数据仓库结构示意图

ETL（Extract Transform Load，即数据抽取、转换、装载的过程）是数据仓库上的专门用语。作为数据仓库的核心和灵魂，能够按照统一的规则集成并提高数据的价值，负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。在技术上主要涉及增量、转换、调度和监控等几个方面的处理。

- 抽取（Extract）：将数据从各种原始的业务系统中读取出来。
- 转换（Transform）：按照预先设计好的规则将抽取的数据进行转换、清洗，以及处理一些冗余、歧义的数据，统一本来异构的数据格式。
- 装载（Load）：将转换完的数据按计划增量或全部导入到数据仓库中。

数据存储（Data Repository）是数据的存储与管理，也是整个数据仓库系统的核心。数据仓库的真正关键是数据的存储和管理。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。要决定采用什么产品和技术来建立数据仓库的核心，则需要从数据仓库的技术特点着手分析。针对现有各业务系统的数据，进行抽取、清理，并有效集成，

按照主题进行组织。数据仓库按照数据的覆盖范围可以分为企业级数据仓库和部门级数据仓库（通常称为数据集市）。传统的数据仓库通常都是建立在商用关系数据库 RDBMS 之上。比如 IBM 的系统是基于自有的 DB2，Oracle 基于自有的 Oracle 数据库，Teradata 基于自有的 Teradata 数据库，SAP 基于 Sybase IQ 而 Microsoft 也是基于自有的 SQL Server。

前端工具主要包括各种报表工具、查询工具、数据分析工具、数据挖掘工具以及各种基于数据仓库或数据集市的应用开发工具。其中数据分析工具主要针对 OLAP 服务器，报表工具、数据挖掘工具主要针对数据仓库。

OLAP，也称联机分析处理（On Line Analytical Processing）系统，OLAP 是数据仓库系统的主要应用。OLAP 服务器对分析需要的数据进行有效集成，按多维模型予以组织，以便进行多角度、多层次的分析，并发现趋势。

报表（Reporting）是企业管理的基本措施和途径，是企业的基本业务要求，也是实施 BI 战略的基础。报表可以帮助企业访问、格式化数据，并把数据信息以可靠和安全的方式呈现给使用者，深入洞察企业运营状况。

## 3.4 OLAP 联机分析处理

联机分析处理（OLAP）系统是数据仓库系统最主要的应用，专门设计用于支持复杂的分析操作，侧重对决策人员和高层管理人员的决策支持，可以根据分析人员的要求快速、灵活地进行大数据量的复杂查询处理，并且以一种直观而易懂的形式将查询结果提供给决策人员，以便他们准确掌握企业（公司）的经营状况，了解对象的需求，制定正确的方案。

http://olap.com/w/index.php/OLAPEducationWiki 是专门为 OLAP 打造的维基网站，里面有一些有用的信息。

OLAP 的概念最早由 E.F. Codd 在 1993 年提出。E.F. Codd 在提出 OLAP 概念时指出 OLAP 必须满足以下的 12 条规则：

（1）有多维度的视角。

（2）对用户透明。

（3）访问性好。

（4）提供报告的性能要稳定，不能因为维度的增加而变差。

（5）采用客户端/服务器架构。

（6）数据的每个维度都相当。

（7）对稀疏矩阵有动态优化功能。

（8）多用户支持。

（9）对于跨域的计算不做任何限制。

（10）直观的数据操作。

（11）灵活的报告体系。

（12）任意多的维度和维度集合。

再仔细重温 E.F. Codd 提出的这 12 条规则，我们可以发现传统的 OLAP 有一点 20 世纪的味道，已经不足以满足大数据中的数据挖掘了。比如第 6 条“数据的每个维度都相当”对于非结构化的数据是没有意义的。

简单来说，OLAP 是由使用者所主导，使用者先有一些假设，然后利用 OLAP 来查证假设是否成立。而数据挖掘则是用来帮助使用者产生假设的。所以在使用 OLAP 工具时，使用者是自己在做探索，但数据挖掘是用工具帮助做探索。

早期的 OLAP 只能做一些简单的统计，而不能发现其中一些深层次的有关系的规则。在 OLAP 使用数据挖掘技术之后，在 OLAP 中挖掘多层次、多维度的关联规则成为一个很自然的过程，因为 OLAP 本身的基础就是一个多层多维分析的工具。

用户往往希望能在数据仓库中随意选择各种相关的数据，在不同的细节层次上进行分析，以各种不同的形式呈现知识。用户希望基于 OLAP 的挖掘能够让有效的数据挖掘方法进行探索性的数据分析。可以提供在不同数据集、不同细节上的挖掘，可以对

数据进行切片、切块、展开、过滤等各种类型的操作。然后再加上一些可视化的工具，就能大大的提高数据挖掘的灵活性和能力。

当我们将 OLAP 和数据挖掘技术结合在一起就形成了一个新的体系 OLAM（On-Line Analytical Mining）。随着数据仓库和 OLAM 技术研究的深入，可以预见越来越多的数据将经过整合、预处理，从而存入数据仓库中。因为在当前，数据仓库上最多的应用就是进行统计、建立多维以及 OLAP 的分析工作。

## 3.5 云存储上的数据仓库

如第 1 章所述，未来大数据将会遵循消费化模式，核心基础设施将作为服务或应用程序来提供，通过互联网实现。数据分析和数据可视化将会在原始数据基础上作为一套标准的服务，并允许用户创建自己的数据模型。

传统的数据挖掘系统都在向海量数据的方向发展。以 Teradata 为代表，在传统的关系数据库上面引入 MapReduce 构建大数据分析平台。它的优势是既可以借鉴 MapReduce 过程并行机制的优势，同时也发挥已有 MPP 数据仓库并行处理的能力。

是海量的，又是通过互联网来实现的，这其实就是云的概念。因为我们认为所谓的云计算，就是两点：

- 可无限延伸扩展的。
- 基于互联网的服务。

只要满足这两点条件，那就是云。基于云的软件服务，就是云服务。基于云的存储系统，就是云存储。而数据仓库正在向云存储这个方向努力。

### 3.5.1　Google 公司的云架构

说起云存储，就必然要提到 Google（谷歌）公司。Google

是最早提出云计算概念的公司，也是大数据时代的先行者。Google 旗下的 Google 搜索、Google 地图、Gmail 邮箱、Picasa 图像存储和 YouTube 视频分享等诸多基于互联网的产品服务都建立在海量数据之上。

面对大数据，Google 采用了分布式存储和并行计算方法，构建了 GFS 文件系统、MapReduce 计算模型和 BigTable 非关系型数据库组成的基础平台，支撑起自家产品和服务。GFS、MapReduce、BigTable 可以称为 Google 基础平台的三个支柱。如图 3-5 所示。

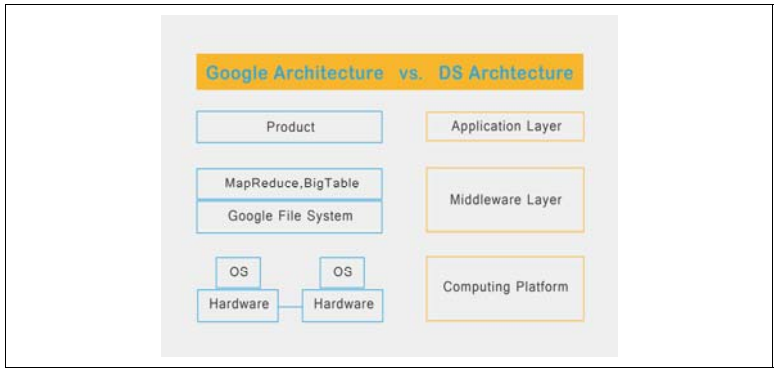



图 3-5　谷歌系统结构和传统系统结构对比图

### 3.5.1.1　GFS 文件系统

佩奇和布森在斯坦福大学时设计和实现了称为 BigFiles 的文件系统，用于存储下载的网页。在 BigFiles 成果的基础上，Google 公司成立之后，面对服务器数量的不断增多和数据量不断增长的情况，谷歌工程师研制出 GFS（Google File System）文件系统。该文件系统解决了大数据在廉价硬件上的分布式存储问题，和早期的文件系统相比有不少新颖之处。如图 3-6 所示。

首先该文件系统认为组件失效不再是意外，而是一种系统接受的正常现象。Google 为了降低成本没有采用超级计算机，而是把所有产品和服务建立在大量廉价服务器甚至普通 PC 组成的集群之上，由集群代替超级计算机提供存储和计算能力。Google

构成集群的单台服务器性能一般、可靠性差且经常出问题。Google 工程师以此为基础假设，在系统的整体处理能力和容错性上精心设计，GFS 文件系统内建数据冗余和容错机制，可以自动应对单台服务器宕机和数据丢失风险。

其次 GFS 文件系统是专为大文件存储而设计的，对大文件读/写操作在参数和通道上进行了优化。GFS 上存储的数据文件通常都很大，一般在 100MB 以上，甚至几个 GB。这也和由大量小文件组成的传统文件系统环境非常不同。GFS 把大文件分块存储，文件块（File Chunk）固定大小为 64MB。

GFS 做的假设是，对大文件最频繁的两项操作是顺序读取和在文件末尾追加新数据，而数据一旦写入之后，改写的需求极少，而文件随机写的操作几乎不存在。如果需要执行随机写的操作，整个大文件都需要重新生成。

GFS 由单一主服务器（Master）和众多存储服务器（Chunk Server）组成，主服务器上存储文件的名字空间等元数据，存储服务器存放具体数据。数据都有多份冗余，以文件块的形式在多台存储器上存放。

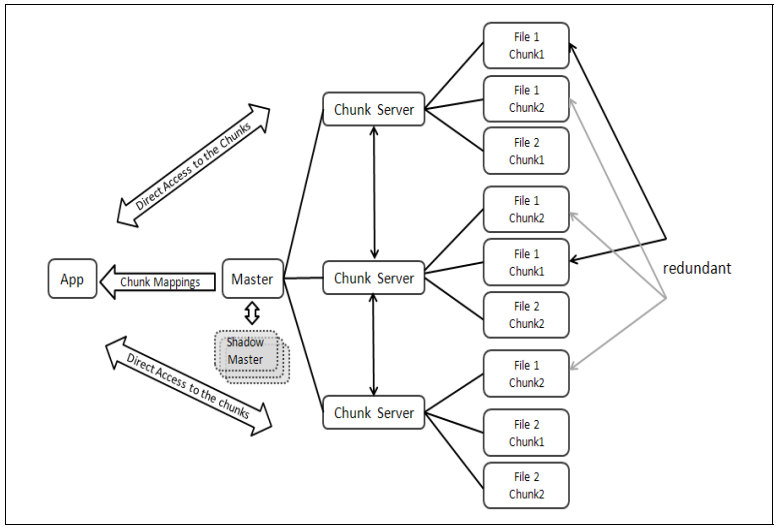



图 3-6　谷歌系统结构示意图

GFS 虽然很美妙，但技术的发展日新月异。在 Google，GFS 已经成为昨日黄花。目前，Google 开发了类似于 GFS，但是全新的实时分布式文件系统 Colossus（大石像），而且 Google 搜索、Gmail 邮箱、Google 文档和 YouTube 视频分享都已经平移到 Colossus 上了。关于 Colossus 的细节不在本书讨论范围之内。

#### 3.5.1.2 MapReduce 模型和框架

数据量的不断增长迫使越来越多的程序采用并行计算来处理问题，然而开发并行计算程序难度不小，一些传统应用，如网页爬取、Web 请求的日志分析等如要实现并行算法，必须解决数据分发、错误处理、负载均衡等一系列问题，相关代码不仅异常复杂且难以维护。

Google 的研究人员发现，大量并行计算应用可以由 Map 和 Reduce 两个过程组成的程序模型描述，只要开发人员完成 Map（映射）、Reduce（简化）两个步骤，其余的事情程序可以标准化处理。于是 Google 的工程师在 GFS 基础上实现了称为 MapReduce 的计算框架，封装了并行计算的复杂性，使普通的开发人员也能较容易地写出并行计算程序。MapReduce 就是一种编程模型，用于大规模数据集（大于 1TB）的并行运算，能够极大地方便不会分布式并行编程的编程人员,将自己的程序运行在分布式系统上。

MapReduce 程序模型和计算框架的用途在 Google 内部非常广泛。Google 利用 MapReduce 实现了分布排序、分布 Grep、Web 连接图反转、Web 访问日志分析、反向索引构建、文档聚类、机器学习、基于统计的机器翻译等众多程序和功能，MapReduce 还成为 Google 的索引更新方式。实践证明，MapReduce 框架的效率和功能都是不错的。

简单来说，MapReduce 改变以往把数据集中在一起的计算方式，而是把计算作为一项任务推送到存放的数据之上。MapReduce 框架运用了一个在计算机算法中最常用的概念——（Divide and Conque）分而治之，也就是把一个大的问题切割成

多个小问题，处理完成之后，再把答案汇总到一起。

- Map（映射）这一步骤所做的就是把在一个问题域中所有的数据在一个或多个节点中转化成 Key-Value（键-值）对，然后对这些 Key-Value 对采用 Map 操作，生成零个或多个新的 Key-Value 对。按 Key 值排序，同样的 Key 值被排到一起。然后合并生成一个新的 Key-Value 列表。
- Reduce（简化）步骤做的是收集工作，把 Map 步骤中生成的新的 Key-Value 列表，按照 Key 放到一个或多个子节点中，用编写的 Reduce 操作处理，归并后合成一个列表，得到最终的输出结果。

换成公式，MapReduce 可以这样表示：

```
Map(k1,v1) → list(k2,v2)
Reduce(k2, list (v2)) → list(v3)
```

输入的是（k1,v1）代表的 Key-Value 对列表，而输出的是 v3 列表。

图 3-7 是 MapReduce 的功能示意图。

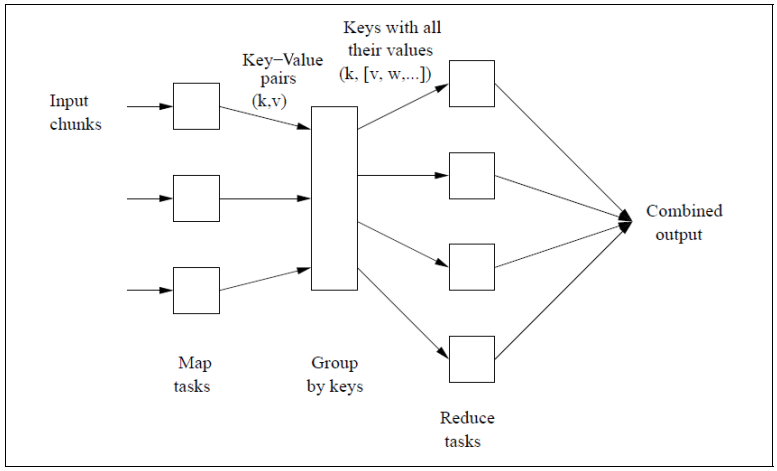



图 3-7　MapReduce 示意图

和 GFS 一样，在 Google 公司 MapReduce 也已经被新的技术所替代。Google 开发了名为 Caffeine（咖啡因）的类数据库系

统取代了 MapReduce 框架。关于 Caffeine 本书就不展开讲述了。

#### 3.5.1.3 BigTable 数据库

大数据对传统关系型数据库（Relational DataBase Management System，RDBMS）造成了很大冲击。关系型数据库的并发能力无法处理每秒上万次的读/写请求，而且关系型数据库在存储了上亿条记录时，查询效率通常会变得低下。最重要的一点是关系型数据库很难横向扩展，在数据量大时无法通过简单添加服务器的方法提高性能和负载能力，而且在对数据库进行升级扩展时又必须停机和迁移数据。

Google 设计和实现了一个名为 BigTable 的数据库，这是一个专门为管理大规模结构化数据而设计的分布式存储数据库。BigTable 放宽了数据事务要求，因为 Google 的大多数 Web 应用程序并不要求严格的数据库事务，对读一致性要求很低，一些场合甚至对写一致性要求也不高。很多 Web 应用程序设计时采用单表主键查询及简单条件分页查询，避免了多表关联查询，所以 BigTable 弱化了 SQL 功能，简化 SQL 的功能有利于存储和性能的提升。

目前 BigTable 已经部署上千台服务器，可靠地处理 PB 级数据，支撑包括谷歌地图（Google Earth）、谷歌分析（Google Analytics）、谷歌金融（Google Finance）等 60 多个应用。受 BigTable 启发，我们在下几节中会讲到的 Cassandra 和 HBase 等非关系型数据库延续着 BigTable 设计思路迅速发展起来。

BigTable 的表本质上是一个 Key-Value 对映射，由行键、列键、时间戳三维定位一条字符记录值。其中行是表的第一级索引，列是表的第二级索引，时间戳是第三级，每行拥有的列是不受限制的，可以每一行都不相同。如果多个列是属于同一族类的，我们可以用一个列族来表示这些列，表达方式为 Family：Qualifier，每个列族都有同样的 Family 值。我们可以用这样的公式表示 BigTable 表中的数据：

```
(row:string,column:string,time:int64)→string
```

图 3-8 是在 Google 论文中的一个示例，描述一张网页在 BigTable 中如何存储。www.cnn.com 是表的行键，以 URL 逆向排列然后按字母顺序存储，www.cnn.com 中最后一段“com”出现在最前面，而第一段“www”出现在最后，所以 www.cnn.com 对应的是“com.cn.www”。contents 是一个列键，对应的是网页内容，而 anchor 是一个列族，包含了 cnnsi.com 和 my.look.ca 两个反向链接的列键。Content 列下存放 t3、t5、t6 三个时间点的页面内容，cnnsi.com 和 my.look.ca 列下存放只有一份链接值，因为链接值并不因为时间而变化。

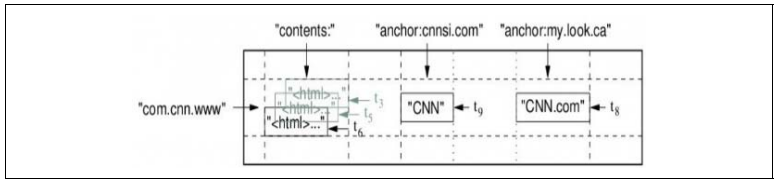



图 3-8　谷歌网页存储示意图

## 3.5.2　开源的分布式系统 Hadoop

本书主要讲述的是大数据挖掘，而近来在互联网上，“大数据”和“Hadoop”这两个词几乎划上等号了，但凡有人说到大数据的构架，就会提到 Hadoop。在本节中我们就来看下 Hadoop 究竟是什么，在大数据上能够起到什么样的作用。

### 3.5.2.1　Hadoop 的由来

和其他在互联网上的术语不同，Hadoop 这个词本身没有什么特殊含义，只是发明人 Doug Cutting 觉得很琅琅上口的一个象声词而已。

3.5.1 节中描述的分布式架构 GFS、MapReduce、BigTable 仅在 Google 内部使用，虽然 Google 无私地把 GFS、MapReduce、BigTable 都以论文的形式公布出来，只可惜这些架构中的代码并未开源。Google 在 2003 年底和 2004 年发表了两篇研究论文。第一篇介绍了 Google File System（谷歌文件系统）；另一篇介绍

的是 MapReduce 编程框架。顺着谷歌论文的思路，Doug Cutting 用 Java 语言“克隆”了一套开源分布式文件系统，取名为 Hadoop。

在 2006 年 1 月，Doug Cutting 加入雅虎公司，而雅虎公司为他提供了一个专门的团队和资源将 Hadoop 发展成一个可在网络上正式运行的完善系统。发展到现在，Hadoop 已经成为 Apache 基金会的一个顶级开源项目。Apache 基金会是一个以支持开源软件为目的的组织，它的官方网址是 http://www.apache.org/。

#### 3.5.2.2 Hadoop 的架构

Google 的分布式系统的两大特征如下：

- 使用大量廉价硬件设备构成集群，并有良好扩展性。
- 有高性能、高可靠性和海量存储的分布式架构。

构成 Hadoop 主要部分的 HDFS 文件系统和 MapReduce 引擎可以看成是 Google 的 GFS、MapReduce 在开源环境下的实现。以此为基础，Hadoop 同样也实现了和 Google 一样的分布式系统，也同样具有了 Google 系统的特征。

除了 Hadoop 主体上的这两个部分之外，在 Hadoop 周围还有各种配套的项目，如 HBase、Hive、Zookeeper、Pig 等。这些项目连同 Hadoop 本身一起构成了一个丰富的生态系统。请看图 3-9 的表示。

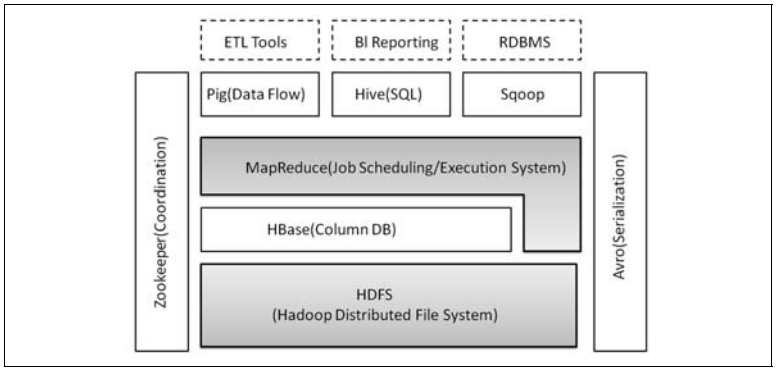



图 3-9 Hadoop 生态系统示意图

下面我们对出现在图 3-9 的 Hadoop 生态系统示意图中的各

个部分一一做简单解释。

- Avro 是为了解决 Hadoop 上 RPC 通信中出现的问题而研制的一个数据序列化系统，用于支持大批量数据交换应用。在 http://avro.apache.org/docs/current/spec.html 上可以找到 Avro 的详细文档。
- HBase 是 Hadoop 和数据库（DataBase）两个英文词合并在一起形成的，是一个在 HDFS 上搭建大规模结构化存储集群分布式存储系统，具有高可靠性、高性能和可伸缩特性。我们可以把 HBase 看成是 BigTable 在 Hadoop 中的实现。
- HDFS 是部署在廉价硬件上提供高吞吐量和高容错性的分布式文件系统，适合有超大数据集的应用程序，可以认为是相当于 Google 的 GFS。这是 Hadoop 的核心组成部分。
- Hive 是由 Facebook 首先研发的基于 Hadoop 的数据仓库工具，可以将结构化的数据映射成数据表并提供类 SQL 数据库查询管理功能，适合于数据仓库的统计分析。我们将在 3.5.2.4 节对 Hive 做展开描述。
- MapReduce 引擎是在 HDFS 上处理大数据集的并行计算框架，基本思想方法来自于 Google。在 Hadoop 上的 MapReduce 框架是由 Hadoop Common 程序包实现的。
- Pig 是在 HDFS 和 MapReduce 上处理大规模数据集的脚本语言，它提供更高层次的抽象并转化为优化处理的 MapReduce 运算。
- Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具。
- Zookeeper 是一个针对大型分布式系统的高可靠性的协调系统，提供的功能包括配置维护、名字服务、分布式同步、组服务等。

所有这些组合在一起的整个系统才是 Hadoop 的全部。图 3-9

中的虚线部分是不属于 Hadoop 系统的一部分、数据库，BI 工具和 ETL 工具可以通过接口与 Hadoop 系统相结合。

雅虎公司和 Doug Cutting 固然居功至伟，但是整个开源社区的努力才是 HadoopHadoop 能够在今天的互联网上发扬光大，被众多互联网公司接受的真正原因。

#### 3.5.2.3 Hadoop 的应用

Hadoop 主要有以下几个优点。而这些优点都是 Google 的分布式架构所拥有的，在下面的列表中，我们把出现的“Hadoop”字样全部转换成“Google 系统”，是完全适合的。

- 高可靠性。Hadoop 按位存储和处理数据的能力值得人们信赖。
- 高扩展性。Hadoop 是在可供使用的计算机集群间分配数据并完成计算任务的，这些集群可以方便地扩展到数以千计的节点中。同样，如果因为资源紧缺而收缩计算机集群或节点数，计算任务也可以随时重新调配。
- 高效性。Hadoop 能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。
- 高容错性。Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

对于 Hadoop 来说，它最大的优点是开放性。因为它是开源的，所以每天都有数以万计的程序员和爱好者在学习和研究 Hadoop 系统，对它做或大或小的更新和改良。

仅通过 6 年的发展，Hadoop 已经形成一套成熟的大数据存储和并行计算处理方案，受到越来越多公司的青睐。Hadoop 在搜索引擎、机器翻译、推荐系统、日志分析、图像视频分析、数据仓库、垃圾邮件识别等领域有广泛应用，在自然语言处理、生物信息学、太空图像处理和海量气候模拟案例上也不断取得成功。

国外有众多知名企业使用 Hadoop，如 Amazon 亚马逊、Facebook 脸谱网、IBM、Yahoo 雅虎、Adobe、AOL、New York

Times 纽约时报等。而在国内，互联网巨头百度和阿里巴巴等都有自己的大规模的 Hadoop 集群，是 Hadoop 的忠实使用者。阿里集团拥有数千台 PC 作为 Hadoop 的子节点，总存储量达到 PB 级别。阿里巴巴集团的 B2B、淘宝、天猫和阿里云等各个子公司都可以分享该集群中的数据。

以运营着世界上最大 CDMA 网络的中国联通来说，用户上网记录是海量数据，联通在一个中等城市公司日均上网记录达到 10 亿条，每月数据近 9TB。海量记录在传统的数据库上无法存储。而且每隔 6 个月中国联通用户整体上网流量会翻一番。所以联通只能保存一段时间内使用的流量总合，没有记录访问的目标 IP 地址明细，也就无法得知流量去向，用户因流量问题引发的投诉就很难处理。更不用说用户上网记录本身也是数据金矿，把用户上网记录存储起来可以发掘出有价值的东西。联通最终选择了采用 Hadoop 数据仓库解决方案，以普通的 PC 服务器部署系统，在数百个数据存储节点上，每个节点有 14TB 容量，在可预见的未来，全国 31 个省市的用户历史上网记录都可以很快随时检索出来。

Hadoop 的出现给很多小规模的 IT 技术公司提供了崭新的机会。Rapid Miner 是数据挖掘领域的一个常用开源工具，而有一群数据挖掘爱好者就成立了一家公司，名为 Radoop（Rapid Miner+Hadoop），在 Hadoop 集群上实现了 Rapid Miner 的功能。关于这家公司的产品，有兴趣的读者可以查阅网站 http://www.radoop.eu/。

#### 3.5.2.4　Hadoop 上的数据仓库工具 Hive

Hadoop 的 MapReduce 程序框架虽然简化了并行计算程序的复杂性，不过对大多数习惯在关系型数据库上搞开发的程序员和搞研究的数据分析师来说，直接在 Hadoop 上编程仍有不小的挑战。

世界上最大的社交网站 Facebook 中有许多人需要在大数据上做分析，这些人员的计算机知识层次不一，有工程师也有分析

员和产品经理。而分析内容也五花八门，有做特定研究的，也有使用商业智能程序的。而同时一系列 Facebook 的产品也建立在大数据分析之上。因此，构建一个能满足各种人员和程序需求的基础设施十分必要。为提高公司在 Hadoop 分布式系统上的大数据利用能力，Facebook 在 2007 年 1 月启动了 Hive 项目，于是，数据仓库工具 Hive 就诞生了 Facebook 的数据基础小组。

Hive 在维持 Hadoop 灵活性和扩展性的基础上提供类似数据库的基本功能，在 Hadoop 文件系统上提供了方便的数据查询和管理功能。在 2010 年 10 月，Hive 升级成为 Apache 开源基金会的一个顶级开源软件开发项目，而这个项目的目的就是为了完整构建基于 Hadoop 的数据仓库。Hive 在 Apache 开源基金会的首页是 http://hive.apache.org/。

我们来看一下 Hive 的架构和特性。Hive 有数据仓库上基础的 Table 表、Column 列、Partition 分区概念，支持 integer、float、double、string 等基础数据类型和 array、list、struct 等可以嵌套使用的复合类型。Hive 还允许用户使用自定义类型和函数，并提供丰富的接口和 API，数据可以通过 API 导入、导出。

不过这里最重要的一点是 Hive 本身并不存储数据，所有的数据都是来自于各数据源，是通过接口从 HDFS 和 HBase 上获取的。作为一个数据仓库工具，Hive 还提供 ETL 工具用于数据抽取、转换、装载。

Hive 是构建在 Hadoop 的 HDFS 文件系统和 MapReduce 引擎计算框架之上，由多个模块组成，其结构如图 3-10 所示。

图 3-10 中的 Metastore（超存储）组件用于储存系统目录和关于表、列、分区等信息的元数据，包含各种模式（Schema）和统计数据，有助于数据探索、查询优化、查询编译。Driver（驱动程序）组件包括编译器、优化器、执行器，把 HQL 语言编译成优化的 MapReduce 作业任务执行，同时管理执行整个生命周期、推护会话和统计信息。Hive 提供的包括 Command Line Interface（命令行界面）和 Web 界面能够方便操作管理数据库。

而 ThriftServer 组件提供 Thrift 接口和 JDBC/ODBC 服务，可以让 Hive 与其他应用集成。

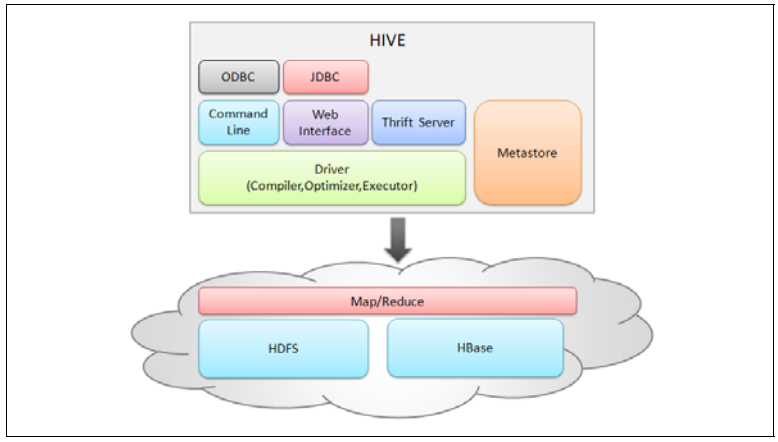



图 3-10　Hive 示意图

Hive 上提供了易于使用的类似 SQL 的 Hive QL 语言（简称 HQL 语言）。HQL 语言能编译成 MapReduce 作业在 Hadoop 上执行，也可以直接写 MapReduce 脚本。

比如我们可以用 HQL 语言创建一张数据库表格，如下所示：

```
CREATE TABLE IF NOT EXISTS table_name
(
--field 1
--field 2
--- …
---field N
)
PARTITIONED BY (string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '...';
```

由上面的这段语句我们可以发现在 Hive 上创建表格的工作和在传统数据库中几乎是一样的。同时，这个语言也允许熟悉 MapReduce 开发框架的开发者自定义 map 映射函数和 reduce 简化函数来处理复杂的分析工作。表 3-1 列出的是在 Hive 上可以使用的 HQL 语句。我们可以看到 HQL 语句的格式与 SQL 语

句基本一致。

表 3-1　HQL 可用命令示例列表

| 功　能 | 命 令 行 |
| --- | --- |
| 创建表 | ■ CREATE TABLE pokes (foo INT, bar STRING);<br>■ CREATE TABLE invites (foo INT, bar STRING) PARTITIONED BY (ds STRING); |
| 查看表 | ■ SHOW TABLES;<br>■ SHOW TABLES '.*s';<br>■ DESCRIBE invites; |
| 修改表 | ■ ALTER TABLE pokes ADD COLUMNS (new_col INT);<br>■ ALTER TABLE invites ADD COLUMNS (new_col2 INT COMMENT 'a comment');<br>■ ALTER TABLE events RENAME TO 3koobecaf; |
| 删除表 | ■ DROP TABLE pokes; |
| 数据载入 | ■ LOAD DATA LOCAL INPATH './examples/files/kv1.txt' OVERWRITE INTO TABLE pokes;<br>■ LOAD DATA LOCAL INPATH './examples/files/kv2.txt' OVERWRITE INTO TABLE invites PARTITION (ds='2008-08-15'); |
| 查询、插入 | ■ SELECT a.foo FROM invites a WHERE a.ds='2010-08-15';<br>■ INSERT OVERWRITE DIRECTORY '/tmp/hdfs_out' SELECT a.* FROM invites a WHERE a.ds='2010-08-15'; |
| 联合、分组 | ■ FROM pokes t1 JOIN invites t2 ON (t1.bar = t2.bar) INSERT OVERWRITE TABLE events SELECT t1.bar, t1.foo, t2.foo;?<br>■ FROM invites a INSERT OVERWRITE TABLE events SELECT a.bar, count(*) WHERE a.foo > 0 GROUP BY a.bar; |

Hive 在 Facebook 中得到广泛使用。目前 Facebook 的 Hadoop/Hive 集群每天运行的作业包括从简单的汇总统计到复杂的 BI 应用、机器学习，数据量在 PB 级别。而在 Facebook 之外，由于它是一个备受追捧的开源项目，也拥有许多拥蠹者。

### 3.5.3　Facebook 的数据仓库

Facebook 是世界上最大的社交网站，到 2010 年就有超过 4 亿用户（2012 年超过 9 亿用户），每月分享 250 亿的内容信息，以及 5000 亿的单月页面浏览量。我们来看 Facebook 上的数据仓库是怎样的。从初创到 2008 年，Facebook 主要使用的还是关系型数据库，但是由于数据增长很快，每天处理数据的量级很快从 GB 上升到 TB 级别，而一些日常数据分析处理程序甚至无法在当天完成。Facebook 不得不重新考虑更新它的数据存储基础架构。Facebook 最终选择了 Hadoop 集群作为数据基础平台，通过堆积廉价服务器扩展数据仓库的规模，Hadoop 终于能够处理 PB 级数据。

在 2010 年 7 月，Facebook 部署的 Hadoop 集群存储了 36PB 未压缩的数据，有超过 2250 台机器和 23000 个核心，每台机器 32GB 内存，日处理 80～90TB 数据。集群平均每天处理 300 多个用户提交的 25000 多个任务。

Facebook 使用开源的 Scribe 工具从 Web 集群和系统记录载入数据日志，数据包括好友信息、新闻推送以及广告推广信息等。每天的数据日志增量都在 TB 级别。数据载入一个生产环境计算机集群，只运行仔细监控的关键性任务。然后 Hive 将数据推送到其他计算机集群，运行关键级别稍低一些的任务。Facebook 的每个计算机集群都是由一系列节点和一个单一核心的交换机组成的，而将数据分割到不同的集群保证了关键性任务的高可靠性。如图 3-11 所示。

Facebook 的数据仓库集群 95%的任务是由 Hive 编写的。Facebook 创建了一个基于 Web 的工具提供给业务分析师来使用 Hive，只需要简单的撰写查询语句就可以查询载入数据仓库的所有数据表格，使原先的每天批处理查询过渡到现在的实时查询。一些简单的任务通常十分钟就可以写成。

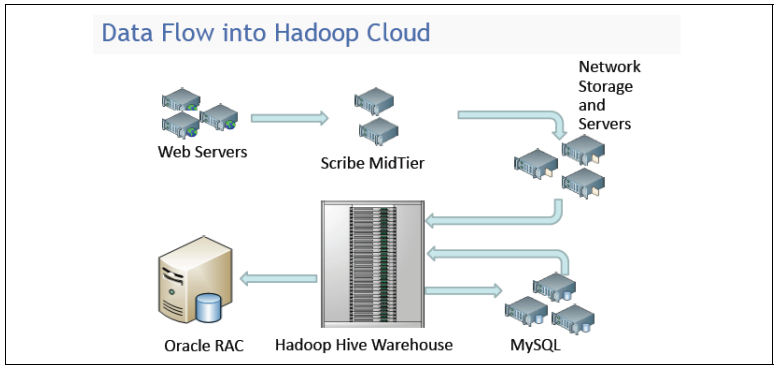



图 3-11　Facebook 数据仓库示意图

### 3.5.4　NoSQL

为了配合 Hadoop 技术，软件开发商们也研发出了各种各样基于数据库的新技术，其中很多也都是出自开源社区，而 NoSQL 是最抢眼的一个技术之一。据悉 NoSQL 一词最早出现于 1998 年，是 Carlo Strozzi 开发的一个轻量、开源，不提供 SQL 功能的数据库。NoSQL 提供的方法相对于 SQL 的数据库来说有巨大的优势。因为它允许应用程序扩展到新的水平。新的数据服务基于真正可扩展的结构和体系构建云、构建分布式。这对于应用开发来说是非常有吸引力的。无须 DBA，无须复杂的 SQL 查询。


NoSQL，指的是非关系型的数据库。随着互联网社会化媒体的兴起，传统的关系数据库在应付超大规模和高并发的 SNS 类型的 Web 2.0 纯动态网站上已经显得力不从心，暴露了很多难以克服的问题，例如以下列出的对于数据库的三高（3 High）要求：

- High Performance——对数据库的高并发读/写的需求；
- Huge Storage——对海量数据的高效率存储和访问的需求；

- High Scalability && High Availability——对数据库的高可扩展性和高可用性的需求。

在上面提到的“三高（3 High）”需求面前，关系数据库遇到了难以克服的障碍。同时，数据库要求我们强行修改对象数据，以满足 RDBMS（Relational Database Management System，关系型数据库管理系统）的需要。而对于 Web 2.0 网站来说，关系数据库的很多主要特性却往往无用武之地，例如：

- 数据库事务的一致性需求；
- 数据库的写实时性和读实时性需求；
- 对复杂的 SQL 查询，特别是多表关联查询的需求。

分布式系统中，有三种重要的属性，详见如下所述。

- 一致性（Consistency）：数据一致性，任何一个读操作总是能读取到之前完成的写操作结果，也就是在分布式环境中，多点的数据是一致的。
- 可用性（Availability）：好的响应性能，每一个操作总是能够在确定的时间内返回，也就是系统随时都是可用的。
- 分区容忍性（Tolerance of network Partition）：可靠性，在出现网络分区（比如断网）的情况下，分离的系统也能正常运行。

美国著名科学家，Berkerly 大学的 Brewer 教授提出的 CAP 原理解释了关于这三种属性的关系。所谓 CAP 原理的意思是，一个分布式系统不能同时满足一致性，可用性和分区容错性这三个需求，最多只能同时满足两个。后来麻省理工学院的两位科学家在理论上证明了 CAP 原理的正确性。如图 3-12 所示为 CAP 原理示意图。

CAP 原理指出一致性、可用性、分区容忍性不可三者兼顾。因此在进行分布式架构设计时，必须做出取舍。而对于分布式数据系统，分区容忍性是基本要求，否则就失去了价值。因此设计分布式数据系统，就是在一致性和可用性之间取一个平衡。对于大多数 Web 应用来说，其实并不需要强一致性，因此牺牲一致

性来换取高可用性，是目前多数分布式数据库产品的方向。

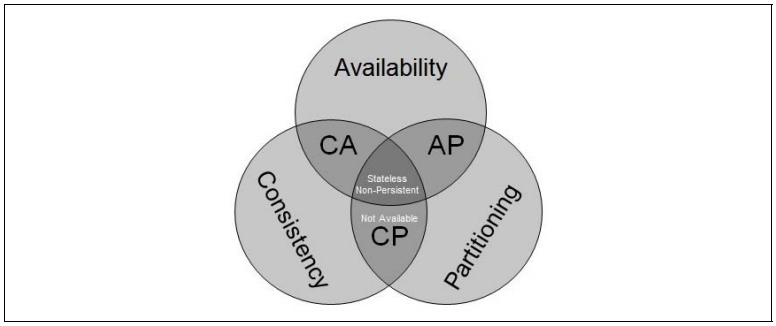



图 3-12　CAP 原理示意图

当然，牺牲分布数据库的一致性，并不是完全不管数据的一致性，否则数据会是混乱的，那么系统可用性再高，分布式再好也没有了价值。所谓的牺牲一致性，只是不再要求关系型数据库中的一贯强一致性，而是只要系统能满足最终一致性即可。关系型数据库要求每次更新过程中的数据都能被后续的访问看到，这是强一致性。如果能容忍更新过程中或者更新过程之后部分或者全部的数据访问不到，则是弱一致性。如果在数据更新过程中或者更新之后的在一段时间内可能访问不到，而经过一段时间后能够访问到全部更新后的数据，则是最终一致性。在此我们所说的牺牲数据一致性，都不是指数据的弱一致性，而是数据的最终一致性，也就是说分布式的数据在更新了的一段时间之后会达到完全一致。

考虑到客户体验，这个最终一致的时间窗口，要尽可能的对用户透明，也就是需要保障“用户感知到的一致性”。通常是通过数据的多份异步复制来实现系统的高可用和数据的最终一致性的，“用户感知到的一致性”的时间窗口则取决于数据复制到一致状态的时间。

目前 Google 的 BigTable 和 Amazon 的 Dynamo 都采用 NoSQL 型数据库，Facebook 的 HBase 也是一种 NoSQL 型数据库。

NoSQL 中最好的代表应当数 Cassandra。Cassandra 是一个混合型的非关系的数据库，类似于 Google 的 BigTable。Cassandra 最初由 Facebook 开发，后转变成了开源项目。它是一个网络社交云计算方面理想的数据库。以 Amazon 专有的完全分布式的 Dynamo 为基础，结合了 Google BigTable 基于列族（Column Family）的数据模型。P2P 去中心化的存储。很多方面都可以称之为 Dynamo 2.0。

另外值得一提的是介于关系数据库和非关系数据库之间的 MongoDB。MongoDB 是一个基于分布式文件存储的数据库。由 C++语言编写。旨在为 Web 应用提供可扩展的高性能数据存储解决方案。MongoDB 是非关系数据库当中功能最丰富，最像关系数据库的。它支持的数据结构非常松散，是类似 json 的 bjson 格式，因此可以存储比较复杂的数据类型。Mongo 最大的特点是它支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。国外的 Craigslist、Foursquare，国内的淘宝网等知名互联网公司都在他们的生产环境部署了 MongoDB。

## 3.6 本章相关资源

- 本章相关参考文献：

[1] 夏火松. 数据仓库与数据挖掘技术[M]. 科学出版社, 2004.

[2]（美）Richard J.R，（美）Michael W.G. 数据挖掘教程[M]. 翁敬农, 译. 清华大学出版社, 2003.

[3] 段云峰等. 数据仓库及其在电信领域中的应用[M]. 电子工业出版社, 2003.

[4]（美）Inmon,W.H 数据仓库（第 4 版）. 王志海等, 译. 机

械工业出版社, 2006 年.

[5] 陈京民. 数据仓库原理、设计与应用[M]. 中国水利水电出版社, 2004.

[6]（美）Lou Agosta. 数据仓库技术指南. 潇湘工作室, 译, 人民邮电出版社, 2000.

[7] Brin, S. and Page, L. *The an atomy of a large-scale hypertextual W eb search eng ine*. In Proceedings of the Seventh international Conference on World Wide Web (WWW-7) (Brisbane, Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1998: 107-117.

[8] L. Page, S. Brin, R. Motwani, T. Winograd. *The Pagerank Citation R anking*: *Brin ging Or der to th e W eb, T echnical Report*. Stanford University, 1999.

[9] Nancy Lynch and Seth Gilbert. *Brewer's conjecture and the feasibility of cons istent, av ailable, par tition-tolerant we b services*. ACM SIGACT News, 2002, 33(2): 51-59.

- 本章相关网址：

[1] http://www.apache.org/

[2] http://www.csdn.net/

[3] http://www.hadoopor.com/

[4] http://dataminingtools.net/blog

[5] http://olap.com/w/index.php/OLAPEducationWiki

[6] http://www.radoop.eu/

[7] http://en.wikipedia.org/wiki/GoogleFileSystem

[8] http://avro.apache.org/docs/current/spec.html

[9] http://hive.apache.org/

# 第 4 章
# 数据挖掘算法及原理

在本章中我们简单介绍数据挖掘的基本算法及其原理，而其中一些常用算法的实际应用会留在本书的最后几章中做评述。因为本书的目的主要是为了讲述大数据挖掘上的应用而不是数据挖掘理论研究，所以算法的起源、原理和具体过程不是我们的重点。

我们在本书中力求非计算机专业的读者能够了解到算法的概念，如何通过这些算法的应用来做好数据挖掘，晦涩高深不是我们的出发点。

## 4.1 数据挖掘中的算法

在实际应用中，面临大规模数据，我们的目标是如何用一个或几个简单而有效的算法或算法的组合来提取有价值的信息，并不是追求算法的复杂和完美。事实上，很多复杂的算法只能停留在理论层面上，面对海量数据时是不现实的。

数据挖掘算法在本书的实例中应用的比较多的是聚类算法、关联算法和分类算法。比如我们在第 8 章中会用 RFM 模型聚类算法来做电子邮件营销相关的数据挖掘。第 9 章中用决策树分类算法来挖掘互联网广告中的作弊行为；第 10 章会用聚类算法和关联算法来做好电子商务网站的客户管理。

## 4.2 数据挖掘十大经典算法

国际权威的学术组织，数据挖掘国际会议 ICDM（the IEEE International Conference on Data Mining）在 2006 年 12 月评选出了数据挖掘领域的十大经典算法。这十大算法分别是 C4.5、K-means、SVM、Apriori、EM、PageRank、AdaBoost、KNN、Naive Bayes 和 CART 算法。

参选的 18 个算法分别来自数据挖掘的不同领域，都有相当的应用历史。下面大致介绍一下参选算法的特点。

- C4.5、CART、KNN、Naïve Bayes 是分类数据挖掘的代表算法。分类算法在数据挖掘中是比较常用的，所以候选算法和最后入围者也是最多的。这四个算法无一例外，全部都入选了最终的十大经典算法。
- K-means、BIRCH 是聚类的代表算法。我们在 4.4 节中会详细介绍聚类算法。
- Apriori、FP-Growth 是关联分析的代表算法。我们在 4.5 节中会详细介绍关联分析的算法原理和概念。
- PageRank、HITS 是链接挖掘的代表算法。链接挖掘是在互联网诞生之后衍生出的一类数据挖掘算法，主要用于处理和分析互联网上的超链接和相关网页信息。
- AdaBoost 是装袋和增强的代表算法。增强（Boosting）和装袋（Bagging），又称 Bootstrap aggregating，都是在人工智能机器学习上的宏算法（Meta-algorithm），也就是说通常装袋和增强算法都需要和另外一个算法合在一起使用。这里的 AdaBoost 就是一个增强的分类器算法。
- GSP、PrefixSpan 是序列模式挖掘的代表算法。序列算法是从一个序列（Sequence）中的数据找出统计规律。在 4.6 节中会对序列挖掘原理和概念做详解。
- SVM、EM 是统计学习的机器学习算法。人工智能机器学

习在数据挖掘中起到很重要的作用，因为很多时候数据内部隐含的意思是我们无法用常理来概括的。

- CBA 是聚合挖掘的代表算法。而 CBA 算法是将关联规则挖掘与分类技术相结合的一种分类算法，在许多领域中都有广泛应用。
- Finding reduct 是粗糙集的算法。粗糙集，从字面上的意思来看就是说数据本身是粗糙的，不完善的。粗糙集算法能从大量不完全的信息中找出符合现有数据的规则。
- gSpan 是图挖掘的代表算法。这里的图（Graph）指的是通过数据对象构建的数学模型，比如图 4-1 就是一个“图 Graph”的表示。从图中找寻频繁出现的一个子图是最近许多数据挖掘学者研究的方向，而 gSpan 是其中做得比较出色的一个算法。

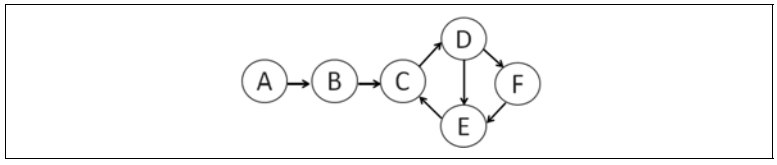



图 4-1 “图 Graph”示意图

这个评选是一个艰难的过程，因为列出的每一个算法在数据挖掘领域都有深远的影响，同时又各自代表了数据挖掘的一个子领域。根据应用的广泛性和引用度等，最后选出的 10 个算法列举如下。

（1）C4.5

C4.5 算法是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。C4.5 算法产生的分类规则易于理解，准确率较高。不过在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，在实际应用中会因此而导致算法的低效。

（2）The K-means algorithm 即 K-means 算法

K-means 算法是一个聚类算法，把所有的数量为 n 的对象，根据它们的属性分为 k 个分割，其中 k < n。K-means 算法假设对象属性来自于空间向量，试图找到数据中各个自然聚类的中

心，而算法的目标是使各个群组内部的均方误差总和达到最小。

（3）SVM（Support Vector Machines）支持向量机

英文为 Support Vector Machine，简称 SV 机（理论书籍中会简称其为 SVM）。它是一种监督式学习的方法，目前广泛的应用于统计分类以及回归分析中。SVM 算法的目的是找到一个最优超平面，使得分类间隔最大。最优超平面就是要求分类面不但能将两类正确分开，而且使分类间隔最大。SVM 算法的理论基础是平行超平面间的距离或差距越大，分类器的总误差会越小。

（4）The Apriori Algorithm

Apriori 算法可以算是最有影响的挖掘布尔关联规则频繁项集的算法。其核心思想很直观，是基于自下而上的递推算法，从一个项目子集扩展到两个项目子集，从两个扩展到三个，直到无法扩展为止。在 Apriori 算法中，频集的概念被第一次提出：所有支持度大于最小支持度的项集称为频繁项集，简称频集。在 4.4 节中我们会举例描述 Apriori 算法的实施，而在关联规则数据挖掘中的很多算法都是从 Apriori 算法衍生出来的。

（5）最大期望（EM）算法

和刚提过的 K-means 算法一样，EM 算法也是一个聚类算法，而该算法的目的是试图找到数据中存在的各个自然聚类的中心。依赖统计学原理，我们聚类模型是一个概率模型，而该模型依赖于无法观测的隐藏变量（Latent Variable）。找出隐藏的变量，我们便能成功构建这个概率模型。而最大期望（EM，Expectation-Maximization）算法就是在概率模型中寻找可能性最大的参数值的算法。在最大期望算法中采用的参数估计方法称为参数最大似然估计（Maximum Likelihood Estimation）。最大似然估计这一术语来自统计学，是用来求一个样本集的相关概率函数的参数的一种统计方法。

（6）PageRank

从商业价值来说，PageRank 算法是最值钱的一个算法，因为 PageRank 算法是 Google 搜索算法的核心内容。PageRank 算法在 2001 年 9 月被授予美国专利，专利人是 Google 创始人之一

拉里佩奇（Larry Page），所以 PageRank 里的 Page（英文是书页的意思）不是指网页，而是指佩奇，因为这个算法是以佩奇来命名的，PageRank 算法可以直译为佩奇分级算法。

PageRank 算法的核心思想是根据网站的外部链接和内部链接的数量和质量两个因素来衡量网站的价值。PageRank 算法背后的概念是，每个到页面的链接都是对该页面的一次投票，被链接的次数越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”，衡量多少人愿意将他们的网站和你的网站挂钩。就像是论文，被别人引述的次数越多，一般就会判断这篇论文的权威性越高。

（7）AdaBoost

AdaBoost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的分类器（强分类器）。AdaBoost 算法根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，如此迭代运营，最后将每次训练得到的分类器融合起来，作为最后的决策分类器。

（8）KNN，K 最近邻分类算法

KNN，英文为 k-nearest neighbor classification，意为 K 最近邻分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的基本思想是：如果一个样本在特征空间中的 K 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。在 2.2.2.1 节中我们已经描述过该算法的细节，此处不再赘述。

（9）Naive Bayes 朴素贝叶斯

在众多的分类模型中，应用最为广泛的两种分类模型是决策树模型（Decision Tree Model）和朴素贝叶斯模型（Naive Bayesian Model，NBC）。朴素贝叶斯模型，简称 NBC 模型，发源于古典数学理论，有着坚实的数学基础和稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简

单。理论上来说，NBC 模型与其他分类算法相比具有最小的误差率。但是实际情况中并非如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在正式的应用中往往是不成立的，这给 NBC 模型的正确分类带来了一定影响。在属性个数比较多或者属性之间相关性较大时，NBC 模型的分类效率比不上决策树模型。而在属性相关性较小时，NBC 模型的性能最为良好。我们在第 8 章描述垃圾邮件分类时会用到朴素贝叶斯算法。

（10）CART，分类与回归树

CART 是 Classification and Regression Trees 的英文首字母缩写，也称为分类与回归树。在分类树下面有两个关键的思想。第一个是关于递归地划分自变量空间的想法；第二个是用验证数据进行剪枝。在 CART 算法中引入了 GINI 指标（Gini Index）方法来衡量分类的纯粹性。

数据挖掘这十大经典算法的评选确实是相当精准，因为如果今天数据挖掘学者们再一次做投票，这十大算法可能依然会是 2012 年的十大经典算法。在 Web 挖掘上，我们在实际案例中经常使用的算法有 K-means 算法、K 最近邻算法、SVM、Apriori 及其衍生算法等。

## 4.3 分类算法（Classification）

在第 2 章中以分类算法为例，我们已正开始为读者们讲述数据挖掘，因为分类是数据挖掘最常用的应用。分类算法反映的是如何找出同类事物共同性质的特征型知识和不同事物之间的差异型特征知识。

最为典型的分类算法是基于决策树的分类算法。决策树其叶节点是类别名称，中间节点是带有分枝的属性，每个分枝对应该属性的某一可能值。我们从实例集中构造决策树，而这棵树的构成过程有多种方式。我们来看一种通过机器学习的有指导的学习方法构建决策树的过程。有指导的学习方法（Supervised Learning）

指的是我们对于每个实例可以很清晰地知道分类结果是否正确，从而对于机器学习的过程进行修改，也就是指导。如 2-2 节中讲到的“高富帅”，当机器学习把训练集中的某个人分成“高富帅”或者“屌丝”，我们可以很清楚地判定分类结果是否正确，从而对于决策树构建过程进行修正。下面我们来看下这个过程：

（1）我们先根据训练子集形成一个初始的决策树。

（2）如果该树不能对所有对象给出正确的分类，那么选择一些例外加入到训练子集中。

（3）重复该过程一直到形成正确的决策集。

最终结果是一棵新生成的决策树。决策树是一种常用于预测模型的算法，它通过将大量数据有目的的分类，从中找到一些有价值的、潜在的信息。决策树的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。最有影响和最早的决策树方法是由 Quinlan 提出的著名的基于信息熵的 ID3 算法。

这里我们需要简单解释一下来自信息论的信息增益的概念。信息增益（Information Gain）是衡量一个属性区分数据样本的能力。信息增益量越大，对信息分类的能力就越强。比如说在电子商务客户分类中，从客户的多个属性中我们选择客户的总成交量作为属性来分类要比选择客户的地区属性分类要有效的多。而用来计算信息增益的公式就需要用到熵（Entropy）。

针对 ID3 算法中的问题，出现了许多较好的改进算法，如 Schlimmer 和 Fisher 设计了 ID4 递增式学习算法；钟鸣、陈文伟等提出了 IBLE 算法等。而 4.2 节提到的 C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了一些改进，使得其实用性更好：用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；在树构造过程中进行剪枝（Pruning）；能够完成对连续属性的离散化处理；能够对不完整数据进行处理。

模拟人脑思维方式的神经网络是个很有趣的分类算法，对于复杂度比较大的分类问题提供了一个相对简单的解决方案。

图 4-2 是一个神经网络的典型示意图，其中①和②代表输

入，可以有任意多个变量的输入，这些变量也称为预测变量。⑥表示输出，通常是一个向量值，但是也可能有多个变量组成。而③、④、⑤表示隐含层，神经网络的隐含层可以有任意多层，每层也可能有多个变量，而隐含层的层数和每层的变量个数决定了神经网络的拓扑结构或复杂度。每个节点的数值都是由它上一层的全部父节点的数值导出的。假设这里的关系全部是线性的，那么整个神经网络就变成一个线性回归函数，如果我们用 $V_X$ 表示节点 X 上的数值，用 $Weight_{XY}$ 表示节点 X 和 Y 之间的权重，会有以下的计算方式：

$$V_2 = V_1 \times Weight_{13} + V_2 \times Weight_{23}$$
$$V_4 = V_1 \times Weight_{14} + V_2 \times Weight_{24}$$

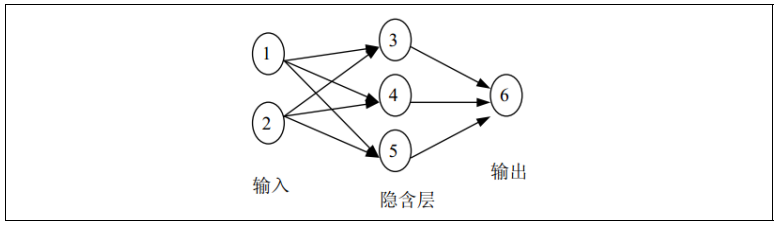



图 4-2　神经网络示意图

设计好神经网络的拓扑架构之后，神经网络的训练就是调整节点间连接的权重。有各种不同的训练方法可以调整节点间的权重。如果我们采用的是一个成熟的数据挖掘软件或工具，神经网络会自动根据训练集来调整神经网络上的各项权重数值。使用神经网络算法最需要注意的问题就是训练适当，如果学习过度，神经网络的效果反而会不好，因为输入时包含了所有的参数，神经网络可能会“记住”训练集中所有不必要的细节。

覆盖正例排斥反例方法也是一种分类的方法。顾名思义，该方法是利用覆盖所有正例、排斥所有反例的思想来寻找规则。首先在正例集合中任选一个种子，到反例集合中逐个做比较，按此思想循环所有正例种子，将得到选取正例的规则。比较典型的算法有 AQ11 方法和 AE5 方法等。

数据分类还有以线性回归和线性辨别分析为典型代表的统

计方法、粗糙集（Rough Set）方法等。

## 4.4 聚类算法（Clustering）

聚类也称分段，是指将具有相同特征的人归结为一组，将特征平均，以形成一个“特征矢量”。而我们通常说的“物以类聚、人以群分”拿来形容聚类是很合适的。聚类（Clustering）是完全可以按字面意思来理解的——将有共同特征的对象实例聚成一类的过程。聚类分析又称群分析，是研究（样品或指标）分类问题的一种统计分析方法。

聚类算法通常用来确定对象一共分多少聚类，并设法从每一聚类中找出最能代表这一聚类的数据特征。聚类算法经常被一些数据挖掘者用来提供不同类对象特征的报告。

简单来说，如果一个数据集合包含 $n$ 个实例，根据某种准则可以将这 $n$ 个实例划分为 m 个类别，每个类别中的实例都是相关的，而不同类别之间是区别的也就是不相关的，这个过程就称为聚类。

我们来看一个最简单的聚类方法，把有 $n$ 个样本的集合分成若干个类别 $\{C_1, C_2, C_3\}$，而这些类别中的样本和每个类别样本中心的距离都不大于阈值 $T$。算法如下：

对于 $n$ 个样本的集合，$Z_s = \{Z_1, Z_2, \cdots, Z_n\}$

给定一个阈值 $T$。

（1）任取一个样本，例如 $Z_1$，把 $Z_1$ 作为第一个类的中心 $C_1$。

（2）然后从样本集中依次取 $Z_i$ $(i = 2,3,\cdots,n)$，计算 $Z_1$ 与 $Z_i$ 的距离 $D_i$。

若 $D_{1i} \leqslant T$，则判定 $Z_i$ 属于 $Z_1$ 为中心的那个类；

若 $D_{1i} > T$，则把 $Z_i$ 作为新的类中心 $C_2$。

（3）然后再对剩下的样本取一个 $Z_i$ 分别计算与 $Z_1$，$Z_2$ 的距离 $D_{1i}$，$D_{2i}$。

若其中较小者≤$T$，则判定$Z_i$属于较小的那一类；

否则，就把$Z_i$作为新的一个类的中心$C_3$。

如此继续循环，直至对全体样本做完处理。完成之后，我们就得到了几个类别$\{C_1, C_2, C_3\}$，而这些类别中的样本和这些样本中心的距离都不大于阈值$T$。如果数据的类别内部间距相对于类别外部间距要小得多（不在一个数量级），那么上面这个简单算法能够快速完成聚类。

在聚类算法中最著名的算法当属 K-means 算法。在 4.2 节中讲述数据挖掘十大算法时我们对 K-means 算法已经做了描述，在此就不赘述了。

聚类 Clustering 与分类 Classification 不同的是，在我们做聚类分析前并不知道会以何种方式或根据来分类。所以在聚类算法完成之后，数据和对象分成若干个群，我们必须要配合专业领域知识来解读这些数据分群的意义。

对于聚类中我们到底应该分出多少类别，这里没有一个最优的方法来判别。我们只能做个大概的估算。估算的类别数字大概是$\sqrt[2]{\frac{n}{2}}$。也就是说如果样本的数量是 100，类别大概是 7；样本的数量是 10000，类别大概是 70。

聚类算法里用来衡量各个数据点之间间隔的一个概念是相异度，或者直观来说就是距离。用数学方式来表达，距离是这样定义的：

设$X=\{x_1, x_2, x_3, ..., x_n\}, Y=\{y_1, y_2, y_3, ..., y_n\}$，其中$X, Y$是两个数据点，各自具有$n$个可度量特征属性，那么$XY$的相异度定义为：$d(X,Y)=f(X,Y)\rightarrow R$，其中$R$为实数域。

也就是说对于任何两个数据点，我们会用一种计算方式$f(X,Y)$把相异值转化成一个实数，对这两个数据点之间作比较。数字越小则越相近。

衡量距离方式可以用欧几里得方式（Euclidean Distance）或是曼哈顿方式（Manhattan Distance）等。在两维平面空间里，欧几里得方式是两点之间的直接距离，而曼哈顿方式是水平方向的

距离加上垂直方向的距离，另一种说法是的士司机方式（Taxicab Geometry），这个说法的原因是纽约曼哈顿街区的街道都是横平竖直的，不可能走斜线。

曼哈顿方式最有趣的事情是从一点到另外一点，只要是对着目标，走水平或垂直两个方向，那么最后的距离是一样的。在图 4-3 中只有对角的这条线衡量的是欧几里得方式下的距离，而其他各种颜色的路线都是曼哈顿方式的。在两维的情况下还不太明显，但是在三维甚至更多维度下，我们采用曼哈顿方式来衡量数据点之间的距离优势就很明显了，因为我们可以采集各个维度上的数值差，它们的绝对值累加之后就是曼哈顿距离。

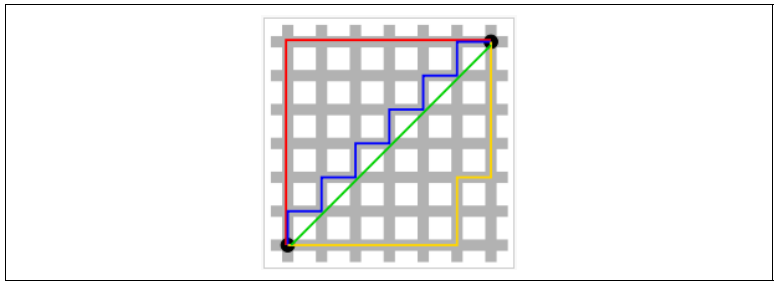


图 4-3　曼哈顿方式示意图

再举一个例子，假设我们所有的数据是在互联网上做的抽样调查，其中的每一个问题都是以“是”或者“否”来做答案的。比如我们有如下问题。

问题 1：你是否在网上玩过网游？

问题 2：你是否注册过淘宝账号？

问题 3：……

在每个用户回答了所有的 10 个问题之后，我们对于这个用户就产生了一个 10-维的数据点，而每个数据点都是以 0 或者 1 来表示的，0 表示用户回答的是“否”，而 1 则表示用户回答的是“是”。

如第一个用户，他对于第 1、2、3、4 个问题的回答都是“是”，而对其他的 6 个问题的答案都是“否”，那么我们可以用 $X=\{1,1,1,1,0,0,0,0,0,0\}$ 来表示他的数据点。而第二个用户，他

对于第 2、4、5 个问题的回答都是“是”，而对其他的 7 个问题的答案都是“否”，那么我们可以用 $Y=\{0,1,0,1,1,0,0,0,0,0\}$ 来表示他的数据点。

如果我们假设这次互联网抽样调查中每个选择的权重都是一样的，而且相互独立，那么我们可以数出不一致的维度，这里答案不相同的是第 1、3、5 个问题，所以这两个用户数据点之间的距离是 $3/10=0.3$。

在大规模数据集中，通常存在着不遵循数据模型的普遍行为的样本。这些样本和其他残余部分数据有很大不同或不一致，叫做异常点（Outlier）。有时候，我们做聚类算法分析的主要原因就是为了找到这些异常点。这些异常点不同于在数据预处理时我们会提到的异常数据。后者的出现是因为数据错误造成的，而异常点的存在有它的特殊原因。例如在电子商务中我们可能会发现有一个客户相关数据游离于所有其他客户之外。经过这个客户相关数据的具体挖掘，我们可能从中发现重要的商机。

还有很多种聚类算法，比如基于 K-means 算法的 K-medoids 和 PAM 算法等。K-medoids 算法和 K-means 算法基本上是一致的，不同之处仅在于每个聚类中心点的选取。在 K-means 算法中，我们将中心点取为当前 cluster 中所有数据点的平均值，而在 K-medoids 算法中，我们将从当前的聚类中选取这样的一个点：到聚类中其他所有点的距离之和为最小的来作为中心点。K-medoids 算法计算出的聚类不容易受到偶然出现的异常数据影响，但是因为每个点的加入都会影响到聚类的中心点位置，显然计算工作量会比较大。因为我们面对的数据量级都比较大，K-medoids 算法的优点比不上它对于运算效率产生的负面影响，所以一般不会考虑。

## 4.5 关联算法

关联规则算法在市场交叉销售（Cross Selling）、向上销售

（Up Selling）、商场布置、产品定价、促销安排、医疗诊断、基因科学研究等领域都有大量的实际应用，而我们在本书中关注的是关联算法在互联网各个领域的实际应用。这一章的任务是让读者能够基本了解关联算法的基本概念和算法原理。

### 4.5.1　关联算法中的概念

所谓关联，反映的是一个事件和其他事件之间依赖或关联的知识。这里有两个英文词是一致的。第一个是相关性 relevance，第二个是关联性 association，两者都可以用来描述事件之间的关联程度。其中前者主要用在互联网的内容和文档上，比如搜索引擎算法中文档之间的关联性，我们采用的词是 relevance。而后者往往用在实际的事物之上，比如我们在电子商务网站上的商品之间的关联度是用 association 来表示的，而关联规则是用 association rule 来表示的。

如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。简单来说，关联规则可以用这样的方式来表示：A==>B，其中 A 被称为前提或者左部（LHS），而 B 被称为结果或者右部（RHS）。如果我们要描述关于尿不湿和啤酒的关联规则（买尿不湿的人也会买啤酒），那么我们可以这样表示：

买尿不湿==>买啤酒

在关联算法中很重要的一个概念是支持度（Support），也就是在数据集中包含某几个特定项的概率。比如在 1000 次的商品交易中同时出现了啤酒和尿不湿的次数是 50，那么此关联的支持度为 5%。

和关联算法很相关的另一个概念是置信度（Confidence），也就是在数据集中已经出现 A 时，B 发生的概率，置信度的计算公式是（A 与 B 同时出现的概率）/（A 出现的概率）。

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。

关联可分为简单关联、时序关联、因果关联等。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有置信度。

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。它在数据挖掘中是一个重要的课题，最近几年已被业界所广泛研究。关联规则挖掘的一个典型例子是购物篮分析。关联规则研究有助于发现交易数据库中不同商品（项）之间的联系，找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。分析结果可以应用于商品货架布局、货存安排以及根据购买模式对用户进行分类。

关联规则的发现过程可分为两步。第一步是迭代识别所有的频繁项目集（Frequent Itemsets），要求频繁项目集的支持度不低于用户设定的最低值；第二步是从频繁项目集中构造置信度不低于用户设定的最低值的规则，产生关联规则（Association Rules）。识别或发现所有频繁项目集是关联规则发现算法的核心，也是计算量最大的部分。

关联规则的概念最早在 1993 年由 R.Agrawal 等人首次提出，而最为著名的关联规则发现方法也是 R.Agrawal 提出的 Apriori 算法。Apriori 算法中的关联规则在分类上属于单维、单层、布尔关联规则。此外比较著名的还有 Eclat 和 FP-Growth 算法。其他的许多关联算法是这几种算法的衍生和改进算法。

前面提到的支持度（Support）和置信度（Confidence）两个阈值是描述关联规则的两个最重要的概念。一个项目组出现的频率称为支持度（Support），反映关联规则在数据库中的重要性。而置信度衡量关联规则的可信程度。如果某条规则同时满足最小支持度（Min-Support）和最小置信度（Min-Confidence），则称它为强关联规则。用概率公式来表示支持度和置信度，如下所示：

$$\text{Support}(AB) = P(AB)$$

$$\text{Confidence}(AB) = P(B \mid A)$$

### 4.5.2　关联规则数据挖掘过程

关联规则数据挖掘一般分成两个阶段：

- 第一阶段必须从原始资料集合中，找出所有高频项目组（Large Itemsets）。高频的意思是指某一项目组出现的频率相对于所有记录而言，必须达到某一水平。以一个包含 A 与 B 两个项目的 2-itemset 为例，我们可以求得包含{A,B}项目组的支持度，若支持度大于等于所设定的最小支持度（Minimum Support）门槛值时，则{A,B}称为高频项目组。一个满足最小支持度的 k-itemset，则称为高频 k-项目组（Frequent k-itemset），一般表示为 Large k 或 Frequent k。算法从 Large k 的项目组中再试图产生长度超过 k 的项目集 Large k+1，直到无法再找到更长的高频项目组为止。
- 关联规则挖掘的第二阶段是要产生关联规则（Association Rules）。从高频项目组产生关联规则，是利用前一阶段的高频 k-项目组来产生规则，在最小可信度（Minimum Confidence）的条件门槛下，若一规则所求得的可信度满足最小可信度，称此规则为关联规则。例如，经由高频 k-项目组{A,B}所产生的规则，若其可信度大于等于最小可信度，则称 AB 为关联规则。

就尿不湿和啤酒这个案例而言，使用关联规则挖掘技术，对交易资料库中的纪录进行资料挖掘，首先必须要设定最小支持度与最小可信度两个门槛值，在此假设最小支持度 Min-upport=5%，且最小可信度 Min-onfidence=65%。因此符合此超市需求的关联规则将必须同时满足以上两个条件。若经过挖掘过程所找到的关联规则{尿不湿，啤酒}，满足上述条件，将可接受{尿不湿，啤酒} 的关联规则。用以下公式可以描述：

Support（尿不湿，啤酒）>=5% and Confidence（尿不湿，啤酒）>=65%。

其中，Support（尿不湿，啤酒）>=5%于此应用范例中的意义为：在所有的交易记录资料中，至少有 5%的交易呈现尿不湿

与啤酒这两项商品被同时购买的交易行为。Confidence（尿不湿，啤酒）>=65%于此应用范例中的意义为：在所有包含尿不湿的交易记录资料中，至少有 65%的交易会同时购买啤酒。因此，今后若有某消费者出现购买尿不湿的行为，我们将可推荐该消费者同时购买啤酒。这个商品推荐的行为则是根据{尿不湿，啤酒}关联规则而定，因为就过去的交易记录而言，支持了“大部分购买尿不湿的交易，会同时购买啤酒”的消费行为。

从上面的介绍还可以看出，关联规则挖掘通常比较适用于记录中的指标取离散值的情况。如果原始数据库中的指标值是取连续的数据，则在关联规则挖掘之前应该进行适当的数据离散化（实际上就是将某个区间的值对应于某个值），数据的离散化是数据挖掘前的重要环节，离散化的过程是否合理将直接影响关联规则的挖掘结果。

### 4.5.3 关联规则的分类

按照不同情况，关联规则可以进行如下分类：

- 基于规则中处理的变量的类别，关联规则可以分为布尔型和数值型。

布尔型关联规则处理的值都是离散的、种类化的，它显示了这些变量之间的关系；而数值型关联规则可以和多维关联或多层关联规则结合起来，对数值型字段进行处理，将其进行动态的分割，或者直接对原始的数据进行处理，当然数值型关联规则中也可以包含种类变量。例如：性别=“女”>=职业=“秘书”，是布尔型关联规则；性别=“女”>=avg（收入）=3300，涉及的收入是数值类型，所以是一个数值型关联规则。

- 基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

在单层的关联规则中，所有的变量都没有考虑到现实的数据是具有多个不同的层次的。而在多层的关联规则中，对数据的多

层性已经进行了充分的考虑。例如，IBM 台式机>=Sony 打印机，是一个细节数据上的单层关联规则；台式机>=Sony 打印机，是一个较高层次和细节层次之间的多层关联规则。

- 基于规则中涉及的数据的维数，关联规则可以分为单维和多维。

在单维的关联规则中，我们只涉及数据的一个维，如用户购买的物品；而在多维的关联规则中，要处理的数据将会涉及多个维。换成另一句话，单维关联规则是处理单个属性中的一些关系；多维关联规则是处理各个属性之间的某些关系。例如啤酒>=尿不湿，这条规则只涉及用户的购买的物品；性别=“女”>=职业=“秘书”，这条规则就涉及两个字段的信息，是两个维上的一条关联规则。

### 4.5.4　Apriori 算法的执行实例

下面我们以数据的实例来看最著名的 Apriori 算法的过程。重申一下 Apriori 算法的核心思想：基于自下而上的递推算法，从一个项目子集扩展到两个项目子集，从两个扩展到三个，直到无法扩展为止。

从数据集中我们找到以下 8 条消费记录，并以 3 作为最小支持度。

| 交易标号 | 销售内容 |
|---|---|
| 1 | 牛奶，冰淇淋，果酱，面包 |
| 2 | 冰淇淋，果酱，面包，咖啡 |
| 3 | 牛奶，面包，果酱 |
| 4 | 牛奶，咖啡 |
| 5 | 牛奶，面包，巧克力 |
| 6 | 冰淇淋，面包，咖啡 |
| 7 | 牛奶，果酱，面包，香蕉 |
| 8 | 咖啡，面包，葡萄 |

所有满足最小支持度 3 的 1-项频集如下，其中的支持度是该产品在整个数据集中出现的次数。比如牛奶就出现了 5 次，而冰淇淋出现了 3 次。

| 支 持 度 | 销售内容 |
|---|---|
| 5 | 牛奶 |
| 3 | 冰淇淋 |
| 4 | 果酱 |
| 3 | 咖啡 |
| 6 | 面包 |

递归执行，所有满足最小支持度 3 的 2-项频集如下，这里出现最多的频集是{牛奶，面包}和{面包，果酱}，各自出现了 4 次。

| 支 持 度 | 销售内容 |
|---|---|
| 3 | 面包，咖啡 |
| 4 | 牛奶，面包 |
| 3 | 冰淇淋，面包 |
| 4 | 面包，果酱 |

再次递归执行，所有满足最小支持度 3 的 3-项频集只剩下一条。

| 支 持 度 | 销售内容 |
|---|---|
| 3 | 牛奶，果酱，面包 |

那么{牛奶，果酱，面包}就是我们要的满足最小支持度 3 的 3-项频集，而且也没有 4-项频集。

### 4.5.5 关联规则挖掘算法的研究与优化

在 Agrawal 提出关联规则之后，诸多的研究人员对关联规则的挖掘问题进行了大量的研究。他们的工作包括对原有的算法进行优化，如引入随机采样、并行的思想等，以提高算法挖掘规则

的效率。对关联规则的应用进行推广。另外也有独立于 Agrawal 的频集方法的工作，以避免频集方法的一些缺陷，探索挖掘关联规则的新方法等。还有一些工作注重于对挖掘到的模式的价值进行评估。

Agrawal 的经典频集方法要求多次扫描可能很大的交易数据库，即如果频集最多包含 10 个项，那么就需要扫描交易数据库 10 遍，如果数据量大，消耗系统资源颇为可观。当我们采用 Agrawal 的算法做大型数据仓库的商品关联时，交易数据动辄上亿条，完整多次扫描这些交易数据太过费时，这就需要引进修剪技术。修剪技术（Pruning）是用来减小数据候选集（用 Ck 表示）大小的一种方法，由此可以显著地改进生成所有频集算法的性能。算法中引入的修剪策略基于这样一个性质：一个项集是频集当且仅当它的所有子集都是频集。那么，如果 Ck 中某个候选项集有任意一个（k-1）-子集不属于频集，这个项集就可以被修剪掉而不再被考虑。这个修剪过程可大幅降低计算关联规则所花的代价。

关联规则的增量式更新问题在实际应用中也经常出现，数据仓库中一直会有新增的内容，如网页信息、商品信息和交易信息等的加入。我们把增量式更新主要归纳成两个问题：

- 在给定的最小支持度和最小置信度下，当一个新的事物数据集添加到旧的事物数据集中时，如何修改和生成新的关联规则。这里常用的经典的解决方案有一个与算法 Apriori 相一致的算法 FUP。
- 给定数据集，在最小支持度和最小置信度发生变化时，如何修改和生成数据集中的关联规则。这里经典的解决方案有算法 IUA 和 PIUA。

对于互联网上海量的数据集，我们需要采用并行处理方法，才能大幅提高效果，在可以接受的时间内完成数据挖掘任务。目前数据挖掘专家和学者们已经提出的并行挖掘关联规则的算法有：Agrawal 等人提出的 CD（Count Distribution）计数分布算法、

CaD（Candidate Distribution）候选集分布算法、DD（Data Distribution）数据分布算法，Park 等人提出的 PDM 算法，以及 Chueng 等人提出的 DMA 和 FDM 算法等。其中 CD 计数分布算法是经典算法 Apriori 在并行环境下的应用，而 DMA 和 FDM 算法是基于分布式数据库的关联规则挖掘算法。

在研究采掘关联规则的过程中，许多学者发现在一些实际应用中，对于很多的应用来说，由于数据分布的分散性，数据比较少，所以很难在数据最细节的层次上发现一些强关联规则。要想在原始的概念层次上发现强的（Strong）和有意义的（Interesting）关联规则是比较困难的，因为好多项集往往没有足够的支持数。当我们引入概念层次后，就可以在较高的层次上进行挖掘。虽然较高层次上得出的规则可能是更普通的信息，但是对于一个用户来说是普通的信息，对于另一个用户却未必如此。所以数据挖掘应该提供这样一种在多个层次上进行挖掘的功能。

概念层次在要采掘的数据库中是经常存在的，比如在一个超市中会存在这样的概念层次：蒙牛牌牛奶是牛奶，伊利牌牛奶是牛奶，王子牌饼干是饼干，康师傅牌饼干是饼干等。如果我们只是在数据基本层发掘关系，{蒙牛牌牛奶，王子牌饼干}，{蒙牛牌牛奶，康师傅牌饼干}，{伊利牌牛奶，王子牌饼干}，{伊利牌牛奶，康师傅牌饼干}都不符合最小支持度。如若上升一个层级，我们会发现{牛奶，饼干}的关联规则是有一定支持度的。

我们称高层次的项是低层次项的父亲层次（Parent），这种概念层次关系通常用一个有向非循环图（DAG）来表示。这样我们就可以在较高的概念层次上发现关联规则。

根据规则中涉及的层次和多层关联的规则，我们可以把关联规则分为同层关联规则和层间关联规则。多层关联规则的挖掘基本上可以沿用“支持度-置信度”的框架。不过，在支持度设置的问题上有一些要考虑的东西。

同层关联规则可以采用两种支持度策略：

- 统一的最小支持度。对于不同的层次，都使用同一个最

小支持度。这样对于用户和算法实现来说都比较容易，但是弊端也是显然的。

- 递减的最小支持度。每个层次都有不同的最小支持度，较低层次的最小支持度相对较小。同时还可以利用上层挖掘得到的信息进行一些过滤的工作。层间关联规则考虑最小支持度时，应该根据较低层次的最小支持度来定。

基于要采掘的数据库中的概念层次和发现单一概念层次中的关联规则的算法，学者们提出了许多高效发现一般或多层关联规则的算法，主要有：Han 等人的 ML_T2L1 和 R.Srikant 等人的 Cumulate、Stratify 等算法。

算法 ML_T2L1 的基本思想是首先根据要发现的任务从原事务数据集生成一个根据概念层次信息进行编码的事务数据集，利用这个具有概念层次信息的新生成的数据库，自顶向下逐层递进地在不同层次发现相应的关联规则。它实际上是算法 Apriori 在多概念层次环境中的扩展。根据对在发现高层关联规则过程中所用的数据结构和所生成的中间结果共享方式的不同，对于算法 ML_T2L1 的研究已经产生了多个变种：ML_T1LA、ML_TML1 和 ML_T2LA 等。

算法 Cumulate 的基本思想与 Apriori 完全一样，只是在扫描到事务数据集某一事务时，将此事务中所有项的“祖先”经过“剪枝”等优化加入到本事务中。

以上我们讨论的基本上都是同一个字段的值之间的关系，比如用户购买的物品。换句话说就是在单维或者叫维内的关联规则，这些规则很多都是在交易数据库中挖掘的。

但是对于实际应用来说，多维的关联规则可能是更加有价值的。例如我们找出的关联规则是如果年龄在 20 到 30 之间的顾客，职业又是学生，那么该顾客会购买笔记本电脑。这条规则可以这样表示：

年龄（X，“20…30”）+职业（X，“学生”）==>购买（X，“笔记本电脑”）。

在这里我们涉及三个维度上的数据：年龄、职业、购买。

根据是否允许同一个维重复出现，可以又细分为维间的关联规则（不允许维重复出现）和混合维关联规则（允许维在规则的左右同时出现）。以下的这个规则就是混合维关联规则，因为顾客 X 购买笔记本电脑的事件出现在左部，而顾客 X 购买打印机的事件出现在右部。

年龄（X，“20…30”）+购买（X，“笔记本电脑”）==>购买（X，“打印机”）

在挖掘维间关联规则和混合维关联规则时，还要考虑不同的字段种类，是种类型还是数值型。对于种类型的字段，本节中前面的算法都可以处理。而对于数值型的字段，需要进行一定的处理之后才可以进行。处理数值型字段的方法基本上有以下几种：

- 数值字段被分成一些预定义的层次结构。这些区间都是由用户预先定义的。得出的规则称为做静态数量关联规则。
- 数值字段根据数据的分布分成了一些布尔字段。每个布尔字段都表示一个数值字段的区间，落在其中则为 1，反之为 0。这种分法是动态的。得出的规则叫做布尔数量关联规则。
- 数值字段被分成一些能体现它含义的区间。它考虑了数据之间的距离的因素。得出的规则叫做基于距离的关联规则。
- 直接用数值字段中的原始数据进行分析。使用一些统计的方法对数值字段的值进行分析，并且结合多层关联规则的概念，在多个层次之间进行比较从而得出一些有用的规则。得出的规则叫做多层数量关联规则。

有很多数据挖掘的理论书籍详细介绍了关联算法的种类和算法细节，我们在此就不多此一举了。这里只是把我们需要了解的概念和算法做了大致的介绍。在第 10 章讲述数据挖掘在电子商务中的应用时，会谈到关联规则算法的实际应用。

## 4.6 序列挖掘（Sequence Mining）

在数据挖掘中的序列挖掘指的是从一个序列（Sequence）中的数据找出统计规律。

如果在序列中可能出现的单元是来自于一个有限的集合，这个集合可以称为 Alphabet（字母表），而对此类序列做挖掘的算法可以称为 String Mining（字符串挖掘）。比如生物遗传学中所有的 DNA 基因序列都是由“A”，“G”，“C”和“T”四个字母组成氨基酸的形成的，不同的排列是生物信息学（Bioinformatics）所研究的课题，而该学科采用的不少数据挖掘算法都是序列算法。在生物信息学（Bioinformatics）上的数据挖掘应用不在本书的讨论范围之内。

而打上时间标签的 Time Series（时间序列）中处理的数据是在不同时间点上收集到的数据，是序列算法中重要的一个峰值。这类数据反映了某一事物、现象等随时间的变化状态或程度。而且在时间序列算法中要求时间段的区分是相同间隔的。如我国国内生产总值从 1949 年到 2009 年的变化，Facebook 股价从 2012 年 9 月 1 日到 10 月 12 日的变化都是时间序列数据。

时间序列数据可作年度数据、季度数据、月度数据等细分，其中很有代表性的季度时间序列模型就是因为其数据具有四季一样的变化规律，虽然变化周期不尽相同，但是整体的变化趋势都是按照周期变化的。

时间序列通常会在连续的时间流中截取一个时间段，然后让时间段在整个时间轴上滑动，从而获得需要的训练数据集。比如金融分析师会根据前面 29 天的货币汇率估计第 30 天的货币汇率变化；电子商务企业会根据前面 99 天的销售情况估计第 100 天和第 101 天的销售数据；网站联盟会根据前面 13 个月的点击率和收入情况估计第 14 个月的总体点击率和收入情况，等等。

时间序列的简单表示大多用 Line Chart（折线图）。每天股市

收盘价作为数据点采集的数据序列是时间序列的一种。如图 4-4 是互联网上两家公司从 2012 年 6 月到 9 月的股价变化对比表。


图 4-4　时间序列数据示意图

时间序列是统计学专业的重要课程之一，对时间序列的研究一般是要建立在一定的计量经济学基础上，因为很多计量经济学的模型对于处理时间序列数据都是相当有效的。计量经济学（Econometrics）是以经济学和数理统计学为方法论作为基础，对于经济问题试图用数量和经验两者进行综合的经济学分支。对于时间序列上的随机过程，在统计学上有三大类算法：Autoregressive Model（自回归模型），Integrated Model（整合模型）和 moving average（MA 移动平均模型）。在微软的 MS SQL Server 2005 中提供的分析工具 SSAS 有一个时间序列算法的程序包，图 4-5 是用该工具对于三个产品销售数据的预测。其中左半边是历史数据，而中间的竖线右面是根据 Autoregressive Model（自回归模型）生成的预测。

根据时间序列型数据，由历史的和当前的数据去推测未来的数据，也可以认为是以时间为关键属性的关联知识。时间序列预测方法有经典的统计方法、神经网络和机器学习等。1968 年 Box 和 Jenkins 提出了一套比较完善的时间序列建模理论和分析方法，这些经典的数学方法通过建立随机模型，如自回归模型、自

回归滑动平均模型、求和自回归滑动平均模型和季节调整模型等，进行时间序列的预测。由于大量的时间序列是非平稳的，其特征参数和数据分布随着时间的推移而发生变化。因此，仅仅通过对某段历史数据的训练，建立单一的神经网络预测模型，还无法完成准确的预测任务。为此，人们提出了基于统计学和基于精确性的再训练方法，当发现现存预测模型不再适用于当前数据时，对模型重新训练，获得新的权重参数，建立新的模型。也有许多系统借助并行算法的计算优势进行时间序列预测。

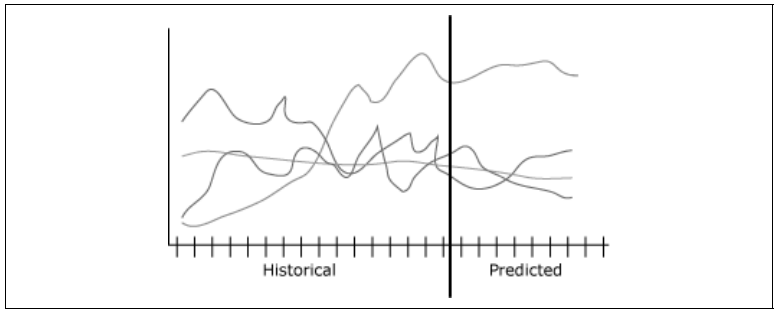



图 4-5　MS SQL 预测分析示意图

我们经常遇到这样的实际问题：

- 如果一个客户{购买了车}，那么他很可能需要在一周内{购买汽车保险}。
- 在股市中，如果{民生银行和工商银行股价上涨}，那么很可能在短期内{农业银行的股价上涨}。

而在此类序列中查找频繁集（Frequent Itemset）的工作其实是 4.4 节关联算法的一种应用。我们可以使用关联算法中的 apriori 算法或 FP-Growth 算法来处理这些应用。

## 4.7 数据挖掘建模语言 PMML

数据挖掘语言标准化是近年来的一个方向，而 XML 格式是一个不错的选择。数据挖掘建模语言 PMML 是 Predictive Model

Markup Language（预言模型标记语言）的缩写，由 Data Mining Group（数据挖掘组织）设计的。PMML 是基于 XML 格式的对数据挖掘模型进行描述和定义的语言，使数据挖掘系统在模型定义和描述方面有法可依，各种数据挖掘系统可以共享模型，又可以在应用程序系统中间嵌套数据挖掘模型，不需要独自开发，就能使数据挖掘达到深度挖掘的目的。预言模型标记语言 PMML 是一种基于 XML 的数据挖掘建模语言，利用 XML 描述和存储数据挖掘模型，使用标准的 XML 解析器对 PMML 解析，可以得到预计的输入和输出数据类型。

如图 4-6 所示，PMML 主要由：标题（Header）、数据字典（Data Dictionary）、数据流（Data Flow）、挖掘模型（Mining Schema）、数据转换（Data Transformation）、模型（Model）、数据挖掘计划（Mining Schema）、目标（Targets）等六个部分组成。其中数据转换有 Derived Values、Statistics、Taxonomy、Normalization 等。模型包括数据挖掘常用的模型，比如决策树、贝叶斯模型、各种回归模型、序列、关联模型、神经网络、聚类等。数据挖掘计划包括了缺失数据的处理和替代能。而目标指的是在数据挖掘处理之后的评估。

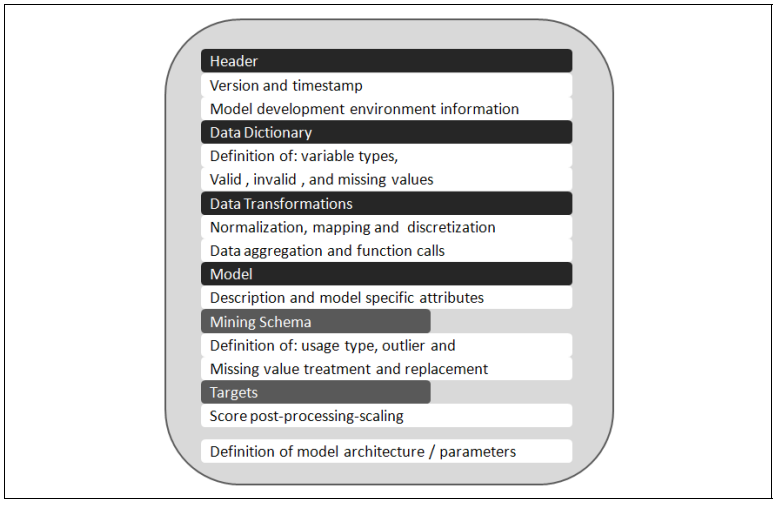



图 4-6　PMML 格式示意图

对于复杂的数据挖掘任务，数据来自多个数据源，可能有多个数据挖掘模块参与数据挖掘过程，需要在各个模块之间交换数据和结果。PMML 的主要组成部分拥有这种灵活的模型交换能力和数据格式转换能力，并实现模型与数据和工具部分分离。因 PMML 是基于 XML 的数据挖掘建模语言，适合部分学习、元学习、分布式学习的数据挖掘应用程序。

在市场上的数据挖掘工具，包括 IBM Intelligent Miner 和 Microsoft SQL Server 2005 等都整合了 PMML，所以在这些工具上可以使用 PMML 来做数据挖掘。

PMML 可以在这个网站下载：http://sourceforge.net/projects/pmml/，现在最新的版本是 Version 4.1。

## 4.8 本章相关资源

- 本章相关参考文献：

[1]（加）Jiawei Han, Micheline Kamber. 数据挖掘概念与技术[M], 范明，孟小峰, 译. 机械工业出版, 2007.

[2]（美）Olivia Parr Rud. 数据挖掘实践[M]. 朱扬勇等, 译. 机械工业出版, 2003.

[3]（美）Richard J.Roiger,（美）Michael W.Geatz. 数据挖掘教程[M]. 翁敬农, 译. 清华大学出版社, 2003.

[4] 陈耿, 朱玉全, 杨鹤标等. 关联规则挖掘中若干关键技术的研究[J]. 计算机研究与发展, 2005, 42（10）.

[5] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg. *Top 10 Algorithms in D ata Mining*. Knowledge Information System (2008) 141-37.

[6] Quinlan, J. R. 1993. *C4.5*: *Programs for Machine Learning*.

Morgan Kaufmann Publishers Inc. L. Breiman, J. Friedman, R. Olshen, C. Stone. *Classification an d Regression Trees*. Wadsworth, Belmont, CA, 1984.

[7] Hastie, T. and Tibshirani, R. 1996. *Discriminant Adaptive Nearest Neighbor Classification*. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 1996, 18(6): 607-616.

[8] Hand, D.J., Yu, K., 2001. *Idiot's Bayes*: *Not So Stupid After All*? International Statistics Rev. 69, 385-398.

[9] L. Page, S. Brin, R. Motwani, T. Winograd. *The Pagerank Citation Ra nking*: *Br inging O rder t o the W eb*. Technical Report, Stanford University, 1999.

[10] R. Agrawal, R. Srikant. *Fast al gorithms f or m ining association r ules. Pr oc.* 20th int'l conf. very large databases, Santiago, Chile, Sept. 1994, 487-499.

[11] J. S. Park, et al. *Using a hash-based m ethod w ith transaction tr imming f or m ining association r ules*. IEEE Transactions on knowledge and data engineering, 1997, 9(5), 813-825.

[12] R. Srikant, R. Agrawal. *Mining association rules with item constrains*. Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997, 67-73.

[13] R. Ng, L. V. S. Lakshmanan, J. Han and A. Pang. *Exploratory Mining a nd Prun ing Optimizations of Constrained Asso ciations Rules*. Proc. of 1998 ACM-SIGMOD Conf. on Management of Data, Seattle, Washington, June 1998, 13-24.

[14] Leo Breiman. *Bagging pr edictors*. Machine Learning, 24(2):123–140, 1996.

[15] Yan, X. and Han, J., *gSpan*: *G raph-Based Substructure*

*Pattern Mi ning*. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02) (December 09 - 12, 2002). IEEE Computer Society, Washington, DC.

[16] Varun Chandola, Shyam Boriah, and Vipin Kumar. *A Framework f or An alyzing Cat egorical Data* (*2009*). In Proceedings of SIAM Data Mining Conference, April 2009, Sparks, NV.

[17] Varun Chandola, Arindam Banerjee, and Vipin Kumar. *Anomaly D etection*: *A Survey* (*2009*) *ACM Co mputing Surveys*. Vol. 41(3), Article 15, July 2009.

- 本章相关网址：

[1] http://citeseer.comp.nus.edu.sg/agrawal94fast.html

[2] http://doi.acm.org/10.1145/342009.335372

[3] http://dx.doi.org/10.1016/S0169-7552(98)00110-X

[4] http://doi.acm.org/10.1145/233269.233324

[5] http://dx.doi.org/10.1006/jcss.1997.1504

[6] http://www.dmg.org/

[7] http://sourceforge.net/projects/pmml/

# 第 5 章
# 在进行数据挖掘之前

在第 3 章数据仓库中我们讨论了数据存储和收集的问题。在本章中主要讨论在数据挖掘过程之前需要的各种准备工作。在 2.5.1 节中提到数据挖掘过程有 8 个步骤，本章主要讲述其中第 2 步到第 5 步，也就是数据集成、数据规约、数据清理和数据变换，其中数据规约、数据清理和数据变换的步骤又合称数据预处理。

在所有的数据挖掘过程中，数据准备工作占用的时间都会在一半以上。没有数据，数据挖掘过程就是巧妇难为无米之炊。而如果没有清理过的合适、正确的数据，数据挖掘则是没有基础的，而且分析出的结果也是没有价值的。

美国学者最早于 20 世纪 80 年代提出了“数据融合”一词。但到目前为止，数据融合尚未有一个统一的定义。在本章中，数据融合仅限于数据集合的范畴，也就是数据层的数据融合，即把数据融合的思想引入到数据预处理的过程中，加入数据的智能化合成，产生比单一信息源更准确、更完全、更可靠的数据进行估计和判断，然后存入到数据仓库或数据挖掘模块中。

常见的数据融合方法有：静态的融合方法，如加权最小平方等；动态的融合方法，如递归加权最小平方、卡尔曼滤波、小波变换的分布式滤波等；基于统计的融合方法，如马尔可夫随机场、最大似然法、贝叶斯估值等；基于信息论算法的方法，如聚集分析、自适应神经网络、表决逻辑、信息熵；基于模糊集理论的聚类方法等。

# 5.1 数据集成

数据集成是将多个数据源中的数据（数据库、数据立方体或一般文件）结合起来存放到一个一致的数据存储（如数据仓库）中的过程。我们做数据集成是为了对数据进行汇总和数据概化。由于不同学科方面的数据集成涉及不同的理论依据和规则，而不同的数据库表格的定义差异也是比较大的，因此，数据集成可以说是数据预处理中比较需要专业知识的一个步骤。

数据集成中包括如下两个部分。

- 数据集成：将多个数据源中的数据整合到一个一致的存储中。
- 模式集成：整合不同数据源中的元数据。

数据集成需要用行业相关知识来处理实体识别问题，以匹配来自不同数据源的现实世界的实体，比如在一个企业的两个数据源中我们分别以 cust-id 和 customer_no 来标识用户，那么在数据集成的时候，我们需要把标识相同的客户合在一起：

```
A.cust-id==B.customer_no
```

在该企业中这两个用户标识定义的方式是相同的，所以这还是相对简单的数据集成。如果在两个数据源中用户标识的定义不同，那么我们可能需要通过程序来做判断。比如在一个数据源中，企业的 ID 是像这样的一个数字“23442”，而在另一个数据源中，企业是地域名称加上这个数字形成的字符串是“SH23442”，那么我们可以这样作比较：

```
string(A.cust-id)== string(B.customer_no).substr(2,length-2)
```

在数据集成中，我们需要检测并解决数据值的冲突。对现实世界中的同一实体，来自不同数据源的属性值可能是不同的。可能的原因有很多种，比如不同的数据表示和不同的度量等。

如两个数据源中，一个是以“男”和“女”来区分性别，而

另一个是以“F”和“M”来表示。那么在集成的时候，我们需要把两个数据源的数据变成一致的。而更加常见的是关于日期的格式，在有的数据源中，日期是一个数值；而有的数据源中，日期是以“××××年×月×日”的字符串方式存储；还有的是以“YY/MM/DD”的格式组成的。

我们做数据集成的目的非常明确，就是把数据从不同的信息源整合到同一个数据平台之上以便于数据挖掘。一般来说，数据集成的步骤结束之后我们才开始进行数据预处理的过程。当然，如果每个数据源需要做的处理工作不尽相同，那么，我们也可以在每个数据源上先做数据预处理，之后再进行数据集成的工作。

## 5.2 为何要做数据预处理

如我们在第 2 章中提到的，数据挖掘的步骤中有 3 步是属于对数据的准备工作——数据规约、数据清理及数据变换，这三个步骤合称数据预处理。

数据预处理的目的是为了提高数据质量，使数据挖掘的过程更加有效，更加容易，同时也提高挖掘结果的质量。数据预处理的作用主要是清理其中的噪声数据、空缺数据、错误数据和不一致数据。数据预处理的工作在数据挖掘之前是必须要做的。

我们要做数据预处理的原因主要有以下两条：

- 现实世界的数据是“杂乱的”——数据多了，什么问题都会出现。有些数据是不完整的，重要属性缺少属性值，或仅包含聚集数据。有些数据包含噪声值，有错误或者“孤立点数据”。有些数据因为在编码或者命名上存在差异，是不一致的。
- 数据挖掘需要高质量的数据。数据质量可以从多个维度来衡量：精确度、完整度、一致性、可访问性、置信度等。没有高质量的数据，就没有高质量的挖掘结果。高

质量的决策必须依赖高质量的数据，数据仓库需要对高质量的数据进行一致地集成。

数据挖掘之前对数据的处理工作要占全过程 60％的工作量，数据集成、数据清理、数据变换和数据规约都是这个处理工作的一部分。这些数据处理技术对于提高数据挖掘模式的质量和结果的准确性，降低实际挖掘所需要的时间是必不可少的。

从图 5-1 中我们可以看到，存储在数据仓库中的数据在飞速增长，但是我们可以看到的是，相关数据（Relevant Data）的增长速度还比不上没有价值的数据增长的速度。

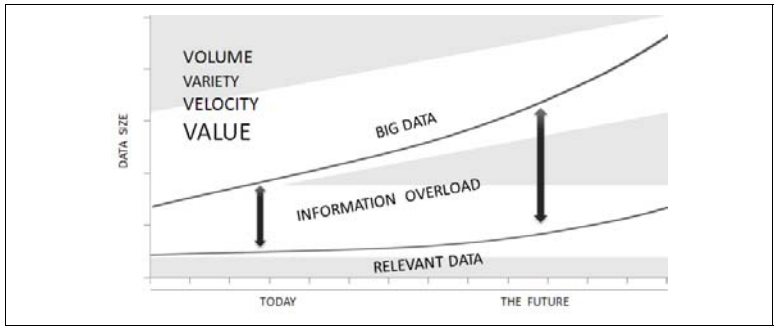



图 5-1　冗余数据的示意图

建立任何实际的知识获取系统，都需要认真研究数据的预处理问题，包括原始数据的采样、收集和整理。不同领域的原始数据，可以通过不同的方法取得，但是，取得的原始数据并不一定就适合直接用于知识获取，通常还需要进行预处理加工，对于原始数据资料中遗漏的信息，需要补充；对于原始数据资料中错误的信息，需要更正；对于原始资料中值域为实数值的数据，可能还需要进行离散化。

数据清洗试图填补训练集中的空缺值、识别孤立点、消除噪声、纠正数据中的不一致。对于空缺值的处理，通常有忽略元组、人工填写空缺值、使用全局常量填充、使用属性平均值填充、使用与给定元组同一类的样本平均值填充，以及使用最可能的值填充等方法。

研究发现，数据挖掘出现错误结果多半是由数据源的质量引起的。因此，重视原始数据的获取，从源头上尽量减少错误和误差，尤其是减少人为误差，尤为重要。

## 5.3 数据预处理

接下来我们来看如何对数据做预处理工作。预处理工作通常要占用整个数据挖掘流程 60%的时间，而当需要做的预处理工作比较繁复时，这一数字通常会更高。而当我们把数据整理完成后，下一步的数据挖掘过程就是一个程序化的工作了。

### 5.3.1 数据清理

数据挖掘的结论依赖于数据质量。有个缩写词 GIGO（Garbage In，Garbage Out），意思是进来的是垃圾，出去的也一定是垃圾。在数据分析和数据挖掘领域尤其如此。无论专家多有经验，也无论数据挖掘算法的实现再完美，也不可能从一堆垃圾中发现宝石。

数据清理过程通过填写缺失的数值、光滑噪声数据、识别或删除离群点并解决不一致性来“清理”数据。主要是为达到如下的目标：

- 格式标准化。
- 异常数据清除。
- 错误纠正。
- 重复数据的清除。

对来自多个系统或数据源的数据生成的数据仓库的数据清理进程中，重要的一步是解决不正确的拼写、多个系统之间冲突的拼写规则和冲突的数据之类的错误。在数据中出现的编码错误或录入资料时的错误，会直接威胁到数据挖掘的效果。

数据清理能解决数据文件中的人为误差，以及数据文件中一

些对统计分析结果影响较大的特殊数值。对原始数据中的缺失数据、重复数据、异常数据进行处理，提高数据质量的过程包括四个环节，如图 5-2 所示。

（1）处理缺失数据。数据并不总是完整的，例如，数据仓库的表中，很多条记录的对应字段没有相应值，比如销售表中的关于顾客的信息除了名字外，其他的各个属性都可能是缺失的。

（2）处理重复数据。除了真正重复的数据外，这里还包括属性冗余和属性数据的冗余。去除了重复记录后，我们可以提高挖掘的速度和质量。

（3）处理噪声数据。所谓噪声数据是指数据中存在着错误或异常的数据。一个测试数据变量中的随机错误或偏差引起不正确属性值，噪声可能是由误差造成的。对有些带有错误的数据元组，结合数据所反映的实际问题进行分析进行更改或删除或忽略。同时也可以结合模糊数学的隶属函数寻找约束函数，根据前一段历史趋势数据对当前数据进行修正。

（4）处理异常数据。在大规模数据集中，通常存在着不遵循数据模型的普遍行为的样本。这些样本和其他部分数据有很大不同或不一致。这些不一致的数据点不一定是噪声。我们需要尽量剔除真正不正常的数据，而保留虽然看起来不正常，但实际上是真实的数据。

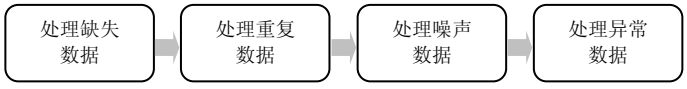



图 5-2　数据清理的四个环节

对于不同的大数据情况，上述的四个环节不是一定都需要的。我们对数据做了合理的数据清理之后，可以进入数据挖掘流程的下一个步骤。

#### 5.3.1.1　处理缺失数据

在很多情况下，我们得到的待处理的信息表并不是一个完备的信息表，表中的某些属性值是被遗漏的，我们无从知道其原始

值，这也是信息系统具有不确定性的一种主要原因。对于这种情况，目前主要通过以下途径来对信息表中的遗漏数据进行补齐。

- 第一种途径是简单地将存在空缺（遗漏）属性值的实例记录删除，从而得到一个完备的信息表。当一个记录中有多个属性值空缺、特别是关键信息丢失时，已不能反映真实情况，它的效果非常差。虽然这种方法不是严格意义上的数据补齐，然而在信息表数据量巨大，并且有遗漏属性值的实例记录的数量远远小于信息表所包含的记录数的情况下，这种方法在删除不完整记录之后并不太影响信息表中信息的完整性，是一种可取的处理方法。但是，当信息表中的信息较少，或者存在遗漏信息的实例对于信息表中的信息相对不是一个可以忽略的比例时，这种方法就会严重影响信息表中的信息量，不能采用这种方法来对信息表的数据进行补齐。
- 第二种途径是采用人工方式填写空缺值：工作量大。而在我们面临的海量数据情况下可行性极低。
- 第三种途径是将空缺（遗漏）属性值作为一种特殊的属性值来处理，比如使用 unknown 或者∞等不同于其他的任何属性值，或者使用默认值。这样，一个不完备的信息表就成为了完备信息表。
- 第四种途径是采用统计学原理，根据信息表中其余实例在该属性上的取值分布情况来对一个遗漏属性值进行估计补充，可以采用属性的平均值填充空缺值；或者使用同类样本预测最可能的值填充空缺值；或者使用贝叶斯公式和判定树这样的基于推断的方法来填充，这样不会影响信息表中包含的信息量。但是如果我们采用的填补方式不合适，那么采用这种方式来填空会影响到算法的结果。

在时间序列中，如果数据属于时间局部性的缺失，则可采用近阶段数据的线性插值法进行补缺；若缺失的时间段较长，则应

该尽量试图采用该时间段的历史数据恢复丢失数据。如果缺失的数据属于数据的空间缺损，我们可以用其周围数据点的信息来代替，且对相关数据作备注说明，以备查用。

#### 5.3.1.2　处理重复数据

在我们做数据清理的过程中，除了真正重复的数据外，这里还包括属性冗余和属性数据的冗余。

- 属性冗余：若通过因子分析或经验等方法确信部分属性的相关数据足以对信息进行挖掘和决策，可通过用专业常识或者相关数学方法找出具有最大影响属性因子的属性数据即可，其余属性则可删除。

我们集成多个数据库时，经常会出现冗余数据，同一属性在不同的数据库中会有不同的字段名，比如“user_birthday”和“用户_生日”，这时我们只需要保留其中的一个字段就可以。或者其中的一个属性可以由另外一个表导出，如“年龄”字段可以从“生日”字段中导出等。当然，这些冗余可能在数据集成的步骤中已经解决。

- 属性数据的冗余：若某属性的部分数据足以反映该问题的信息，则其余的可删除。若经过分析，这部分冗余数据可能还有他用，则先保留并作备注说明。

举例来说，在数据仓库中的“地址”一栏中包含了“国家”、“省份”和“城市”信息，而这些信息其实和其他属性是重复的。我们可以将“地址”中的“国家”、“省份”和“城市”信息删除而不影响数据挖掘工作。

之前我们在数据集成的工作中，如果能够仔细将多个数据源中的数据集成起来，可以减少或避免结果数据中的冗余与不一致性，从而可以提高挖掘的速度和质量。

#### 5.3.1.3　处理噪声数据

为了避免 GIGO，我们需要把数据中的噪声尽量多地去除。特别是当我们需要从数据中发现异常点或者偏离常规数据模型

时，这个问题尤其突出。当我们在寻找百万分之一的模型时，第二个小数位的偏离就会起作用。处理噪声数据，目前最广泛的是应用数据平滑技术。

处理噪声数据的方法，主要有以下几种：

- 采用分箱技术来检测周围相应属性值进行局部数据平滑。
- 利用聚类技术，根据要求检测孤立点数据，并进行修正。
- 利用回归函数和时间序列分析的方法进行修正。
- 采用计算机检测出怀疑噪声数据，然后对它们进行人工判断的方式等。

下面我们介绍一下在数据去噪过程中最常用的数据平滑分箱（Binning）算法。分箱算法的主要目的是去噪，将连续数据离散化，增加粒度。

首先排序数据，并将他们分到等深的箱中，然后可以按箱的平均值平滑、按箱的中值（Median）平滑、按箱的边界平滑等聚类。

下面我们拿商品价格做实例进行分箱。商品价格排序后的数据（单位：元）：10.01，80，150，209.99，210，240.02，250，280，339.04。

- 划分为（等深的）箱并取整之后：

箱 1：10，80，150

箱 2：210，210，240

箱 3：250，280，340

- 我们如果用箱平均值平滑并取整之后：

箱 1：80，80，80

箱 2：220，220，220

箱 3：290，290，290

在分箱后，我们就用箱中新的数值来代替原有的数值，这样数据中不应该有的噪声就被去除了。因为当我们在做数据分析时，并不关心数据值原来究竟是 10.01 还是 10.00。

#### 5.3.1.4　处理异常数据

在大规模数据集中，通常存在着不遵循数据模型的普遍行为的样本。这些样本和其他部分数据有很大不同或不一致。

对于不同类型的数据，我们会采用不同方式来做异常数据处理：

- 针对时间序列数据，我们可以采取基于移动窗口理论等方法实现对异常点的检测；
- 针对空间数据，可以采取基于移动曲面拟合法的方法实现对异常点的检测；
- 针对多维数据，可以采取聚类分析法实现异常点的检测。

当对检测出来的异常点经过验证，判定为误差时，剔除后可提高数据挖掘算法的效率和准确度，特别是如果我们在数据挖掘过程中选择的算法对误差比较敏感。

这里特别要注意的问题是有些和其他部分不一致的数据样本并不是噪声，而是异常点（Outlier）。异常点不同于噪声，其实是数据固有的可变性的结果。而当对检测出来的异常点判定为正常点时，重点分析这些异常点有时能发现隐含着重要的信息，甚至有些数据挖掘的算法其主要目的就是为了找出这些异常点。

### 5.3.2　数据转换

数据预处理中的数据转换，根据数据对象不同可以分成两类对于传统常规数据的数据变换和对于非常规数据的数据变换：

- 常规数据转换通常通过线性或非线性的数学变换方法等方式将数据转换成适用于数据挖掘的形式。常用的传统数据规范化方法有最小-最大规范化、Z-score 规范化也就是零-均值规范化、小数定标规范化等。
- 非常规数据的数据变换，根据数据的特性会有比较多的形式各异的转换方式。比如把音频和视频数据转换成系统指定的格式等。

数据转换和 5.3.3 节讲述的数据规约最主要的不同点在于数据转换基本都是无损的，而数据规约对于原始数据通常都是有损的。

常规数据转换采用线性或非线性的数学变换方法消除它们在空间、属性、时间及精度等特征表现的差异。常见的常规数据变换方法有以下几种，而具体采用哪种变换方法应根据涉及的相关数据的属性特点，根据研究目的可把定性问题定量化，也可把定量问题定性化进行数据的操作变换。

- 为了减少数据复杂度，用高层概念替换底层概念；
- 专注于数据规范化，使数据按比例缩放，落入特定区域；
- 做属性构造，通过一个或多个属性的变换计算构造出新的属性等。

数据转换对数据挖掘模型和输入数据集的要求有较强的依赖，针对不同的数据挖掘模型需要进行不同类型的数据转换。在数据转换阶段，我们通过一些数据转换工具来变换数据到我们可以做后续处理的范围内，或者说把数据标准化。

所谓数据标准化是把区间较大的数据整合到一个相对较规则的区间中，包含标准差标准化、极筹标准化和极差正规化等。我们来看几个最常用的数据转换方法。

（1）标准差标准化。

所谓标准差标准化是将各个记录值减去记录值的平均值，再除以记录值的标准差，即：

$$x'_{ij} = \frac{x_{ij} - x_{ia}}{S_i}$$

其中，$x_{ia}$ 为平均值，其表达式为：

$$x_{ia} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

设 $S_i$ 是标准差，有：

$$S_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - x_{ia})^2}$$

经过标准差标准化处理的所有记录值的平均值为 0，标准差为 1。

（2）极差标准化。

极差标准化是数据标准化的另外一种常用方式。对记录值进行极差标准化变换是将各个记录值减去记录值的平均值，再除以记录值的极差，即：

$$x'_{ij}=\frac{x_{ij}-x_{ia}}{\max(x_{ij})-\min(x_{ij})}$$

经过极差标准化处理后的观测值的极差等于 1。

（3）极差正规化。

极差正规化又是另外一种常用的数据标准化方式，可以把所有的观测值转化到[0,1]的区间之内。对记录值进行极差正规化变换是将各个记录值减去记录值的极小值，再除以记录值的极差，即：

$$x'_{ij}=\frac{x_{ij}-\min(x_{ij})}{\max(x_{ij})-\min(x_{ij})}$$

经过极差正规化处理后的每个观测值都在 0～1 之间。

（4）最小-最大规范化也是一样的数据标准化转换，把所有的数据转化到我们新设定的最小值和最大值的区间内。

而对于序列数据，我们还有两个常用的数据转换方法，分别计算数据的差值和比值。所谓数据差值，是用 $S(t+1)-S(t)$ 的相对改动来代替 $S(t+1)$。而数据比值是采取 $S(t+1)/S(t)$ 的相对改动来代替 $S(t+1)$。

我们再来看下对于非常规数据的数据转换。比如哥伦比亚大学有一个名为 Million Song（一百万首歌），做音乐识别的研究项目，目的是分析世界上各种不同的歌曲。读者可以到下面的网址上访问他们的数据仓库：http://labrosa.ee.columbia.edu/millionsong/。Million Song 项目把每首歌转换成 55 个数据点，包括 Artist Name（音乐家名字）、Audio MD5（整首歌的 Hash 编码）、Title（标题）、Loudness（整首歌的平均分贝）等。在进

行了转换之后，每一首歌就变成了这 55 个数据点的综合表示。

比如下面就是歌手 Rick Astley（里克阿斯特里）的“Never Gonna Give You Up”这首歌在数据转换之后存入数据仓库的部分数值：

```
artist_mbid: db92a151-1ac2-438b-bc43-b82e149ddd50
       the musicbrainz.org ID for this artists is db9...
artist_mbtags: shape = (4,)
       this artist received 4 tags on musicbrainz.org
artist_mbtags_count: shape = (4,)
       raw tag count of the 4 tags this artist received on
       musicbrainz.org
artist_name: Rick Astley
       artist name
artist_playmeid: 1338
       the ID of that artist on the service playme.com
audio_md5: bf53f8113508a466cd2d3fda18b06368
       hash code of the audio used for the analysis by The Echo
       Nest
duration: 211.69587
       duration of the track in seconds
end_of_fade_in: 0.139
       time of the end of the fade in, at the beginning of the
       song, according to The Echo Nest
```

### 5.3.3 数据规约

对于大部分数据集，一般的数据预处理步骤已经足够，而且数据仓库的存在就是为了保存大量的数据。但有些数据挖掘算法比较复杂，即使在少量数据上进行挖掘分析仍然需要很长的时间，在大型数据集全集上做数据分析需要的时间是天文数字，不可能实现。在这样的情况下，在应用数据挖掘技术以前，可能需要采取一个中间的、额外的步骤来做数据规约。这里应用的技术就是数据规约技术。

数据规约技术可以用来得到数据集的规约表示。在规约之后的数据集相比于原数据集要小得多，但仍然接近于保持原数据的完整性，并且结果与规约前结果相同或几乎相同。我们尽量不使用数据规约的原因是数据规约对于原始数据通常都是有损的。采用维数消减模型将多维数据压缩成较少维数的数据，这类方法虽

然对原始数据通常都是有损的，但其结果往往具有更大的实用性。

数据规约的主要问题是是否可在没有牺牲成果质量的前提下，丢弃这些已准备和预处理的数据。能否在适量的时间和空间里检查已准备的数据和已建立的子集。数据规约算法会有以下这些结果：

- 更少的数据，提高挖掘效率；
- 更高的数据挖掘处理精度；
- 简单的数据挖掘处理结果；
- 更少的数据特征。

对数据的描述、特征的挑选、规约或转换是决定数据挖掘方案质量的最重要问题。在现实情况下，数据特征的数量可达到数百项。我们在 Web 挖掘流程中做的数据规约主要是维规约：一方面，如果我们只需要数据仓库中的相关部分数据特征用于分析，就需要进行维规约，以挖掘出可靠的模型；另一方面，高维度所引起的数据超负载（Overload），会使一些数据挖掘算法不实用，唯一的方法也就是进行维规约。

用来做预处理的原始数据集的 3 个主要维度通常以平面文件的形式出现：列（特征）、行（样本）和特征的具体数值。数据规约过程也就是这三个基本操作：删除列、删除行、减少列中的值。

从方法上讲，数据规约中最重要的方法是特征规约方法，数据通常不只为数据挖掘的目的而收集，有时候单独处理相关的特征可以更有效，从应用的角度来讲，我们希望选择只留下与数据挖掘应用相关的数据，以达到用最小的测量和处理量同时获得最好的性能。

数据规约特征集的最主要的标准任务是特征选择，即分析者基于应用领域的知识和挖掘目标，可以选择初始数据集中的一个特征子集，此子集在数据挖掘的性能上可以比得上整个特征集。在特征选择方面，一种可行性较高的技术是基于平均值和方差的比较来选择特征的，但是此方法有一个主要缺点——特征的分布

未知性较强。另外需要注意的是，数据规约需要在减少数据存储空间的同时尽可能保证数据的完整性，获得比原始数据量小得多的数据，并将数据以合乎要求的方式表示。

在做数据规约的过程中，如果我们针对的数据是非常规数据，比如音频视频等多媒体形式，那么我们可以做的数据规约对于数据本身一定是有损失的，但并不影响其数据的主要信息。比如我们可以把高清的 1024×768 的图片格式转化成 640×480 的格式，或者把每秒 30 帧的视频用每秒 10 帧的视频来代替。虽然精度有所降低，但是主要图像和视频信息并没有受到损失。

## 5.4 本章相关资源

- 本章相关参考文献：

[1]（加）Jiawei Han; Micheline Kamber 数据挖掘概念与技术[M]. 范明，孟小峰, 译. 机械工业出版, 2007.

[2]（美）Olivia Parr Rud. 数据挖掘实践[M]. 朱扬勇等, 译. 机械工业出版. 2003.

[3]（美）Richard J.Roiger，（美）Michael W.Geatz. 数据挖掘教程[M]. 翁敬农, 译. 清华大学出版社, 2003.

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

[5] Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar. *Enhancing D ata Analysis w ith Noise R emoval*. IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol. 18, No. 3, pp. 304-319, March 2006.

[6] Levent Ertoz, Michael Steinbach, and Vipin Kumar, Finding Clusters of Different Sizes, Shapes, and Densities

in Noisy, High Dimensional Data (2003). SIAM International Conference on Data Mining (SDM '03).

- 本章相关网址：

http://labrosa.ee.columbia.edu/millionsong/

# 第 6 章
# R 语言和其他数据挖掘工具

在第 4 章中我们介绍了数据挖掘的基本原理和算法，而对于应用来说，更重要的要看我们如何实现这些数据挖掘的算法。在这一章中，我们首先介绍的是在数据挖掘领域最新和使用最广泛的 R 语言，然后再大致讲述在数据挖掘过程中常用的一些其他工具。

## 6.1 R 语言的历史

本节将用较大的篇幅来介绍 R 语言，之所以会对 R 语言专门介绍不仅仅是因为这个语言的热度，更重要的是希望能通过本章的阐述为那些想实施或想学习数据挖掘的读者提供一个有力的辅助工具。R 语言简称 R，在本书中将交替使用这两种说法，并无任何差别。

首先我们需要感谢图 6-1 中两位 R 语言的创始人，左边是 Robert Gentleman 博士，右边是 Ross Ihaka 博士，两位都曾经是新西兰奥克兰大学的统计学教授，而 R 语言也是在这所大学诞生的。由于两位发明人名字的开头字母都是 R，所以这个新的语言被命名为 R 语言。人们对 R 语言的描述有很多，通常的定义为一个能够自由有效地用于统计计算和绘图的语言和环境，又或者是用于统计分析、数据挖掘等各个数据领域应用的软件。在外界对于 R 语言的众多介绍和定义中，笔者比较喜欢 Google 首席经济学家 Hal Varian 对它的描述：R 语言的美妙之处在于你能用

它做各种各样的事情，这归功于 R 上可以免费运用的程序包，所以有了 R，你可以真正站在巨人的肩膀上。


图 6-1　R 语言创始人 Gentleman 和 Ihaka 博士

作者强烈建议想在数据挖掘领域做一些工作的读者马上登录 R 语言的镜像站下载最新版本的 R 来体验它强大的功能，http://cran.r-project.org/，如图 6-2 所示。这里的 CRAN 是 Comprehensive R Archive Network（R 综合典藏网）的英文首字母缩写。它不仅收藏了 R 的执行档下载版、源代码和说明文件，还收录了各种用户撰写的软件包。到目前为止，全球有超过一百个 CRAN 镜像站。

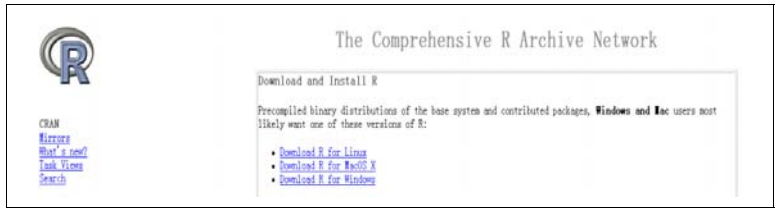



图 6-2　R 语言 CRAN 网站示意图

另一个由美国 Bell Lab（贝尔实验室）开发的，曾在数据分析领域叱咤风云的 S 语言可以算是 R 语言的前身，因为这二者在语法结构上几乎完全一致。贝尔实验室统计研究部的统计学家们曾普遍采用一种需要大量编程的 SCS（Statistical Computing Subroutines）系统进行统计分析，由于统计学家本身对程序语言的掌握程度都有限，而且编写大量程序的工作容易让整个分析变

得混乱，统计学家常常会陷于大量的额外任务而无法专注于数据分析本身。在这种情况下，统计研究部的负责人 John Chambers 发明了 S 语言。由于 S 语言由统计学家发明，所以语言的整个语法结构相对简单，功能上也更适合于做数据分析。因此 S 语言一经诞生，就非常受整个数据分析行业的欢迎，1993 年 S 语言的许可证被 MathSoft 公司买断，S 语言的商业产品 S-PLUS 正式发布，从此 S 语言正式成为商业化产品。由于其独特的优势，S 语言曾与同为商业软件的 SAS 和 SPSS 并称为三大统计分析软件。值得一提的是，1998 年美国计算机协会（ACM）曾将软件系统奖授予 S 语言，而获得这个奖项的其他得奖语言或系统也都赫赫有名，1983 年、1995 年、2002 年的得主分别是 UNIX、WWW 万维网和 Java。可惜的是，S 语言和 SAS、SPSS 一样，都被商业化了，统计学家和程序设计师无法把新的算法实现加入到 S 语言中让大家使用。

自 1992 年 R 语言诞生之后，其创始人不想让该语言步 S 语言的后尘，立志于让更多的人参与到它的开发和改进中。R 语言在继承了 S 语言所有优势的同时，保持完全开源的特性，在 1997 年终于正式加入 GNU 项目。GNU 的英文缩写很有意思，它是“GNU 不是 UNIX 系统”（GNU's Not UNIX）的递归缩写，所谓递归，是因为在 GNU 的解释中又出现了一次 GNU。GNU 计划，是由理查德 • 斯托曼（Richard Stallman）在 1983 年公开发起的，而 GNU 最早的目标是创建一套完全自由的操作系统。之后在 1985 年创立的自由软件基金会（Free Software Foundation）专门为 GNU 计划提供无偿的技术和财政支持。GNU 发布的最重要的系统是 Linus Torvalds 所编写的与 UNIX 兼容的 Linux 操作系统，被称为 GNU/Linux。

在 R 语言加入 GNU 之后，世界各地的数据爱好者积极的加入到 R 语言的开发中，他们聚集在 R 的开源社区，默默的贡献着自己的劳动成果，最先进的统计方法和数据挖掘算法都能很快在 R 上找到相关程序包，更重要的是这些程序包是免费开源的，

你可以在不用花一分钱的情况下去理解和运用这些世界一流的数据科学家们的思想结晶。欧美的软件水平可以持续领先，与这些开源爱好者们的贡献是有莫大关系的。

2009 年 1 月 7 号《纽约时报》记者 Ashlee Vance 在该报科技版刊登了一篇文章 *Data Analysts Captivated by R's Power*，这篇文章详细介绍了 R 语言的历史、特点以及业界对它的评价，文章还将其与分析软件巨头 SAS 进行了比较。虽然从数据分析、数据挖掘本身来说，用什么样的工具并不是最重要的，但是由于现今整个数据挖掘、数据分析都已进入相对成熟的阶段（针对美国而言），从业者越来越依赖于数据挖掘工具，工具性能的优良与否往往能影响项目的最终实施效果。该篇文章当时引起了很大的轰动，SAS 和 R 语言的使用者们聚集在 SAS 和 R 各自的社区内进行着激烈的讨论，对于 R 语言用户来说，这么多年默默使用的小众语言能够得到《纽约时报》这种主流媒体的关注实属不易，大家都在为 R 语言在不久的将来战胜 SAS 信心十足。而对于无论在商业份额还是学术应用都占据领导地位的 SAS 来说，一个开源免费的软件能够引起这么多人的好评甚至威胁到自己的地位也让他们颇为不屑。SAS 公司一位市场总监 Anee Milley 当时说：我们有一些为飞机设计引擎的客户，我很高兴我乘坐他们的飞机时，它的引擎不是用免费软件设计的。在接下来几天里又有几篇文章相继发表，双方阵营的代表人物都发表了各自的看法。无论如何，自此以后 R 语言的知名度大大增强，一些其他分析软件的用户包括 SAS、SPSS 的用户都开始关注它，也正是这场争论让那些非专业人士也了解到分析软件，了解到 R 语言。

而数年之后，R 语言的免费和开放使得它最终取得了压倒性的胜利。今天在互联网上最强大的几家公司包括 Facebook、Google、LinkedIn 和 Microsoft（Bing 搜索引擎）都在用 R 语言做数据分析。而据不完全统计，在 2012 年，全世界约有 200 万数据分析师和学者在使用 R 语言。我看到的另外一个数据表明在数据挖掘领域，R 语言已经被约一半的数据挖掘研究者用来做

数据分析和图形展示。由美国数据挖掘社区 KDNuggets 在 2012 年做的调查，统计在数据挖掘领域常用的所有编程语言中，R 语言从 2011 年的 45.1%上升到 2012 年的 52.5%，增加了 7%。而 Java 和 MATLAB 从 2011 年到 2012 年是下降的。如表 6-1 所示。

表 6-1　2012 年数据挖掘常用语言调查

| What programming/statistics languages you used for analytics / data mining in the past 12 months? [579 voters] % users in 2012 % users in 2011 | |
|---|---|
| R（304 voters in 2012） | 52.5%<br>45.1% |
| Python（209） | 36.1%<br>24.6% |
| SQL（186） | 32.1%<br>32.3% |
| Java（123） | 21.2%<br>24.4% |
| SAS（114） | 19.7%<br>21.2% |
| UNIX shell/awk/sed（85） | 14.7%<br>10.4% |
| C/C++（83） | 14.3%<br>12.8% |
| MATLAB（76） | 13.1%<br>14.6% |
| Perl（52） | 9.0%<br>7.9% |
| Pig, Hive, or other Hadoop-based languages（39） | 6.7%<br>6.1% |

续表

| What programming/statistics languages you used for analytics / data mining in the past 12 months? [579 voters] % users in 2012 % users in 2011 | |
|---|---|
| GNU Octave（34） | 5.9%<br>N/A for 2011 |
| Lisp/Clojure（25） | 4.4%<br>0.7%（Lisp only） |
| Ruby（22） | 3.8%<br>N/A for 2011 |
| Scala（14） | 2.4%<br>N/A for 2011 |
| Julia（2） | 0.3%<br>N/A for 2011 |
| Other（66） | 11.6%<br>12.3% |
| None（4） | 0.7%<br>1.2% |

到今天，R 语言已经成为美国统计学毕业生必修的一门软件。在实际应用领域的流行程度也已经大大增强。如表 6-2 所示，在 2012 年 1 月份，TIOBE 公布的编程语言排行榜，作为统计计算语言的 R 首次进入前 20 名，超过常与之比较的 MATLAB 和 SAS。而在 R 语言之前，MATLAB 是最受统计学和数据挖掘学者青睐的语言。R 的一切功能均是免费的。而由于 R 本身的开源性，再新的统计模型也很快能被开源支持者们实现，所以在众多应用场景上，算法出现之后，在 R 上很快就有新的统计软件包，或者说在 R 上会最快有新算法和研究成果的实现。同样，由于 R 是完全开源，我们可以很快地基于研究者已经开发出的算法编写更适合自己情况的算法。

表 6-2　美国统计专业选用语言排名

| Position Jan 2012 | Position Jan 2011 | Delta in Position | Programming Language | Ratings Jan 2012 | Delta Jan 2011 | Status |
|---|---|---|---|---|---|---|
| 1 | 1 | = | Java | 17.479% | -0.29% | A |
| 2 | 2 | = | C | 16.976% | +1.15% | A |
| 3 | 6 | ↑↑↑ | C# | 8.781% | +2.55% | A |
| 4 | 3 | ↓ | C++ | 8.063% | -0.72% | A |
| 5 | 8 | ↑↑↑ | Objective-C | 6.919% | +3.91% | A |
| 6 | 4 | ↓↓ | PHP | 5.710% | -2.13% | A |
| 7 | 7 | = | (Visual) Basic | 4.531% | -1.34% | A |
| 8 | 5 | ↓↓↓ | Python | 3.218% | -3.05% | A |
| 9 | 9 | = | Perl | 2.773% | -0.08% | A |
| 10 | 11 | ↑ | JavaScript | 2.322% | +0.73% | A |
| 11 | 12 | ↑ | Delphi/Object Pascal | 1.576% | +0.29% | A |
| 12 | 10 | ↓↓ | Ruby | 1.441% | -0.34% | A |
| 13 | 13 | = | Lisp | 1.111% | +0.00% | A |
| 14 | 14 | = | Pascal | 0.798% | -0.12% | A |
| 15 | 17 | ↑↑ | Transact-SQL | 0.772% | +0.01% | A |
| 16 | 24 | ↑↑↑↑↑↑↑ | PL/SQL | 0.709% | +0.15% | A |
| 17 | 20 | ↑↑↑ | Ada | 0.634% | -0.05% | B |
| 18 | 39 | ↑↑↑↑↑↑↑↑↑ | Logo | 0.632% | +0.29% | B |
| 19 | 25 | ↑↑↑↑↑↑ | R | 0.609% | +0.07% | B |
| 20 | 21 | ↑ | Lua | 0.559% | -0.08% | B |

### 6.1.1　R 语言的特点

前面在阐述 R 语言的历史时曾提到，R 是由统计学家设计的，所以 R 的特性处处都打上了发明者把它作为统计工具的烙印，即为了更好地方便没有计算机编程基础又渴望对数据进行分析挖掘的统计学者。所以 R 语言拥有完整体系的数据分析工具，而且为数据分析结果的显示提供了强大的图形功能。

#### 6.1.1.1　R 语言语法

R 语言汇集了面向对象语言（Objected Oriented Language）和数学语言的特点。它的基本语法结构主要有以下这些内容：

- 标准的和基于各种设备的输入/输出；
- 面向对象编程方式和数学编程方式；
- 分布式计算结构；
- 引用程序包；
- 数学和统计学各种函数包括：基本数学函数，模拟和随机数产生函数，基本统计函数和概率分布函数；
- 机器语言学习功能；
- 信号处理功能；
- 统计学建模和测试功能；
- 静态和动态的图形展示。

R 语言的整个语法结构完全来自 S 语言，突出的两个特点是函数式编程（Functional Programming Language）和向量化计算。

函数式编程是将计算机运算视为函数的计算。和指令式编程相比，函数式编程强调函数的计算比指令的执行重要。和过程化编程相比，函数式编程里，函数的计算可随时调用，这点在分析过程中很有用，它能有效地将分析过程通过函数的方式记录下来，下次如果再进行同样的分析过程就不需重复写代码，只需调用相应的函数即可。R 语言镜像站上的各个程序包都是用这种方式写的，用户直接调用程序包里的函数很方便。我们以一个实际数据应用中经常用到的简单例子来展现函数式编程在数据分析中的好处：

```
basic.stats<-function(x,more=F){
       stats<-list()
       wanzheng.x<-x[!is.na(x)]
       stats$n<-length(x)
       stats$na<-stats$n-length(jvzheng.x)
       stats$mean<-mean(jvzheng.x)
       stats$sd<-sd(jvzheng.x)
       stats$median<-median(jvzheng.x)
       if(more) {
       stats$stew<-sum(((jvzheng.x-stats$mean)/stats$sd)^3)
```

```
/length(jvzheng.x)

stats$kur<-sum(((jvzheng.x-stats$mean)/stats$sd)^4)/length
(jvzheng.x)-3}
       unlist(stats)}
```

这个函数用于统计数据集的样本个数、缺失值个数、均值、标准差、中位数，并能根据用户自己的选择来决定是否计算偏度和峰度，最后以列表的形式展现出来。这个函数在应用于对数据进行初步分析时非常有用，当碰到类似分析时，用户只需调用函数即可，而不用重复的一个个的输入命令。

R 语法结构上的另一个特性在于 R 语言大量运用向量化计算而不是 C 语言和 Java 常用的循环运算，这里并不是说 R 语言不支持循环运算，而是因为 R 语言所有运算都在内存中进行，所以使用循环会大量消耗内存使得 R 语言的性能迅速减慢，向量化运算在一定程度上能缓解这方面的不足，这点与 MATLAB 类似，向量化计算常被定义为一种特殊的并行计算，相比于一般程序在同一时间只执行一个操作的方式，它可以在同一时间执行多次操作，通常是对不同的数据执行同样的一个或一批指令，或者说把指令应用于一个数组、向量。

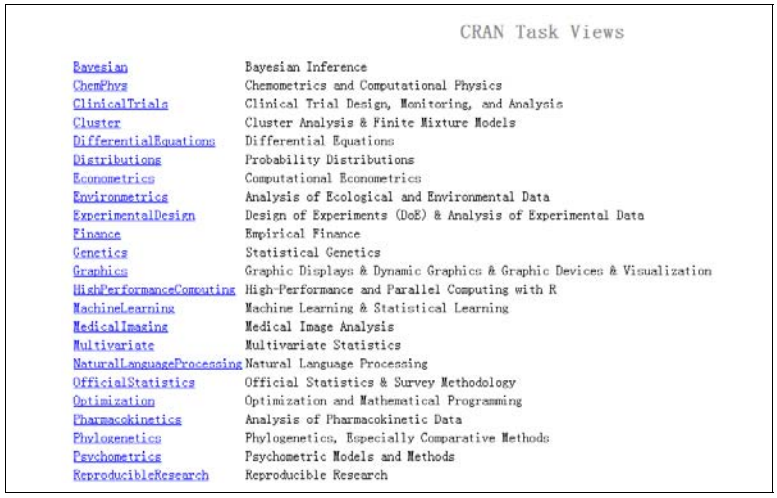
apply 函数就是一个很典型的向量化运算，这个函数的意思是通过将一个函数应用到矩阵或数组中，返回一个向量或数组。例如 apply（x,2, mean）表示的是矩阵 x 按列求均值，正是由于 R 语言中向量化的计算特点使得 R 的代码量会比一般语言实现同一功能要少很多。

#### 6.1.1.2 R 语言程序包

R 社区镜像站上的程序包一直是 R 使用者引以为豪的宝贵财富。截至 2012 年 9 月份 R 镜像站上的程序包已经达到 4045 个，这些包都是由世界各地的 R 爱好者提供并经过严谨筛选后发布的。镜像站上对程序包做出过相应的分类，从分类的结果中我们可以发现这些包的应用范围十分广泛，从化学计量学和计算物理学中的试验分析到金融学里的各种投资模型，从计量经济学到机器学习、自然语言处理，从数据可视化到高性能运算，真可

谓是应有尽有，包罗万象。感兴趣的读者可以自行查看。图 6-3 中列出了在镜像站上程序包的分类。

有一些数据挖掘工程师开始使用 R 语言就是因为在 R 语言中有他们需要的某一个数据挖掘算法实现的程序包，而一旦开始使用 R，就被 R 语言给收编成忠实的用户了。


CRAN Task Views

| | |
|---|---|
| Bayesian | Bayesian Inference |
| ChemPhys | Chemometrics and Computational Physics |
| ClinicalTrials | Clinical Trial Design, Monitoring, and Analysis |
| Cluster | Cluster Analysis & Finite Mixture Models |
| DifferentialEquations | Differential Equations |
| Distributions | Probability Distributions |
| Econometrics | Computational Econometrics |
| Environmetrics | Analysis of Ecological and Environmental Data |
| ExperimentalDesign | Design of Experiments (DoE) & Analysis of Experimental Data |
| Finance | Empirical Finance |
| Genetics | Statistical Genetics |
| Graphics | Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization |
| HighPerformanceComputing | High-Performance and Parallel Computing with R |
| MachineLearning | Machine Learning & Statistical Learning |
| MedicalImaging | Medical Image Analysis |
| Multivariate | Multivariate Statistics |
| NaturalLanguageProcessing | Natural Language Processing |
| OfficialStatistics | Official Statistics & Survey Methodology |
| Optimization | Optimization and Mathematical Programming |
| Pharmacokinetics | Analysis of Pharmacokinetic Data |
| Phylogenetics | Phylogenetics, Especially Comparative Methods |
| Psychometrics | Psychometric Models and Methods |
| ReproducibleResearch | Reproducible Research |

图 6-3　R 语言程序包列表示意图

在第 4 章中提到的各种标准数据挖掘算法都可以在一个或多个 R 语言的程序包中找到可靠的实现方法。我们在 6.1.2 节中会简单介绍一些在 R 语言中常用的数据挖掘包。

正是由于 R 的完全开放性，无数的爱好者为 R 语言做了狂热的贡献。到了今天，几乎所有你熟知和曾经学到的数学模型和统计学模型都可以在 R 语言中找到开源的实现。而如果你对某个程序包中提供的解决方案不满意，你完全可以对它做修改，并提交给 R 社区。

为了解决 R 语言本身的效率问题使它可以处理大数据，许多学者和程序员贡献了一些程序包使得其在大数据上也能用 R 语言分析。美国田纳西大学的 Remote Data Analysis and Visualization Center（远程数据分析和可视化中心）专门有一个项目的目的是为 R 开发大数据的相关程序包。在 2012 年 10 月

他们发布了 pbdMPI，pbdSLAP，pbdBASE 和 pbdDMAT 四个程序包，统称为 pbdR（Programming with Big Data in R，在大数据上用 R 编程），他们声称用 pbdR 程序包组合已经可以在 12 000 台机器上并行运行 R。在 http://r-pbd.org/网站上可以找到更多相关信息。pbdR 程序包现在只是经历了简单的测试，对于它们是否能够真正提升 R 语言的扩展性，我们拭目以待。

#### 6.1.1.3 R 语言接口

R 常常出现在各个领域的应用中，这就需要 R 能够适应各个行业不同数据类型的要求，而且需要能方便的与其他程序语言进行交互。如果想要最有效地用好 R 语言，你必须用好 R 语言的接口。可喜的是 R 能够很好地做到这点。

例如在与数据库连接上，R 提供了两种方式，一种是计算机普遍都有的 ODBC 开放式连接，可以通过调用程序包 RODBC 实现。另一种是使用 DBI 程序包，与需要访问的某一数据库的专门程序一起使用。而通过 API 则可以与新浪微博、Twitter 等社交网络进行连接，R 也能通过程序包 rpy2、rjava 很好地和 python、java 互相调用，进行混合性编程。诸如 SPSS、SAS 等其他统计分析软件的文件也可以在 R 中打开。

#### 6.1.1.4 R 语言数据可视化功能

我们发现 R 在数据可视化上的应用很有潜力，首先它做的图往往很绚丽，这点是其他分析软件无法比拟的，而且可供选择的工具也有很多，从传统的可将图形分隔成格子状的 lattice 包到新近的改变传统绘图模式的 ggplot2 包，R 能用数据画出千姿百态的图形。

我们这里简单介绍一下 ggplot2 包，其思想来自于 *Grammar of Graphics* 一书，而 gg 代表的就是“Grammar of Graphics”。传统方法包括 lattice 包和其他分析软件在作图时都会先去定义具体的图形，然后再在相应的图形框架下去增添相应的参数。ggplot2 的特点却在于它是先定义各种底层组件（如线条、方块），

再合成复杂的图形，这使得用户在使用时能够更方便地根据自己的需要画出各种图形。图 6-4 是用 R 语言作出的图形。

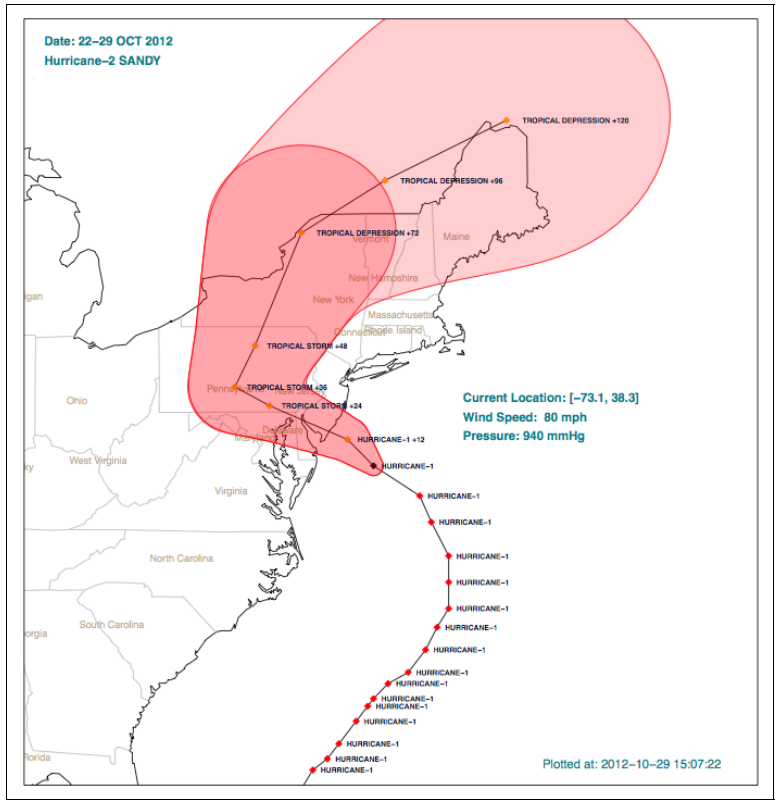



图 6-4　R 语言生成的美国飓风 Sandy 的轨迹图

下面我们用 R 来绘制在网站分析中非常实用的热力图。热力图是以亮点颜色的深浅来显示访客热衷的页面区域，运用热力图，网站分析者可以清楚地看到页面上每一个区域的访客兴趣焦点，这种方法非常直观，使用者无须数据分析和任何页面分析经验就可以对自己网站的经营情况有大体了解。热力图的目标变量往往是点击量之类的能反映点击热度的指标，例如图 6-5 就是一家网店的点击热力图。

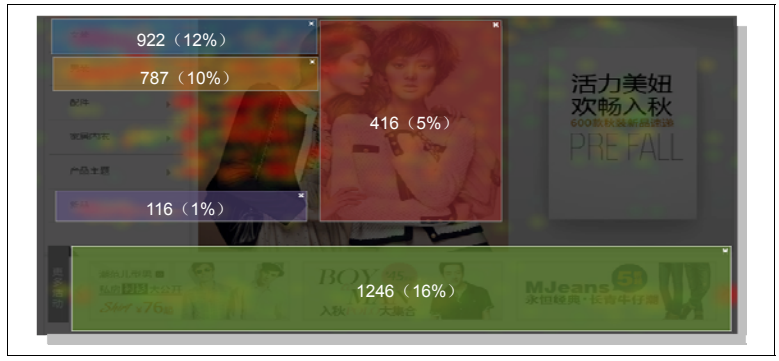



图 6-5　网络热力图

为展示热力图的绘制过程，我们采用点击量作为目标变量，从点击位置和访客区域两个维度进行热力图的绘制。图 6-6 中深色的部分是点击量最多的，而由深至浅的色块表明点击量的由多至少，而空白的色块表示该访客区域没有人点击的对应位置。

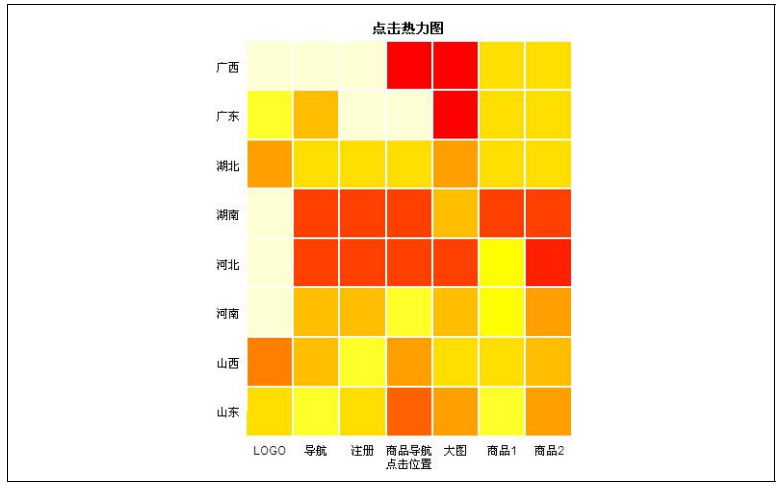



图 6-6　R 语言生成的网络热力图

为了绘制热力图，我们采用了一些函数，主要有以下 3 个，我们先来看第一个函数命令。

```
image(x=1:nrow(click_matrix),y=1:ncol(click_matrix),z=clic
k_matrix,axes=FALSE,xlab="点击位置",ylab="",main="点击热力图")
```

image 函数是 R 语言中常用的一种高级图形函数，image 的

特点是可以按大小将数据用不同的颜色表现出来。函数的具体形式是 image(x,y,z)，其中 x，y 要求是向量，z 要求是矩阵。所以我们把原始数据转化成了矩阵形式 click_matrix，由于图形的横纵轴坐标在后续命令中要单独实现，这里暂时停止坐标轴的绘制，axes=FALSE 就表明了这点。xlab 和 ylab 表示横坐标和纵坐标的标签，这里横坐标标签为“点击位置”，为了使图形更加美观，这里没有对纵坐标标签命名，main 表示图形的标题，位于正上方。

```
axis(1,at=1:nrow(click_matrix),labels=rownames(click_matri
x),col="white",las=1)                                    ,
axis(2,at=1:ncol(click_matrix),labels=colnames(click_matri
x),col="white",las=1)
```

axis 函数用于对 image 已经设置好的图形框架下加入横纵坐标轴，用矩阵的行作为横坐标，行名为横坐标的标签。矩阵的列为纵坐标，列名为纵坐标的标签。坐标本身的颜色设为白色，las=1 表示坐标刻度标签的方向总是水平。

```
abline(h=c(1:ncol(click_matrix))+0.5,v=c(1:nrow(click_matr
ix))+0.5, col="white",lwd=2)
```

为了让图形的层次感加强，我们使用了 abline 函数，abline 函数对整个图形用线进行划分，h，v 表示分别在纵坐标和横坐标画水平线和垂直线。线的颜色是白色，lwd 指线的粗细程度。

### 6.1.2　R 语言和数据挖掘

“工欲善其事，必先利其器”。在数据挖掘的过程中，工具的采用能够让我们的工作更好更快的进行下去，由于数据挖掘本身的复杂性，在选择数据挖掘工具时，要全面考虑多方面的因素，包括分类、聚类关联这些常规算法是否可以实现，数据存取能力以及和其他产品的接口等。R 语言的众多特性表明它有作为数据挖掘工具的潜力，下面讨论一下如何在数据挖掘过程中运用 R 语言。

如果您从学术领域来研究数据挖掘，那 R 语言绝对是非常

不错的工具。它能保证你在完成任务（学术研究的数据量往往不大）的前提下，更方便快速的实现你的算法，你还可以参考 R 语言上那些出色的程序包源代码，这无疑对你的工作有很大的帮助。如果你的任务只是需要从大数据集中挖掘某些规律为决策提供辅助信息，我们也推荐你使用 R 语言，正如前文所述，R 语言在数据可视化的潜力应该更多的挖掘出来，通过 R 语言你能绘出众多能够表达数据含义的图形，而从图形中，能很快地发现你所要的答案，而且在这种情况下，如若数据量大也可以通过抽样的方式解决。

总结一下，在笔者看来，和其他数据挖掘工具相比，R 语言主要有以下的优势：

- 最廉价（免费）；
- 最全的算法；
- 最完美多样的数据展示；
- 最狂热的爱好者社区。

在 R 镜像站上关于数据挖掘的包有很多个，这里只介绍几个常用的包作为读者参考，更多的包读者可以在网上自行查看。

- CORElearn 包原本是 Marko Robnik-Sikonja 和 Petr Savicky 两位教授构造的一个自成一体的数据挖掘系统，在经过发明人同意后将系统整合进 R 语言中，这个程序包集合了多种分类算法和回归模型，例如朴素贝叶斯、随机森林、决策树、回归分析等，除此之外，这个包里包含了大量的特征选择算法和模型评估方法。这样综合性的包在实际应用中往往能为用户提供很大的便利，用户无需自己去寻找相应的包，而只需用一个包就可以完成任务，何乐而不为呢。读者可以在下面这个网站链接上找到很多与这个系统相关的文章，http://lkm.fri.uni-lj.si/rmarko/papers/。
- e1071 包也是综合了众多数据挖掘算法的包，其中被使用的较多的 svm() 函数实现了支持向量机。
- tm 是文本挖掘（text mining）的缩写。文本挖掘在数据挖

掘中一直处于非常重要的地位，文本挖掘的用处越来越多，流行程度也越来越深。在 R 语言中与文本挖掘有关的包主要是 tm 包，tm 包最早发表于 2008 年的 Journal of Statistical Software，包含了语料库建立、词-文档矩阵转化等文本挖掘中常见的数据处理函数，将文本数据转化成常规数据后，就可以与 R 语言中任何数据挖掘工具包结合加以研究了。

- rpart 包很好地实现了分类算法中的深受用户喜爱的算法——决策树算法。
- arules 包提供了有效处理稀疏二元数据的数据结构，而且提供函数用 Apriori 算法和 Eclat 算法来挖掘频繁项集、最大频繁项集、闭频繁项集和关联规则。
- randomforest 包，如名字所示，实现了随机森林算法。
- ROCR 包则是专门用于做模型评估的，可以很方便的绘出 ROC 图。

另外还有许多和数据挖掘相关的程序包可供使用，例如 nnet 包和 RSNNS 包实现了神经网络。其他的包请读者自己参阅 CRAN 网络：

http://cran.r-project.org/web/packages/available_packages_by_name.html

我们在赞扬 R 的无所不能时，也不得不承认 R 的一大缺点，就是在处理大数据问题上 R 的性能确实不是太好。虽然现在也有相应的程序包来试图解决这一问题，但是在实际应用中 R 的性能依然是个不容忽视的短处。也许在不久的未来这一问题能够真正得到突破，但目前，我们只能根据我们的工作来有区别的使用 R。如果你的任务是要建立一个像 Google、百度、Facebook 这些公司所需要的超大规模数据集的数据挖掘系统，那么 R 可能就不是一个好的选择了。并不是说 R 在这些公司毫无用处，在 2012 年由 R 用户组织的主题为“R 与预测分析科学”的讨论会议上，Google 的数据科学家们对 R 在 Google 中的应用做了介绍。Google 主要使用 R 进行数据探索和构建模型原型，它并不

是应用于生产系统，而主要是运行在桌面环境中。Google 使用 R 的流程是：使用其他的工具提取数据，将数据加载到 R 中，使用 R 建模分析，在生产环境中使用效率最高的 C 语言，或者 python 实现结果模型。

从 R 在 Google 的应用中我们看出，虽然 R 在面对大数据时表现的往往并不如人意，但这并不能表示我们可以在数据挖掘领域拒绝 R 语言，我们可以通过抽样的方法来缓解大数据对 R 的压力，也可以用 Google 的做法，在 R 不擅长的地方用其他语言和工具代替，在 R 擅长的地方用 R，这样做无疑是非常明智的。读者可以借鉴。

统计学家 Andrew Gelman 在他的博客中提到他非常希望在使用 R 的过程中能够不为 R 所费的时间和内存担心，而解决方案的其中一种就是“在云中”运行 R。如果能够大规模实现并行处理的方式，对 R 语言在大规模数据上的应用能够如虎添翼。作者期待能够出现一个像 R+Hadoop 这样的好产品。

## 6.2 其他数据挖掘工具

在介绍完 R 语言以后，作为补充，这里再介绍几个在应用领域非常普遍的一些数据挖掘工具供读者参考。

在数据挖掘领域有很多各有特点的数据挖掘软件和工具，其中比较著名的有商用的数据挖掘工具 IBM Intelligent Miner、SAS Enterprise Miner、SPSS Clementine 和开源工具 Weka 等，它们都能够提供常用的挖掘过程和挖掘模式。

其他常用的数据挖掘工具还包括 Statistica、Rapid Miner、KNIME、Salford System、LEVEL5 Quest 、MineSet（SGI）、Partek 、SE-Learn 、WINROSA 和 XmdvTool 等。其中 Statistica 的易用性是相对较好的，而 Rapid Miner 工具近来备受青睐的原因之一是在于它很好地提供 R 语言插件，之二是因为它也是一个开源工具。

### 6.2.1　MATLAB

可以用来做数据分析的工具很多，其中之前多次提到的MATLAB，和 Mathematica、Maple 并称为三大数学软件，因为它们在数值计算方面的功能做得比较出色。

如前面的表 6-2 所示，在 2012 年，约有 14%的数据挖掘从业者在数据挖掘过程中使用了 MATLAB。MATLAB 是由美国 Mathworks 公司发布的主要面对科学计算、统计学研究的交互式计算环境。

和 R 语言不同的地方是，发明 MATLAB 的 Cleve Moler 是数学家，曾在密歇根大学、斯坦福大学和新墨西哥大学任教。可想而知，MATLAB 更注重的是数学运算，而 MATLAB 工具的名字本身也是 MATtrix LABoratary（矩阵实验室）两个英文词各三个字母拼成的。

MATLAB 将数值分析、矩阵计算、科学数据可视化以及非线性动态系统的建模和仿真等诸多强大功能集成在一个易于使用的视窗环境中。MATLAB 最大的特点是高效的数学表达式表现方式、数值计算及符号计算功能。MATLAB 语言是简化版的类 C++语言，因此语法特征与 C++语言相似，但是强化了数学表达式的书写格式。使之更利于数学和统计专业的分析人员使用。图 6-7 为 MATLAB 的界面示意图。

我们用一个简单的例子来看 MATLAB 的矩阵处理功能，比如我们用 MATLAB 来解一个三元一次方程组，求解可以把方程组转化成为矩阵运算的方式进行计算，其中的 $a$ 和 $b$ 都是系数矩阵：

$$ax = b \Rightarrow x = a^{-1}b \Rightarrow x = a \backslash b$$

我们在 MATLAB 里解下面的方程组：

$$\begin{cases} 5x_1 + x_2 + 4x_3 = 6 \\ 3x_1 + 3x_2 + 7x_3 = 6.5 \\ x_1 + 4x_2 + 2x_3 = 10 \end{cases}$$

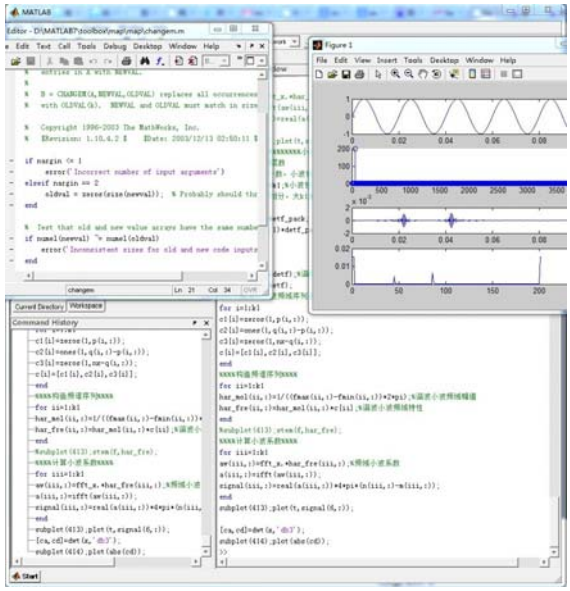


图 6-7　MATLAB 界面示意图

MATLAB 的表达形式及结果如下：

```
>>a=[5 1 4; 3 3 7; 1 4 2];      %建立系数矩阵
>>b=[6; 6.5; 10];               %建立常数项矩阵
>>x = a\b                       %求方程组的解
```

计算结果显示为：

```
x=
    1.2466
    2.5342
   -0.6918
```

也就是说答案为：

$$\begin{cases} x_1 = 1.2446 \\ x_2 = 2.5342 \\ x_3 = -0.6918 \end{cases}$$

矩阵上的其他操作也是非常容易，我们来看几个在线性代数中常用的操作：

- $a'$ 是求 $a$ 的矩阵转置；
- 函数 inv()是对矩阵求逆；

- 函数 det()是求矩阵行列式的值；
- 函数 diag()可以用来抽取对角线上的元素或者重新构建对角矩阵；
- 函数 sqrt()是用来求矩阵的开方；
- 可以使用符号（“.+”、“.-”、“.*”、“./”、“.^”）让两个矩阵中的每一个元素直接进行加减乘除和乘方的操作。

除了刚才演示的命令行操作以外，MATLAB 也可以实现在 C 语言或 FORTRAN 语言上的大部分功能，包括 Windows GUI（图形用户界面）的设计。虽然略逊于 R 语言，利用 MATLAB 的图形功能也可以轻易地绘制二维和三维图形。

自 1984 年 MATLAB 1.0 发布至今，已经有了 10 多个稳定版本，而最新的是 2012 年 9 月发布的 MATLAB 8.0。

在 R 语言出现以前，统计学和数学专业的大学经常用 MATLAB 来做线性代数、信号处理、数值分析等课程的教学工具，让同学们在 MATLAB 上应用数学和统计方法分析给定的数据集。

## 6.2.2　其他商用数据挖掘工具

其他商用数据挖掘工具还有很多，但是最富盛名的当属 IBM 的 SPSS Modeler。

### 6.2.2.1　SPSS Modeler

SPSS Modeler，或者说是 IBM SPSS Modeler，支持整个数据挖掘流程，包括从数据获取、转化、建模、评估到最终部署的全部过程。值得一提的是在行业中应用较多的 CRISP-DM（跨行业数据挖掘标准流程）是由 SPSS 公司提出，而且这也是 SPSS Modeler 的设计理念。SPSS Modeler 的可视化数据挖掘使得“思路”分析成为可能，即将精力集中在要解决的问题本身，而不是局限于完成一些技术性工作。提供了多种图形化技术，有助于理解数据间的关键性联系，指导用户以最便捷的途径找到问题的最终解决办法。

SPSS Modeler 就是原来的 SPSS Clementine，是由 SPSS 公司研制的商业数据挖掘平台。SPSS 公司在市场研究和市场调查领域有很高的市场占有率，在全球约有数十万家产品用户，分布于通信、银行金融、保险证券、制造业、市场调研、政府税务、教育科研、医疗卫生、化工行业、零售业、电子商务等多个领域和行业，据官方宣传全球 500 强中约有 80%的公司使用 SPSS 工具。SPSS Modeler 作为 IBM 重点推出的数据挖掘工具，包含了复杂的统计方法和机器学习技术。在 2009 年 SPSS 公司被 IBM 全资收购，而 SPSS Modeler 的名字就是在那个时候被更改的。2012 年 Rex Analytics 分析公司做的调研表明 IBM 的 SPSS Modeler 是目前被使用频率最高的数据挖掘工具。

图 6-8 为 SPSS Modeler 的用户界面。

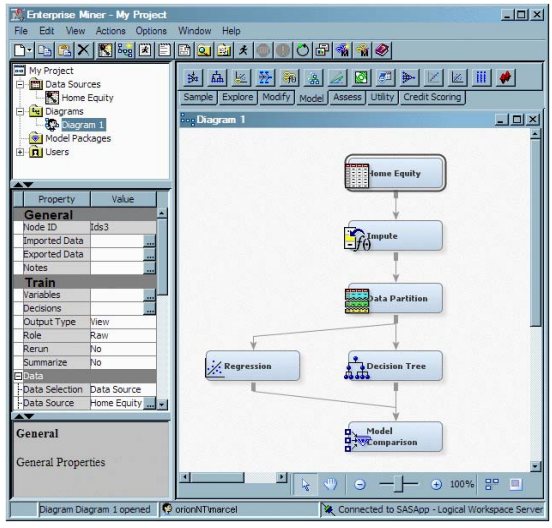



图 6-8　SPSS Modeler 用户界面示意图

### 6.2.2.2　SAS Enterprise Miner

SAS Enterprise Miner，简称为 SAS EM，是一种在中国的大型企业及政府机关中得到广泛应用的数据挖掘工具，比较典型的应用包括上海宝钢配矿系统应用和铁路部门在客运研究中的应

用，还有上海通用汽车，中国招商银行等也都是 SAS 的客户。

这个数据挖掘工具，SAS Enterprise Miner 是数据分析领域另一巨头 SAS 公司的数据挖掘产品，SAS（Statistical Analysis System）本身是美国北卡罗来纳州州立大学在 1966 年开发的统计分析软件。SAS 公司成立于 1976 年，总部位于美国北卡罗来那州的凯瑞（Cary），是全球最大的私有软件公司。SAS Enterprise Miner 是一种通用的数据挖掘工具，作为行业巨头，SAS 自然不甘心使用竞争对手倡导的数据挖掘过程，所以 SAS Enterprise Miner 采用的是 SAS 自己提出的 SEMMA 过程进行数据挖掘。

作为 SAS 公司力推的数据挖掘产品，SAS Enterprise Miner 在经过几代的改进后日趋成熟。可以与 SAS 数据仓库和 OLAP 集成，实现从提出数据、抓住数据到得到解答的“端到端”知识发现。与 SPSS Modeler 类似，SAS Enterprise Miner 也可利用具有图形化的模块将数据挖掘单元组成处理流程图，并依此来组织数据挖掘过程。这一过程在任何时候均可根据具体情况进行修改、更新、存储，以便此后重新调出来使用。这种图形化界面的优势是可引导那些数理统计经验不多的用户，而对于有经验的专家，SAS Enterprise Miner 也提供了大量的选项，可让有经验的人士进行精细的调整分析处理。

图 6-9 为 SAS Enterprise Miner 的操作界面。

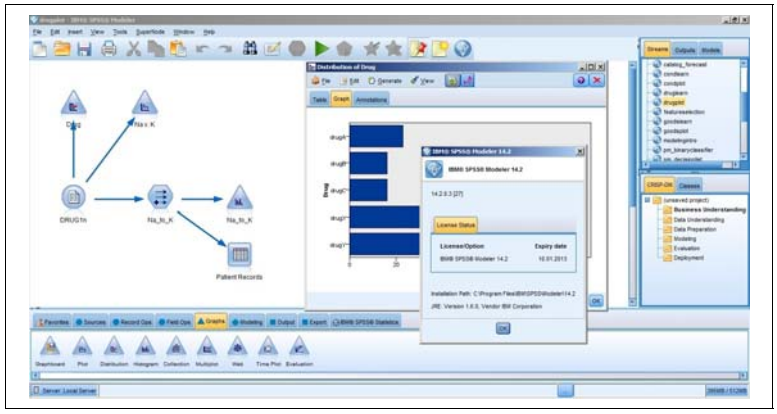


图 6-9　SAS Enterprise Miner 界面示意图

#### 6.2.2.3 IBM Intelligent Miner

由美国 IBM 公司开发的数据挖掘软件 Intelligent Miner，是一个分别面向数据库和文本信息进行数据挖掘的软件系列。我们平时说的 Intelligent Miner 主要是指 Intelligent Miner for Data，是用来挖掘包含在数据库、数据仓库和数据中心中的隐含信息，帮助用户利用传统数据库或普通文件中的结构化数据进行数据挖掘。它在市场分析、诈骗行为监测及客户联系管理等方面都有一定程度的应用。

Intelligent Miner 的分支产品 Intelligent Miner for Text 允许企业从文本信息进行数据挖掘，文本数据源可以是文本文件、Web 页面、电子邮件、Lotus Notes 数据库等。在 IBM 收购 SPSS 之前是 IBM 做商业智能的主打产品，但即使是收购之后，Intelligent Miner 还是依赖之前积累的老客户和与 DB2 的紧密耦合在市场上占领一席之地。美国花旗银行和最大的健康保险公司埃特纳（Aetna Healthcare）都是 IBM Intelligent Miner 的忠实客户。

图 6-10 是 IBM Intelligent Miner 的用户界面。

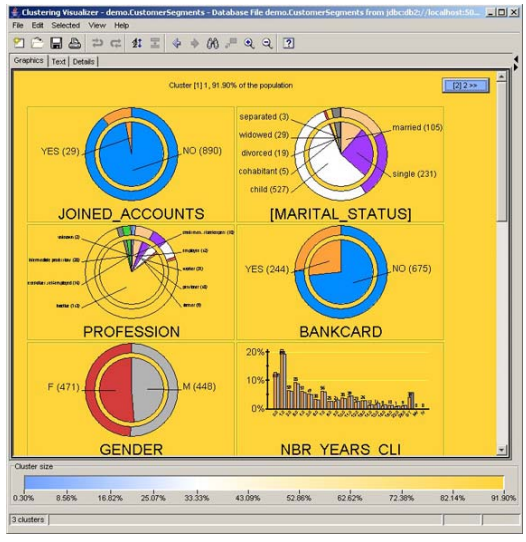



图 6-10　IBM Intelligent Miner 界面示意图

### 6.2.3　开源数据挖掘工具 Weka

对于数据挖掘行业的研究人员来说，开源的数据挖掘工具最近受到的关注比较多：一方面是由于商用数据挖掘系统较为昂贵；另一方面是因为数据挖掘领域新的技术和挖掘过程出现得较快，而开源工具因为众多从业人员的兴趣所致跟风的速度可以令他们满意。开源的数据挖掘工具的代表工具主要有 Rapid Miner、KNIME 和 Weka。

在数据挖掘开源领域里，Weka 属于知名度比较高的一款软件，它的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是用 Java 语言开发的开源的数据挖掘软件。Weka 也是使用的最广泛的机器学习开源工具之一。我们可在其官方网站下载其软件和源代码：http://www.cs.waikato.ac.nz/ml/weka/。Weka 是由新西兰怀卡托 Weka 小组在 1997 年研制的软件。2005 年 8 月在第 11 届 ACM SIGKDD 国际会议上，怀卡托大学的 Weka 小组荣获了数据挖掘和知识探索领域的最高服务奖。

Weka 软件有四个组成部分，如图 6-11 所示。

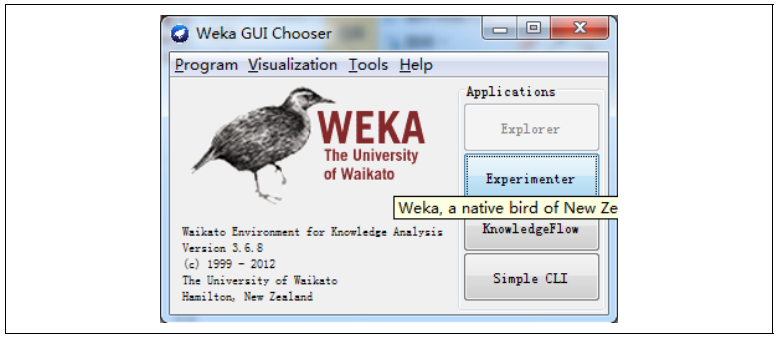



图 6-11　Weka 界面示意图

（1）Explorer：在该环境中，我们可以实现各种数据挖掘算法，并提供可视化结果。

（2）Experimenter：用来做算法实验的环境。在该环境中，用户可以创建、比较、修改和分析算法。

（3）KnowledgeFlow：在“知识流”的环境中，用户可以把不同组件按照一定顺序连接起来，组成知识流用以处理和分析数据。

（4）SmpleCLI：简单的命令行界面。

在 Weka 中可以使用的数据挖掘算法主要有三类：分类算法、聚类算法和关联算法，正好对应我们在本书中使用最多的数据挖掘算法。

Weka 在 R 语言中时，我们可以使用 Rweka 程序包调用 Weka 中的所有算法。

而在开源社区中有一种呼声，要求有人可以用 Hadoop 作为基础把 Weka 重新写一遍，使得机器学习工具可以在大数据平台上得到充分的应用。

## 6.3 数据挖掘和云

本章前面两节介绍的内容都是基于企业需要构建自己的数据挖掘平台，而其实亚马逊、Google 都提供了基于云的数据挖掘，也就是 Data Mining in the Cloud（云中的数据挖掘）。

亚马逊 Amazon 的 EMR（Elastic MapReduce，弹性 MapReduce）框架是基于 Hadoop 的互联网服务。您可以在 http://aws.amazon.com/elasticmapreduce/上查看该服务的细节。EMR 运行在 Amazon 的 EC2（Elastic Computing Cloud，弹性计算云）和 S3（Simple Storage Service，简单存储服务）基础平台之上。

图 6-12 是 Amazon 的 EMR 过程示意图。

图中数字标记的过程解释如下：

（1）把需要处理的数据采用 Amazon S3 API 或其他方式上传到 Amazon S3，同时上传相应的 Map 和 Reduce 可执行程序，然后给 Amazon EMR 发送一个请求开始工作流。

（2）Amazon EMR 启动一个 Hadoop 集群，并在每个节点上启动 Hadoop。

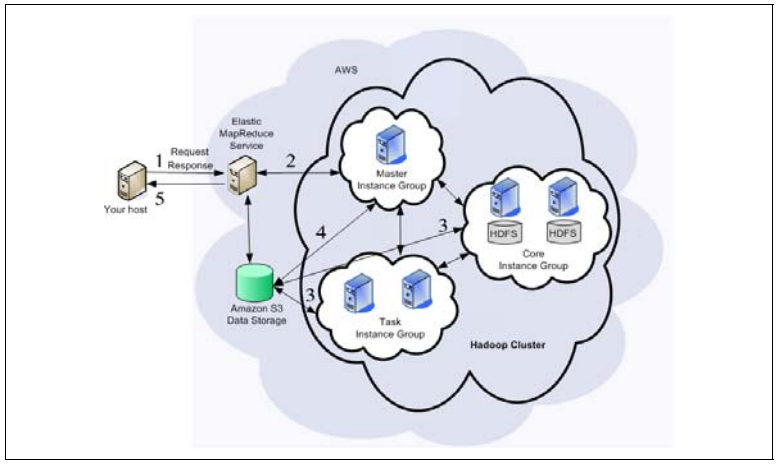



图 6-12　Amazon 的 EMR 过程示意图

（3）Hadoop 开始执行工作流，把数据从 Amazon S3 分配到各个核心和任务节点，而 Map 任务动态地把数据分解并加载到各个子节点上。

（4）Hadoop 执行数据处理任务然后把结果数据从集群上传到 Amazon S3。

（5）工作流完成，您可以从 Amazon S3 上取回数据。在 Amazon EMR 上的 HBase 和 S3 存储相关联，使得我们很容易把放在 HBase 上的数据备份进 S3，或是把数据从 S3 中恢复出来。

Amazon EMR 上实现了第 3 章的图 3-9 所示 Hadoop 生态系统中的其他部分，包括 HBase、Hive 和 Pig 等，所以我们可以在 EMR 上运行更加复杂的数据分析和挖掘工作。

您也可以尝试 Google 为数据挖掘提供的 Prediction API 接口。Google 的 Prediction API 是谷歌提供的机器学习工具云服务。您可以用 Google API 做客户情感分析、垃圾邮件监测、文档分类、过滤 RSS feeds、用户评论或反馈和电子商务推荐系统等。

https://developers.google.com/prediction/是 Google Prediction API 的官方网站。Google Prediction API 允许预览用户通过 RESTful Web 服务采用 Google 的高级机器学习算法构建更智能的应用。

REST（Representational State Transfer，表现状态转移）是Roy Fielding 博士在 2000 年写作的博士论文中提出来的一种软件架构风格，在此风格中，每个资源是由全球唯一的 URI 来指定，资源本身和其表现方式是完全独立的。当一个用户拿到资源的表现方式时，他有足够的信息可以修改或者删除服务器上相应的资源而且每条消息都包含了足够的信息可以描述消息的处理。而 RESTful Web 服务（又称为 RESTful Web API）是一个使用 HTTP 并遵循 REST 原则的 Web 服务，有以下四个方面：

- Web 服务有一个基础的全球唯一的 URI；
- Web 服务支持一种互联网媒体类型，比如 XML；
- HTTP 方式支持的一系列操作，比如 GET、PUT、POST 和 DELETE。
- API 必须是超文本驱动的。

Prediction API 从数种可用的机器学习算法中选取最佳算法，支持结构化数据和无结构的文本作为输入，可以输出数百种离散类或连续值。Prediction API 可以与 Google App Engine、网络等众多平台整合。

## 6.4 本章相关资源

- 本章相关参考文献：

[1] William N Venables and David M Smith，An Introduction to R，Second Edition，Network Theory Ltd, 2009.

[2] Robert I. Kabacoff. R in *Action:Data A nalysis an d Graphics with* R. August, 2011, ISBN:9781935182399.

[3] Leland Wilkinson, D. Wills, D. Rope et al. *The Grammar of Graphics (Statistics and Computing*. Springer, 2005.

[4] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D Dissertation,

University of California, Irvine

[5] Mark Hall，Eibe Frank et al. *The WEKA data mi ning software: a n up date*. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18.

- 本章相关网址：

[1] http://www.r-project.org/

[2] http://www.statmethods.net/

[3] http://cran.r-project.org/

[4] http://lkm.fri.uni-lj.si/rmarko/papers/

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://cos.name/

[7] http://r-ke.info/

[8] http://www.rdatamining.com/

[9] http://www.mathworks.cn/products/matlab/

[10] http://aws.amazon.com/elasticmapreduce/

[11] https://developers.google.com/prediction/

[12] http://andrewgelman.com/

[13] http://r-pbd.org/

# 第 7 章
# 互联网上的日志分析

通过网站的日志分析，我们可以清楚地得知用户在什么 IP、什么时间、用什么操作系统和什么浏览器访问了网站的哪个页面，是否访问成功，在网站上面点击了哪个链接进入了网站的另一个什么页面，以及离开网站去往哪里。

而在网站上可以有的信息除了日志之外，还有各种访客的会话信息，网上商城访客在网站中购物交易或者填写的各种资料信息和评论信息，访客的搜索查询信息等，不过这些信息都是我们可以直接收集的，在本章中就不做讨论了。

如果将多个网站上的日志分析综合起来，我们能够得到更加有价值的数据。后面我们在第 8、9、10 章节中讲述的数据挖掘具体案例，包括网站分析、网站联盟作弊分析等都需要以日志文件的分析作为基础。

如果您的网站相对比较简单，那么我们不一定需要使用本章中的日志分析，可以直接采用 Google 分析（Google Analytics）等网站分析工具。不过本书的对象是大数据，对于普通网站的小规模的网站分析本来就不在本书的讨论范围之内。对于海量数据上的分析，目前我还没有看到第三方服务公司有很好的类似 Google 分析的产品，我们必须要自力更生，从网站日志入手。关于 Google 分析这一工具的局限性，我们在第 9 章互联网广告分析中会提及。

# 7.1 网站日志简介

各家互联网公司使用的 Web 服务器五花八门，常用的有基于 Apache、Ngnix 和 IIS 的服务器等，通常都保存了对 Web 页面的每一次访问的 Web log 项（日志项）。它能够忠实记录所有访问该 Web 服务器的数据流信息。我们可以从服务器管理员这里随时拿到所有的文件。

如果您的服务器是托管在虚拟机上，那么一般虚拟主机提供商不一定会开通网站日志统计功能。这时需要您自己到服务器管理后台开通这个日志统计功能，如果还是不行，那么有可能要专门申请开通。

根据网站的访问需要和总体流量情况，Web 服务器群组按照一定的时间间隔和存储空间来开启新的日志文件，每个文件中记录的可以是一天、一个小时或者一分钟的日志数据。为了保证性能和传输方便，我们通常控制 Web 文件的大小最多不超过 1GB。

为了有效地分析和处理这些日志文件，我们可以采用数据挖掘技术。对于简单的网站结构，可能分析处理一个 Web 的日志文件就可以了，但是通常对于一些比较大的网站来说，往往是好几十个甚至上百台 Web 服务器组成一个集群来对外服务的，在分析这些网站的日志文件时，就可能需要采取分布式的 Web 数据挖掘技术。图 7-1 就是一个服务器群组中一台服务器某一天的部分日志文件列表，其中文件的编号和服务器名称，与在网络拓扑架构中的位置以及时间相关。

对于这种每天都会产生这么大量的数据日志文件的系统，数据挖掘可以起到很好的效果。

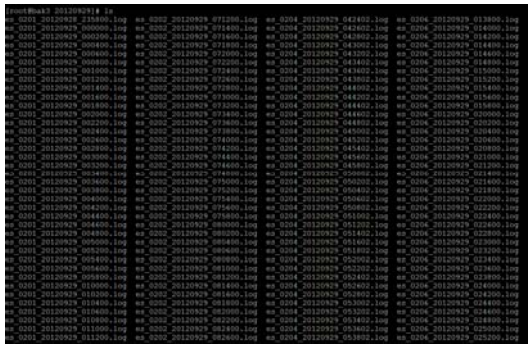


图 7-1　日志文件列表示意图

日志文件的格式通常都是很直观的。例如 Windows 的 IIS log（IIS，Internet Information Service，互联网信息服务），它可支持的 Web 日志格式有 Microsoft IIS 日志文件格式、NCSA 公用日志文件格式和 W3C 扩展日志文件格式等。最早的 Microsoft IIS 日志文件格式信息记录是以逗号分隔的 ASCII 文本文件，而且数据固定，不能自定义。从 IIS5.0 起，W3C 扩展日志文件格式成为默认的日志文件格式。

在 W3C 扩展日志文件格式中，系统管理员可以根据客户的不同需要，来调整在日志文件记录哪些内容和信息。例如 IIS5.0 的 W3C 扩展日志文件格式中，除了默认的常用属性，包括所请求的 URI 资源、客户端 IP 地址和时间戳等，还有多达 19 项可以选择记录的扩展属性。表 7-1 列出了 W3C 扩展日志文件格式中可选的属性。

表 7-1　W3C 扩展日志文件格式常用属性说明表

| 字 段 名 | 描　　述 |
| --- | --- |
| 客户端 IP 地址 | 访问服务器的任何客户端的 IP 地址 |
| 用户名称 | 访问服务器的用户名称 |
| 服务名 | 在客户机上运行的 Internet 服务 |
| 服务器名称 | 生成日志项的服务器名称 |
| 服务器 IP | 生成日志项的服务器的 IP 地址 |
| 服务器端口 | 客户端连接到的端口号 |

续表

| 字 段 名 | 描　　述 |
| --- | --- |
| 方法 | 客户端试图执行的操作（例如，GET 命令） |
| ServiceStatus | 简单邮件传输协议（SMTP）回复代码 |
| URI 查询 | 客户端试图执行的查询（如果有）。在日志中记录了客户端搜索以进行匹配的一个或多个搜索字符串 |
| 协议状态 | 以 HTTP 术语表示的操作的状态 |
| 发送的字节数 | 服务器发送的字节数 |
| 接收的字节数 | 服务器接收的字节数 |
| 所用时间 | 操作所需的时间长短 |
| 协议版本 | 客户端使用的协议（HTTP，FTP）版本。对于 HTTP，是 HTTP 1.0 或 HTTP 1.1 |
| 主机 | 计算机名 |
| 用户代理 | 在客户端使用的浏览器 |
| Cookie | 发送或接收的 Cookie 的内容（如果有） |
| 引用站点 | 将用户指向当前站点的站点 |

在个人电脑市场上，微软的 Windows 占据了绝对的优势，但是在 Web 服务器市场中，Windows 就明显不敌 UNIX 和 Linux。我们下面来看一下主要运行在 Linux 平台上的两种 Web 服务器上的日志文件是怎样的。

Apache 是一种开源的 Web 服务器，主要是在 UNIX 和 Linux 平台上的。新兴的互联网 Web 2.0 公司很多的基本架构都基于 LAMP（Linux、Apache、MySQL 和 PHP 这四种 Web 主要技术的缩写）技术或者 LNMP（Linux、Nginx、MySQL 和 PHP 这四种 Web 主要技术的缩写）技术，其中的 A 指的就是 Apache，而 N 指的是我们后面要提到的 Nginx。

Apache 服务器的 HTTP 服务器 log 有以下的格式：

```
%h %l %u %t "%r" %>s %b "%{Referer}i" "%{User-agent}i"
```

上面参数的解释如下：

```
%h   =  客户访问端的 IP 地址
%l   =  RFC 1413 格式的客户标记
```

```
%u   =  请求文档的客户标识 userid
%t   =  服务器处理完请求的时间
%r   =  用户的请求
%>s  =  服务器返回给客户的服务器代码
%b   =  返回给客户的数据尺寸
```

我们来看几个 Apache 日志文件的实例。下面是某个 Apache 日志文件中的记录：

```
66.249.65.107  -  -  [08/Oct/2011:04:54:20  -0400]  "GET
/support.html HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible;
Googlebot/2.1; +http://www.google.com/bot.html)"
```

在这行记录中，66.249.65.107 是访客的 IP 地址。

知道了用户的 IP，就可以通过查询来得知用户是来自哪个国家、哪个省份、哪个城市的。可以通过 http://www.ipchecking.com/?ip=66.249.65.107&check=Lookup 来查看访客的 IP 地址来源。查询得知，该 IP 地址是在美国加州的 Mountain View 市，地址是 1600 Amphitheatre Parkway，而且该 IP 是谷歌拥有的。如图 7-2 所示。


图 7-2　IP 查询示意图

这条记录的访问时间是 2011 年 10 月 8 日凌晨 4 点 54 分 20 秒。

在记录中的时间数据之后是“GET”，这是服务器的处理动作，一共只有两种：GET 和 POST。在网站日志中绝大部分都是

GET，只有在进行 CGI 处理时才会出现 POST，否则服务器在响应时都是 GET，也就表示用户从服务器上获取了页面或者其他的文件。GET 后面的“/”代表的是用户访问的页面。“/support.html”表示访问的是该网站的技术支持页面。

HTTP/1.1——这个代表用户访问该页面时，是通过 HTTP1.1 协议进行传输的，也就是超文本传输 1.1 版本协议。除了个别提供 FTP 下载的网站之外，普通用户基本都是通过 HTTP 协议来进行访问的。

200 代表的是用户访问页面的时候返回的状态码。服务器返回代码如果不是 200，就表示有错误发生。下面列出常用的服务器错误代码：

```
200——OK
206——Partial Content，部分内容
301——Moved Permanently，用户所访问的某个页面 url 已经做了 301 重定向（永久性）处理
302——Found，内容被暂时重定向，已经找到
304——Not Modified，未修改，采用缓存（cache）拷贝
401——Unauthorised（password required），需要密码
403——Forbidden，不可访问
404——Not Found，没有内容
408——Request Timeout，请求超时
500——Server Error，通常是服务器发生错误，比如在维护或者下线了
11179 说明传输了 11179 字节。
Googlebot/2.1; +http://www.google.com/bot.html 这说明来访的是 Google 的爬虫（spider）。Google 的爬虫程序会把我们的网页抓下来整理到谷歌的搜索引擎中。Googlebot 是谷歌爬虫的标记。
```

我们再看下面的一个实例，在 Apache 服务器的日志文件中有以下两条记录：

```
202.109.65. ××× - - [08/Oct/2011:11:17:55 -0400] "GET / HTTP/1.1" 200 10801 "http://www.google.com/search?q=blah+blah&ie=utf-8&oe=utf-8 &aq=t&rls=org.mozilla:en-US:official &client=firefox-a" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"

202.109.65. ××× - - [08/Oct/2011:11:17:55 -0400] "GET /style.css HTTP/1.1" 200 3225 "http://www.blahblah.net/" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"
```

这两条记录的 IP 地址都是 202.109.65.×××。出于对用户个人的隐私考虑，我们把 IP 地址最后的三位隐掉了。

根据 IP 地址查询 http://www.ipchecking.com/?ip=202.109.65.107&check=Lookup，结果如图 7-3 所示。访客来自上海市中心。不同的网站其用户群会有比较明显的区别，比如某个宁波的地方性网站的大多数访客肯定是来自宁波的，而有的网站其用户没有什么明显的地域区别。用户 IP 配合其他信息可以让我们更加有效的分析网站的用户体验是不是做得够好。

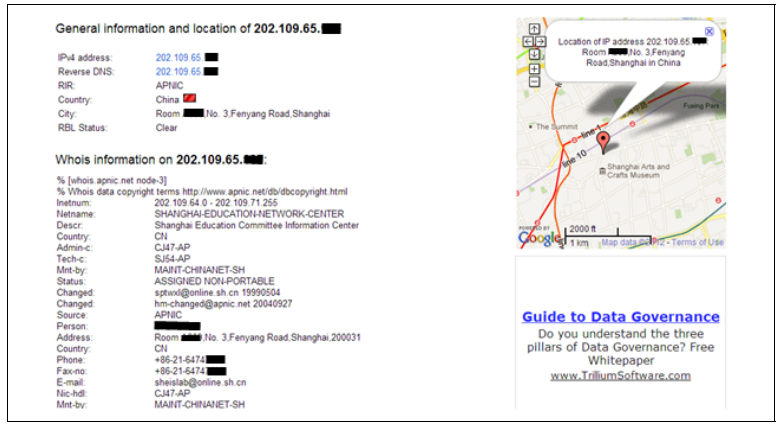



图 7-3　IP 查询示意图之二

这两条记录的访问时间都是 2011 年 10 月 8 日上午 11 点 17 分 55 秒。

服务器的处理动作是 GET，而 HTTP/1.1 表明访问是通过 HTTP1.1 协议进行传输的。

200 代表的是用户访问页面的时候返回的状态码是成功的。

10801 和 3225 代表的是传输的字节分别是 10 801 和 3 225 字节。

“Firefox/2.0.0.7”说明该用户使用的是 Firefox 浏览器。

这两条从 202.109.65.×××的访问是在谷歌上搜索“blah blah”，然后从搜索结果中点击并查看了我们 http://www.blahblah.net 的主页。

如果我们发现在日志中存在大量的 404 错误代码，那么就说明可能有个互联网链接断了，或者有人指向一个已经不存在的网页。如果这种情况发生，那么在日志文件中会有大量类似下面的内容：

```
173.173.108.××× - - [20/May/2012:13:34:45 -0700] "GET
/index.php?-dsafe_mode%3d0ff+-ddisable_functions%3dNULL+-d
allow_url_fopen%3dOn+-dallow_url_include%3dOn+-dauto_prepe
nd_file%3dhttp%3A%2F%2F81.17.24.82%2Finfo3.txt HTTP/1.1" 404
531 "-" "Mozilla/4.0 (compatible; MSIE 6.0b; Windows NT
5.0; .NET CLR 1.0.2914)"
```

另一类使用比较频繁的 Web 服务器是 Nginx 服务器。Nginx 是开源的高性能 HTTP 服务器，和 Apache 服务器相比，在处理大规模数据的情况下性能更加好。Hulu、Pinterest 和 Zynga 等互联网新贵都采用的是 Nginx 服务器。

和其他 Web 服务器不同，Nginx 的日志文件中把普通访问信息和错误信息分开，所以日志文件会分为两类：错误日志（Error Log）和访问日志（Access Log）。

错误日志可以在 Nginx 的终端上用下面的命令行看到：

```
sudo less /var/logs/error.log
```

之所以加上“sudo”是因为要有管理员权限才能访问错误日志。当然，安全的作法是设置合适的权限，因为能够有权限查看错误日志文件的人不一定是系统管理员。不过系统安全不是本书讨论的内容范围。

错误日志中的每一条都是类似这样的。我们来看 Nginx 服务器上一条错误日志信息：

```
2010/08/23 15:25:35 [error] 19997#0: *1 open() "/var/www/
nginx-default/phpmy-admin/scripts/setup.php" failed (2: No
such file or directory), client: 80.154.42.54, server:
localhost, request: "GET /phpmy-admin/scripts/setup.php
HTTP/1.1", host: "www.example.com"
```

Nginx 上的错误日志文件信息内容和 Apache 服务器的日志内容类似，最开始的是时间值：“2010/08/23 15:25:35”。

在时间值之后的[error]表明这条记录记录的是一条错误信息。以上这条信息“(2: No such file or directory)”是表明在 Nginx 服务器上发生了“找不到文件或目录”错误。

访问的 IP 地址是 80.154.42.54。

“HTTP/1.1”表示这个代表用户访问该页面时，是通过 HTTP1.1 协议进行传输的，也就是超文本传输 1.1 版本协议。

“www.example.com”是这次用户访问的网站。

在 Nginx 上的错误信息种类很多，如果我们只关心最严重的错误，我们可以在 Nginx.conf 中加上下面这条，那么非严重错误都会不再记录在错误日志文件中了。

```
error_log  logs/error.log  crit;
```

再看这条错误信息：

```
 "accept() failed (53: Software caused connection abort) while
accepting new connection on 0.0.0.0:80"
```

这条信息表示的是用户企图打开一个网页，而在网页尚没有完全成功打开时就取消了。这条信息不是错误信息，但是也给我们提供了有用的资料。如果这样的信息比较多，说明我们的网页打开速度可能太慢，用户经常会失去等待的耐心。

在 Nginx 服务器上，访问日志也是存放在和错误日志的同一位置。可以用下面的命令行看到访问日志的最后 100 行：

```
    sudo tail -n 100 /var/log/nginx/access.log
```

Nginx 上的日志文件可以以各种形式来组织，而且可选用的属性也是多样的，比如下面这一条就在 Nginx 上定义了一个专门的日志类型“speciallog”。

```
log_format   speciallog   '$remote_addr   -   $remote_user
[$time_local]  '
                '"$request" $status $body_bytes_sent '
                '"$http_referer" "$http_user_agent"';
```

为了做好 7.2 节中的网站日志处理，根据网站的具体分析需要来调整一下 Web 服务器的记录文件的记录字段是十分必要的，这样既可以将不要的数据去掉，也可以增加一些可以帮助我

们后面用来分析“可以行为”可能需要的字段，使原始数据的采集更加便于后续对日志信息的处理工作。

重新定义了日志之后，我们来看一下新的日志。下面分别是我们新定义的日志文件格式和一条按照这一格式产生的数据访问信息：

```
$remote_addr - $remote_user [$time_local] "$request" $status
$body_bytes_sent "$http_referer" "$http_user_agent"
123.45.678.90 - - [23/Aug/2010:03:50:59 +0000] "POST
/test/wp-admin/admin-ajax.php HTTP/1.1" 200 2
"http://www.blahblah.net/test/wp-admin/post-new.php"
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US)
AppleWebKit/534.3 (KHTML, like Gecko) Chrome/6.0.472.25
Safari/534.3"
```

我们一一对应来看相关的信息：

```
$remote_addr 对应的 IP 地址是 123.45.678.90
$remote_user 对应的对象为空
$time_local  对应的对象是 23/Aug/2010:03:50:59 +0000
$request     对应的命令是 POST /test/wp-admin/admin-ajax.php
HTTP/1.1
$status      对应的是 200
$body_bytes_sent 对应的是 2
$http_referer    对应的是 http://www.blahblah.net/test/wp-
admin/post-new.php
$http_user_agent 对应的是 Mozilla/5.0 (Macintosh; U; Intel Mac
OS X 10_6_4; en-US) AppleWebKit/534.3 (KHTML, like Gecko)
Chrome/6.0.472.25 Safari/534.3
```

在访问日志中还有一项很重要的数据：Cookie。Cookie 的英文原意是小饼干，但在互联网上指的是网站为了辨别用户身份而储存在用户本地终端浏览器上的一类数据，包含每次用户访问站点时 Web 应用程序都可以读取的信息。起源据说是这样的：你再次进入以前登录过的某一个网站时，在网页中可能会出现这样的字样：“Hello，Sandy. 欢迎回来”，感觉很舒服，就好像是吃了一个小饼干一样的感觉。

为什么会有 Cookie 这个概念呢？因为用来在互联网存取信息的 Http 协议本身是无状态的，对于一个浏览器发出的多次请求，承载互联网应用的服务器无法区分是否来源于同一个浏览器。所以，需要额外的数据用于维护会话，而 Cookie 正是这样

的一段随 Http 请求一起被传递的额外数据，通常由服务端或者客户端写入和读取。大多数浏览器支持的 Cookie 最大为 4096 字节，所以通常 Cookie 是用来存储少量类似于用户 ID、密码、页面默认选择之类的标识符。浏览器还限制同一网站可以在用户计算机上存储的 Cookie 数量。大多数浏览器只允许每个站点存储 20 个 Cookie。

现在的互联网服务为了体现个性化，通常都使用 Cookie。当我们给用户分配了 Cookie 之后，在日志处理中的用户识别和路径识别问题的解决就可以得到缓解，因为用户使用移动终端（笔记本或者平板电脑）的原因，Cookie 要比 IP 地址能更加精准地定位独立访客。为什么说是缓解而不是解决？原因在于个人隐私的问题使很多用户在他们的浏览器上禁止使用 Cookie 或者禁止使用永久保存 Cookie（Persistent Cookie）。

还有一种日志是由 Web 服务器和服务器上的 Web 服务程序自定义的。程序可以把所接受到的指令和所做操作的相关数据录入到日志文件中，而这些数据之后会由线下的程序来做专门处理，或是提取相关信息选择性存档，或是用以做失效回滚的保障等。比如下面来自某家互联网广告公司的日志，其中的 ad_num，adbar_id，uid，referer_url，unit_id 等都是自定义的项，是运行在该服务器上的 Web 服务程序写入的。

```
s:20120706000359^*^115.174.94.184^*^ad num=1&adbar id=7771
&uid=883&referer url=http%3A%2F%2Fwww.yjsjl.org%2Fhtml%2Fg
erenjianlifanwen%2F2012%2F0629%2F87480.html&unit id=109585
&charge mode=1&idea id=1426326&owner id=22445&plan id=3822
9&price=1000^*^a=8966 1 0%2C6024 1 0%2C6025 1 0;
p=34691 1 0%2C1 2 0;          u=97438 1 0%2C23 1 0%2C1 1 0;
i=1129965 1 0%2C82 1 0%2C9 1 0;
panshi user=2j0TAk/updJV4xbkBZEgAg==^*^Mozilla/4.0
(compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; QQDownload
691; .NET CLR 2.0.50727)
s:20120706000359^*^115.174.94.184^*^ad num=1&adbar id=7771
&uid=883&referer url=http%3A%2F%2Fwww.yjsjl.org%2Fhtml%2Fg
erenjianlifanwen%2F2012%2F0629%2F87480.html&unit id=109582
&charge_mode=1&idea_id=1426242&owner_id=22445&plan_id=3822
```

```
9&price=1000^*^a=8966 1 0%2C6024 1 0%2C6025 1 0;
p=34691 1 0%2C1 2 0;          u=97438 1 0%2C23 1 0%2C1 1 0;
i=1129965 1 0%2C82 1 0%2C9 1 0;
panshi user=2j0TAk/updJV4xbkBZEgAg==^*^Mozilla/4.0
(compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; QQDownload
691; .NET CLR 2.0.50727)
```

## 7.2 网站日志处理

Web 日志记录的是 Web 使用记录，实际上也是流水操作记录的一种，它机械而忠实地记录着访客对 Web 服务器访问的细节情况。因此，分析这些原始数据，可以对其进行一些研究工作，如系统性能分析，通过 Web 缓存改进系统设计，使得页面缓存机制更加适合实际的需要，并且可以动态适应访客访问行为模式。这些分析还可以有助于建立针对个体用户的定制 Web 服务。在这些分析结果的驱动下，可以使 Web 具有智能性，能快速、准确地找到用户所需信息，为不同用户提供不同的服务，为用户提供产品营销策略信息等。

但这只是 Web 日志分析的一半工作。另外一半工作是我们需要把网站日志整理成半成品，以供其他的数据挖掘工具和系统使用。比如我们在第 9 章中会重点讲述的网站联盟广告分析，其中大量的数据都是来自于网站日志。在这种情况下，Web 日志处理需要做的工作不是分析日志记录本身，而是整理、转换和规约在 Web 日志中的数据，形成半成品。

### 7.2.1 Web 日志预处理

如果我们的 Web 服务器是 Nginx 服务器，那么把日志文件重新定义好之后，日志的组织格式可以很方便的做好数据预处理和后续的数据挖掘。

一般普通的网站通常使用简单结构的 Web 服务器架构，通常这些网站的访问量不是很多，一般在同一角色下只有一个 Web 服务器。对于这些简单结构的 Web 服务器，在一段时期内用来分析的原始数据往往就是一个 Web 日志文件。一般来说，对于这种情况如果我们做日志分析和传统的数据挖掘的处理手法有类似的地方，也大致可以分开原始数据预处理、挖掘算法和模式分析几个主要的步骤。

Web 日志挖掘的原始数据预处理包括依赖域的数据净化、用户识别、会话识别和路径补充等。对日志进行预处理的结果直接影响到挖掘算法产生的规则与模式。因此，预处理过程是保证 Web 日志挖掘质量的关键。

在日志文件上做数据净化指删除 Web 服务器日志中与挖掘算法无关的数据。所谓有关无关是和数据挖掘应用场景有关。如果我们主要做网页访客分析，那么只有日志中 HTML 文件与用户会话相关（但有些以浏览图片或者查询其他媒体为主的网站可能是例外），因此可以通过检查 URI 资源的后缀删除认为不相关的数据。经过数据净化，数据可以十分集中。

如果日志分析需要和其他数据来源对接，那么我们需要注意在数据净化的过程中不能丢弃一些在分析中暂时用不到的信息。比如，我们的网站可能和其他的一些网站做数据交换，而他们所关注的信息和我们所关注的不一定完全一致。

为了满足我们需要完成的一些任务，可以对这些数据进行一些简单的数据转换，比如如果我们对于时间戳中的时间部分不感兴趣，可以把时间戳中的时间部分去掉，只留日期。同样，我们也可以把用户代理中的一长串文字转化成实际的浏览器。表 7-2 是转化后的一张表，我们只取了前面 8 行作为演示。

表 7-2　数据转化后的日志表

| IP | 时　间 | 响应代码 | 访问页面 | 用户代理 |
|---|---|---|---|---|
| 66.249.65.××× | 7/1/2012 | 200 | /news/20120630.htm | Firefox |
| 66.249.65. ××× | 7/1/2012 | 403 | /products/24421.htm | Firefox |

续表

| IP | 时　间 | 响应代码 | 访问页面 | 用户代理 |
|---|---|---|---|---|
| 66.249.65.××× | 7/1/2012 | 200 | /news/20120630.htm | Firefox |
| 66.249.65.××× | 7/1/2012 | 200 | /news/20120629.htm | Firefox |
| 66.249.65.××× | 7/1/2012 | 200 | /news/20120628.htm | Firefox |
| 201.201.65.××× | 7/1/2012 | 200 | /index.htm | Chrome |
| 201.201.65.××× | 7/1/2012 | 200 | /about.htm | Chrome |
| 201.201.65.××× | 7/1/2012 | 200 | /index.htm | Chrome |

而通过转化之后，我们也可以从数据中直接了解很多信息。这里可以告诉我们的是有一个来自于 66.249.65.×××的用户可能对产品感兴趣，因为他在同一天内多次访问了网站，当他单击了/products/24421.htm 的页面而没有收到任何信息（403 错误）之后，他还是继续看了我们的两个新闻页面。
我们再来看一个例子，如表 7-3 所示。

表 7-3　可能的黑客攻击表

| IP | 时　间 | 响应代码 | 访问页面 | 用户代理 | Cookies |
|---|---|---|---|---|---|
| 55.210.77.××× | 6/17/2012 | 403 | /movies/us/black.htm | IE6 | 无 |
| 55.210.77.××× | 6/17/2012 | 200 | /movies/us/black.htm | IE6 | 无 |
| 55.210.77.××× | 6/17/2012 | 200 | /movies/us/black.htm | IE6 | 无 |
| 55.210.77.××× | 6/17/2012 | 200 | /movies/us/black.htm | IE6 | 无 |
| 55.210.77.××× | 6/17/2012 | 403 | /movies/us/black.htm | IE6 | 无 |
| 55.210.77.××× | 6/17/2012 | 403 | /movies/us/black.htm | IE6 | 无 |

表 7-3 比表 7-2 多了一列“Cookies”，也就是用户浏览器上的 Cookie 表示。我们可以看到来自 IP 地址 55.210.77.×××的六次访问都是访问同一个电影文件，而且都没有 Cookie。在第 9 章的网盟广告中我们可以看到，用户是否有 Cookie 对于是否有作弊行为是有一定作用的。这里来自 IP 地址 55.210.77.×××的访问不但没有 Cookie，而且对于同一电影网址多次访问。那么这里很可能有黑客攻击的嫌疑。

假如我们排除表 7-3 的作弊嫌疑，那么这里出现了 50%的 403

响应代码或者 403 错误，这是值得我们敲响警钟的。完整排查日志我们可能会发现这里出现的大量的 403 错误不是偶然现象。通过与后面章节中的表 7-6 类似的方法，我们可以统计在某一天或者某一个时间段出现的 403 错误。

403 错误出现的原因很多：可能是因为过多的用户连接导致我们的服务器超负载；可能是因为有大量没有权限的用户试图访问某些资源；可能是网站上的某些证书已经到期；甚至可能是因为用户的浏览器不支持加密算法所导致。通常经过进一步的日志文件分析可以判断到底是哪一种 403 错误。在表 7-3 的例子中，因为出现 403 错误的多数是在播放电影的页面，那么很可能是因为过多的用户连接造成的。我们需要迅速调整，或者增加带宽，或者增加和/movies/us/black.htm 相关的 Web 服务器资源。不然，用户如果太频繁地遇到访问错误，很可能会去其他网站上找寻类似资源，从而造成用户的流失。

如果我们对用户代理信息做一个简单的字符串识别，比如把前文出现过的“Mozilla/5.0(compatible; Googlebot/2.1; +http://www.google.com/bot.html)”转化成“谷歌爬虫”，把“Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7”转化成“Windows 访客”。我们可以得到类似表 7-4 的列表，表 7-4 也只列出了 7 条数据作为示意。

表 7-4　网站日志简单用户识别分析表

| IP | 时　间 | 响应代码 | 用户识别 | 访问页面 |
|---|---|---|---|---|
| 66.249.65.107 | 7/1/2012 | 200 | 谷歌爬虫 | /index.htm |
| 66.249.65.107 | 7/1/2012 | 200 | 谷歌爬虫 | /news/main.htm |
| 66.249.65.107 | 7/1/2012 | 200 | 谷歌爬虫 | /news/20120630.htm |
| 220.181.7.13 | 7/1/2012 | 200 | 百度爬虫 | /index.htm |
| 130.210.77.××× | 7/1/2012 | 200 | Windows 访客 | /movies/us/black.htm |
| 130.22.102.××× | 7/1/2012 | 200 | Windows 访客 | /movies/us/black.htm |
| 111.17.118.××× | 7/1/2012 | 200 | Macintosh 访客 | /movies/us/black.htm |

从表 7-4 的日志分析表中我们可以得到很多信息。例如我们

从谷歌爬虫的访问频率来看，发现谷歌对我们网站的访问频率是恰当的，使得我们的网站在谷歌上有合理的收录。但是百度爬虫的记录在一段时间内网站日志记录很少或者不存在，那么我们就需要分析为什么百度不对我们的信息进行抓取或者减少了收录。是因为百度搜索引擎的规则改变？是我们网站被百度降权？还是因为网站的外链出现了问题？

在 Web 日志挖掘的原始数据预处理中，用户识别是很重要的一个环节。除了上文所述对于搜索引擎爬虫的判别之外，还需要能够识别出独立访客。由于本地缓存、代理服务器和防火墙的存在，使有效识别用户的任务变得十分复杂。用户识别是个难题，因为我们目前没有一种方式可以保证准确识别用户。只用看 IP 地址的方式是不够的，因为同一 IP 地址上可能会有多个用户存在。比如盘石公司总部的近千人和数千台机器，在访问外网时，显示的 IP 是四个 IP 中的一个。同样，在同一幢大楼中用相同 ISP 的客户在上网时可能显示的是同一 IP。另外，现在通过代理服务器（Proxy Server）访问互联网的人越来越多，那么通过同一个代理服务器的客户也会显示同样的 IP。有效识别客户目前一般被采用的方法是基于一些启发性规则。其中一种规则是如果 IP 地址相同，但是代理（User Agent）信息变了，表明用户可能是在某个防火墙后面的内网的不同用户，则可以标记为不同的用户。

另一种用户识别方式是将访问信息、引用信息和网站拓扑结构结合，构造出用户的浏览路径。如果当前 IP 地址请求的页面同用户已浏览的页面没有链接关系，则认为是存在 IP 地址相同的多个用户。但是这一方式不能分辨访问浏览路径中的用户。

在用户识别基础之上的另一个问题是会话识别。如果 Web 服务器日志跨越时间区段比较大，比如超过一周，那么同一用户可能多次访问该站点。而会话识别的另外一种情况是同一客户开始回话的时间点是在上一个被处理的服务器日志中，而在本次服务器日志有上次访问的延续。会话识别的目的就是将用户同一次

访问的记录放到单个会话中，但是把不同的访问分开。

最简单的方法是用超时的技术，如果两个页面之间请求的时间差值超过了一定界限就认为用户开始了一个新的会话，而两次访问的时间在界限之内就算是同一次会话。例如，可以设置 30 分钟，那么所有间隔在 30 分钟以上的同一用户访问都算是一次新的会话。

在用户识别和会话识别的基础之上还有一项需要做的工作是用户访问路径的补充。举例来说，通过网站拓扑结构分析发现当前请求的页面与用户上一次请求的页面之间没有超文本链接，那么用户很可能使用了浏览器上“BACK”的功能调用缓存在本机中的页面。从历史记录中检查引用信息以确定当前请求来自哪一页，如果在用户的历史访问记录上有多个页面都包含当前请求页面的链接，则将请求时间最接近的作为当前请求的来源。如果引用信息不完整或者无法分辨，我们需要利用站点的拓扑结构来完成用户的会话。

用户识别，会话识别和访问路径补充是 Web 日志挖掘中数据预处理常用的步骤，其目的就是尽量使域处理后的数据比较真实和完整，为后面的数据分析和挖掘打好基础。

在实际应用时，在做预处理的过程中，要完全准确地识别出全部的独立访客，准确分析出全部访客的行为是不切实际的。

我们的困难一是在于复杂的网络环境，Web 服务器一般不会直接暴露在外网，而中国的网络质量在各地的差异很大，导致很多访问的路径与预想情况差别很大，想从原始的日志文件中准确、直接地辨认用户和行为是十分困难的。

二是由于不同层次的访客浏览网站行为的复杂性、不确定性和不连贯性。我们应该根据分析的需要，首先确定一定需要的行为，再确定这些行为出现的一些条件和特征，从而确立一些分析规则，将这些行为尽量挖出来，对于不需要的行为数据应该尽量过滤。

只有预处理数据做好了，后面的分析和挖掘才会比较准确，

因此，多安排一些时间放在 Web 日志数据的预处理阶段是十分必要的。

### 7.2.2　Web 日志分析和数据挖掘

对于预处理后的数据，就可以进一步进行识别用户浏览行为的序列模式了。在 Web 日志上做序列模式数据挖掘的主要算法是频繁遍历路径（Frequent Path Travel）。

所谓遍历路径就是在用户会话时请求页面所组成的序列。由于用户会话中，既包含请求页面又包含路径补充时添加的页面，因此挖掘频繁遍历路径时，首先在每个用户会话中找出所有的最大向前路径。挖掘频繁遍历路径问题转化为在所有用户会话的最大向前路径中发现频繁出现的连续子序列的问题。要寻找这些频繁遍历路径，必须定义这些连续子序列的长度和支持度（Support），所谓支持度就是包含频繁遍历的用户会话数目。

当 Web 服务器具有复杂结构时问题就变得麻烦起来了。所谓 Web 服务器的复杂结构主要是具有复杂拓扑分布式结构的网站，同时存在多台 Web 服务器，而且日志文件存在于各个服务器上的情况。对于这种分布式的结构，因为不可能把所有的日志文件用同一台数据挖掘服务器来处理。我们一般可以采用多代理技术的分布式 Web 日志挖掘技术来解决。多代理分布技术用到 Web 日志挖掘系统主要基于多代理的三重体系结构，包括用户访问层、代理层和计算处理层。用户代理层根据用户访问层传过来的用户请求进行分析，根据已定义好的相关规则和算法把已存在的数据发送到不同计算处理层按相应的算法进行分析挖掘。代理技术的使用能够有效地对多个异构 Web 服务器同时进行分析和处理。

我们用最简单的方法来解释多代理的分布式 Web 日志挖掘，可以按照用户来分类，根据 IP 地址的 Hash 表示（哈希，一种数据压缩映射方法），让负责不同计算的机器来处理具有相同 Hash 数值的用户访问信息，结束之后把计算结果再通过代理汇总到一

个数据仓库之中。

最后，对于一个可用性的系统，当然需要将挖掘结果直观而明了地展示给用户，以便于用户理解。而这些频繁遍历路径正可以给网站提供一些宝贵的资料来改进网站的结构和性能，例如，对这些被频繁访问的相关网页配置专门的 Web 服务器，适当增加缓存，使用户访问速度加快等。

除了对网站上的用户和访问路径做分析之外，我们还可以对于日志做一些数据统计工作，从而对网站的运营产生帮助。

下面我们以国内某个论坛访问为例来看如何对网站日志做数据统计。我们先来看表 7-5。

表 7-5　页面访问统计列表

| 日期 / 访问页面 | 6 月 1 日 | 6 月 2 日 | 6 月 3 日 | 6 月 4 日 | 6 月 5 日 | 6 月 6 日 |
|---|---|---|---|---|---|---|
| /index.htm | 37781 | 45523 | 38842 | 45511 | 69966 | 38811 |
| /bbs/main.htm | 32244 | 37778 | 28666 | 41128 | 61108 | 31112 |
| /about.htm | 155 | 202 | 137 | 129 | 286 | 98 |
| /bbs/gold.htm | 19973 | 27882 | 12241 | 22551 | 48543 | 17227 |
| /bbs/messages/21234.htm | 832 | 745 | 701 | 676 | 632 | 601 |
| /bbs/messages/19982.htm | 772 | 713 | 689 | 656 | 884 | 611 |
| /bbs/messages/17351.htm | 233 | 201 | 189 | 165 | 342 | 145 |

表 7-5 列出了一个论坛的访问统计表。我们选取了包括网站首页（/index.htm）、论坛首页（/bbs/main.htm）、论坛精华页（/bbs/gold.htm）在内的 7 个页面。从页面访问统计列表中可以看到网站首页和论坛首页的访问量是最大的，而除了偶尔的情况，所有页面的访问基本上是成正比的。如果网站首页 index.htm 的访问量大，那么论坛首页/bbs/main.htm 的访问量和论坛精华页/bbs/gold.htm 的访问量也会较大。但是，每个单独论坛页面/bbs/messages/××××.htm 的访问量并不和总访问量成正比，而是随着时间推移呈总体下降趋势，从论坛的特性来说，这是可以理解的，旧的讨论话题对人的吸引力总是越来越小的。

我们再来看一下在这个论坛上某一个月页面访问响应代码的统计。在表 7-6 中列出了某一个月前 14 天除表明正常响应代码 200 之外的所有其他响应代码的统计数据。比如在这个月的 3 号，服务器出现了 9 次 301 响应代码，228 次 304 响应代码，3 次 403 响应和 2 次 404 响应代码，而响应代码 500 和 502 均为零。

表 7-6　页面访问响应代码统计列表

| 日期 / 响应代码 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 301 | 2 | 5 | 9 | 0 | 0 | 17 | 23 | 10 | 3 | 2 | 3 | 1 | 0 | 4 |
| 304 | 376 | 189 | 228 | 144 | 172 | 302 | 498 | 224 | 210 | 133 | 99 | 43 | 173 | 201 |
| 403 | 4 | 3 | 3 | 3 | 3 | 29 | 53 | 17 | 5 | 0 | 0 | 4 | 2 | 3 |
| 404 | 1 | 2 | 2 | 2 | 2 | 2 | 107 | 97 | 2 | 0 | 0 | 3 | 2 | 2 |
| 500 | 0 | 0 | 0 | 0 | 0 | 11 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 502 | 0 | 0 | 0 | 0 | 0 | 9 | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

在 7.1 节中我们对于网站响应代码做过介绍。对于这个论坛来说，我们最关注的两个值是 403 和 404。我们又可以把响应代码 403 和 404 称作 403 错误和 404 错误。403 错误表示不可访问，而 404 错误表示找不到访问页面。从表 7-6 中我们看到在这个月的 6 号、7 号和 8 号三天，服务器出现较多的 403 错误和 404 错误，但是从 9 号起基本又恢复正常。系统在这几天发生状况的原因还需要进一步分析数据才能得出结论。

## 7.3 邮件日志

邮箱的日志记录给出的分析信息更加丰富。

除了和网站日志分析信息相关的内容之外，通过对邮箱中邮件内容的信息挖掘以及用户对不同邮件的处理，包括回复和点击等，我们可以总结出更多和该用户相关的个人信息。

比如你收到的邮件和回复的邮件中多次提到从上海到美国

的旅游，那么如果给你推送和“中美廉价机票”相关的广告信息，效果应当是不错的。又比如你在邮件中多次提及股市和股票的选择，那么和“优质股票基金”或是“高回报投资”相关的广告信息就可能有不错的效果。

Gmail 根据对每个邮箱信息的日志分析和数据挖掘，使得他们的邮件营销功能相当有针对性。请看图 7-4。

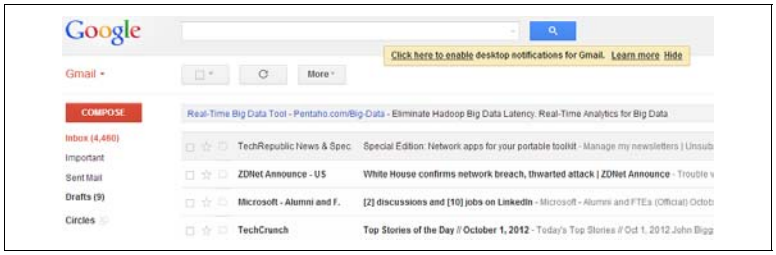



图 7-4　Gmail 内容图

图 7-4 是我的 Gmail 邮箱的截图，而在最上方的“Real-Time Big Data Tool - Pentaho.com/Big-Data”就是一家提供实时数据挖掘工具公司做的广告。他们做的广告确实是我感兴趣的内容。如果有使用 Gmail 邮箱的客户，不妨关注一下 Gmail 上方的关联广告，看这些广告是否是你所感兴趣的内容。

除了和个人资料相关的信息之外，在邮件日志中还有不少信息也是值得邮件提供商关注的。比如我们可以对用户常用的一些操作进行“频繁路径分析”，如果发现执行某一类操作的用户相对比较多，那么我们可以把这些操作加入到商业邮箱的常用操作中去。

## 7.4 本章相关资源

- 本章相关参考文献：

[1] 杨怡玲，管旭东，陆丽娜等. 一个简单的 Web 日志挖掘系统[J], 上海交通大学学报, 2000, 34（7）.

[2] Pang-Ning Tan, and Vipin Kumar, Mining Association Patterns in Web Usage Data (2002). International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet.

[3] Pang-Ning Tan, and Vipin Kumar. *Discovery of Web Robot Sessions based on their Navigational Patterns (2002)*. Data Mining and Knowledge Discovery, 6(1): 9-35.

- 本章相关网址：

[1] http://www.ipchecking.com/

[2] http://nginx.org/en/docs/

[3] http://wiki.nginx.org/Main

# 第 8 章
# 数据挖掘和电子邮件

全世界每天发送 1500 亿封邮件，其中至少有三成是带有广告性质的，75.8%的全球 500 强企业使用邮件营销，仅在 2012 年邮箱使用量将近 5 亿。电子邮件营销是市场营销方案中的一个重要环节，如何做好这个环节是至关重要的，结果好坏会直接影响到企业的市场效果和收益。如何用数据挖掘手段通过海量数据的分析做好电子邮件营销是本章讨论的第一个专题。而作为双刃剑的另外一面，在平时每天收到的邮件中，大量的垃圾邮件也让人觉得头痛不已。从企业邮箱的角度，如何通过数据挖掘分析辨别垃圾邮件是本章讨论的第二个专题。

## 8.1 邮件营销与垃圾邮件过滤

1994 年 4 月 12 日，美国两位律师坎特和西格尔夫妇把一封“绿卡抽奖”的广告信发到他们可以发现的 6500 个新闻组，这个新奇的“邮件炸弹”在当时引起了疯狂的下载与转发，也使得很多服务器商的服务器处于瘫痪的状态。在这对夫妇所著的 *How to make a fortune on the internet superhighway*（如何在互联网高速公路上赚钱）一书中，详细介绍了这次的辉煌经历：通过邮件发布广告信息，只花了不到 20 美元的通信费用就吸引来了 25000 个潜在客户，其中有 1000 个转化为新客户，从中赚到了 10 万美元。坎特和西格尔夫妇也因此被称为通过 EDM 赚钱的第一人。

在今天，经过用户许可的 EDM（Email Direct Marketing，电

子邮件营销）已经和数据库营销（DataBase Marketing）一样，成为市场营销中的一项重要组成部分。如图 8-1 所示为收到的营销邮件。


图 8-1　邮件营销示意图

然而，类似坎特和西格尔夫妇的这种未经用户许可而“狂轰滥炸”式的邮件发送模式所产生的垃圾邮件不算是真正的邮件营销。这些无孔不入的垃圾邮件也正困扰着很多人（如图 8-2 所示）。


图 8-2　垃圾邮件示意图

在进行本章的正式叙述前，我们先对垃圾邮件和邮件营销这两个概念进行区分。

2003 年美国政府通过了历史上第一部反垃圾邮件法——CAN-SPAM（控制非自愿色情和促销攻击法案）。这部法律划清了垃圾邮件和营销邮件的界限。如果你发送包含市场推广信息或推销信息的邮件时必须严格遵守这部法律的相关规定。若违反了其中任何一条规定，单人罚款额度最高可达 16000 美元。同年，

《中国互联网协会反垃圾邮件规范》出台，在该规范中，垃圾邮件被界定为包括下述属性的电子邮件：

（1）收件人事先没有提出要求或者同意接收的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件；

（2）收件人无法拒收的电子邮件；

（3）隐藏发件人身份、地址、标题等信息的电子邮件；

（4）含有虚假的信息源、发件人、路由等信息的电子邮件。

与垃圾邮件相对应的邮件营销则指在用户事先许可的前提下，通过电子邮件的方式向目标用户传递有价值信息的一种网络营销手段。

从两个定义可以看出正规的广告邮件的首要标准就是要经过用户许可，也就是说那种滥发广告邮件的模式并不算做邮件营销。从营销角度来说，虽然发一封邮件的成本低，但是毫无目标式的发送并不能产生效果，客户要对你的产品产生兴趣，至少他要对您的产品有一定印象。例如，如果用户注册了你的网站并订阅了相关的邮件服务，此后发送的邮件包括促销邮件就不属于垃圾邮件的范围，相反如果用户从未与发邮件的广告厂商发生过任何接触，这些邮件往往是这些广告的厂商通过低价购买用户邮箱地址或者自行配对、地址猜测的方式群发的，自然算作垃圾邮件而不能作为正规广告邮件。

图 8-3 为卓越亚马逊在用户注册后发送的促销广告邮件附带的说明，从说明中可以看出卓越亚马逊之所以会发送这则邮件是因为用户在该网站注册过。从这个角度说，这则邮件的发送是经过用户许可的（在注册时会附带邮件订阅选择的服务），从中我们还看出如果用户不愿再接受自动订阅类邮件，可以点击退订邮件。这引出第 2 个评价标准用户可以退订，如果没有这一条，用户无法拒绝接受，则邮件就变为垃圾邮件，所以邮件营销的邮件也务必要加上这一条。

防垃圾邮件规范中的第 3、4 条主要从邮件的信息透明度和准确性来判断垃圾邮件和正常邮件的区别。不管在邮件的什么地

方，您都必须清楚地告诉收件人这封邮件是一个广告。规范规定必须在邮件中将其直截了当地说出来。决不能隐藏身份、地址、标题等标示信息。


本自动订阅类邮件信息的收件地址是：
您之所以收到这封邮件，是因为您曾经注册成为亚马逊的用户。我们保证仅向您发送关于亚马逊的产品、促销优惠以及服务的电子邮件。卓越亚马逊尊重并保护您的隐私。
如果您不愿意继续接收来自亚马逊的自动订阅类邮件，请点击退订邮件。如果您想订阅更多或更新您所订阅的邮件列表，请点击更新我的邮件订阅。
商品价格或促销内容如有调整，以亚马逊上最终页面为准。任何问题欢迎您致电或邮件联系亚马逊客服中心：400-810-5666。如有问题也请访问帮助中心。
amazon.com. 旗下公司
Copyright © 1999 - 2012 Amazon.cn All Rights Reserved

图 8-3　亚马逊邮件订阅示意图

有些邮件为了吸引用户去打开经常会假装自己是另外一家知名网站或公司，这是垃圾邮件发送者为通过垃圾邮件过滤器或吸引用户打开邮件惯用的伎俩。这也是不合法的，也就是说，为了使你的广告邮件不被判为垃圾邮件，你发信的邮件地址必须是你自己的。你邮件中推广的网站域名必须是真实的，客户在邮件中看到的信息必须是真正关于你或你所经营的业务。

总之，在进行邮件营销之前，你必须保证自己所发的邮件在法律上是符合规定的，只有做到这一步，后期的邮件营销活动才称得上是有效。即使做到了合法，EDM（电子邮件营销）在很多时候也不是很有效，而主要的原因是因为缺乏较清晰的发送目标。我们在本章会描述如何通过数据挖掘的手段来提升邮件营销的效果。

## 8.2 数据挖掘和邮件营销

### 8.2.1　如何有效地进行邮件营销

除了垃圾邮件之外，我们常常会收到一些无关痛痒的所谓促销邮件。有些内容或者商品完全不是我们所感兴趣的，还有些虽然可能是我们感兴趣的内容或者商品，但是促销不够吸引人，更

加离谱的是有的时候我们收到的折扣券是关于我们刚刚在某家网站购买的同一商品。这往往是有些商家不加分析的全部或随机发送促销邮件的结果，应该说这种没有针对性的撒网式的营销方式效果不但不好，而且可能会起到副作用。

本节中，我们会从数据挖掘角度来讲解怎样提高邮件营销的效果。

作为消费者在进行网络购物之前都会进行注册，这些注册信息对卖家来说是很重要的，我们可以根据注册信息来跟踪客户的后续行为，待到一定量的数据产生后就可以利用这些数据来对客户进行分类。此时如果再碰到促销活动，我们只需对那些更有可能响应促销的客户发邮件，这样促销效果就得到了保证。比如说我们从客户的行为分析中发现他多次把某类商品加到购物车中，但是最终一直都没有成交，那么我们就可以给客户发关于类似商品的促销邮件，把他转化成消费客户。

美国 2011 年的邮件广告收入是 6.3 亿美元。2011 年 GROUPON 做得风生水起的时候拥有 3000 万个邮件地址，并且都是经过用户许可的，像微软、IBM 等做得也都非常不错。

图 8-4 截自亚马逊（Amazon）公司给接受者发送的电子营销邮件，邮件的发送者是亚马逊的本地部门（Amazon Local）。邮件中为接受者量身定做了两个机会供选择。一是七折的夏威夷旅游，二是五折的开飞机学习班。这两个选择都不错。我仔细翻阅了最近几个月来自亚马逊的电子营销邮件，发现不少还是相当诱人的选择，我相信这是亚马逊参考了笔者在亚马逊上的历史消费记录的结果。

图 8-5 截自 Trip Advisor 发给我的一封营销邮件。Trip Advisor 是一家上市的互联网公司，是全球最大最受欢迎的旅游社区之一，也是全球排名第一的旅游评论网站。Trip Advisor 每隔一段时间就会给我发营销邮件，而每隔几封邮件我就会发现感兴趣的内容。

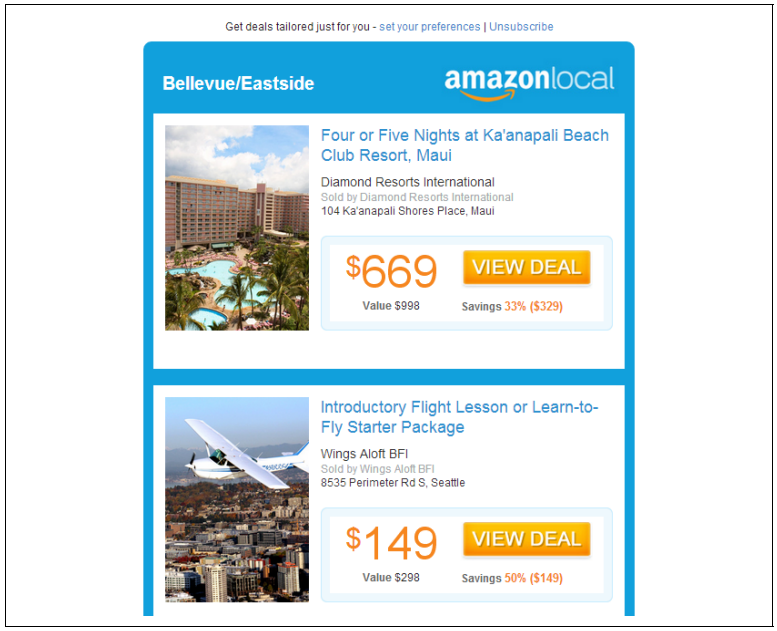



图 8-4　Amazon Local 营销邮件


图 8-5　Trip Advisor 营销邮件

在 7.3 节中我们提到谷歌的 Gmail 做的数据挖掘，但是在本章中我们并不打算以此为案例来展开。原因是绝大部分的公司都不可能像谷歌公司这样拥有客户邮箱的直接访问权限，而通过其

他手段获取到用户邮箱的内容是不合法的。在本章中我们讨论的邮件营销是通过我们可获取的用户信息和数据来做的，应当是有推广价值的。

邮件营销在网络营销里的作用到底有多大？

对于每个销售产品的公司而言，所有的客户都处在一个生命周期中。当客户第一次访问网站或者第一次对产品产生兴趣的那一刻，这个生命周期就开始了，可以把这些客户判定为潜在客户。而当客户开始购买产品或频繁访问网站时，就处在活跃期。活跃了一段时间或是完成了第 N 次购买之后，如果客户在一定的时间内没有再回来访问，那么这个客户就处在流失期。时间较短的（时间因产品而异）称为短期流失期，而时间较长的称为长期流失期。因为兴趣的转移或者搬家等其他原因，最后客户迟早会淡化乃至忘记产品，这样客户生命周期就陷入衰退期。

对于处在生命周期不同阶段的客户，如果能在适当的时候与客户进行沟通，重新点燃客户的兴奋点，就有可能延长其生命周期。而在这个过程中，电子邮件作为一种非常有效且廉价的客户沟通方式，无疑是一个不错的选择。例如，可以在客户注册后通过发送邮件招揽客户购买，也可以在客户完成第一次购买后继续发邮件向其推广您的产品，还可以定时发送电子邮件提醒、提供特别优惠或公司新闻，让客户不会对公司产品丧失印象等。

下面我们针对不同生命周期的客户来阐述 EDM 的处理方法。

#### 8.2.1.1 潜在客户

潜在的客户可能只是在网站访问或者注册了，他们对品牌和产品并不十分了解。这样的客户购买的意向往往不大而且很快就会对产品产生遗忘，所以对于这类客户的重点是激发他们对产品的兴趣。此时可以通过优化邮件的形式和内容或者为其提供一些优惠活动来吸引其购买。当然如果这些邮件并未产生预想效果也没关系，因为客户即使暂时未产生购买行为，但是发送的邮件至少会使客户加深对产品的印象，当他下次真的需要这类产品时，

首先想到的就是公司的产品了。

客户第一次去到你的网站往往只是通过一个广告的点击，但是在访问之后客户如果留下了他的电话号码，这个客户的质量是要高于一个普通的来访者的。如果我们的网站上有在线沟通工具，而该在线沟通工具和邮件营销系统在后台是打通的，那么下一次该客户在我们的网站上出现时，系统应当给客服以提示，请他们主动和客户联系，以提高转化率。

#### 8.2.1.2　首次购买客户

当客户首次产生购买行为后，就从潜在客户转化成为真实客户。一方面说明我们前期的工作产生效果了，另一方面，最重要的客户维护步骤才算是真正开始。对于首次购买客户，关键是要让客户进行二次购买，只有客户完成了两次以上的购买，我们才能把他们转化成公司的忠实客户。要做到这点就需要让客户感觉到自己对本公司的重要性，公司是很在意他的。

在邮件营销方面，我们可以在客户首次购买之后自动发送感谢邮件。在这封邮件中，不仅要感谢客户购买了公司的产品，还要询问客户有什么反馈意见，提出使用产品中所需要注意的事项，需要明确告诉客户一旦发生情况，如何联系客服。邮件同时还需要详细介绍公司，让客户感觉到他们之前的购买决定是如何的英明。

#### 8.2.1.3　活跃客户

当客户不断产生购买行为或频繁访问公司网站时，即是其活跃期，此时要做的是尽量延长其活跃时间和尽量提高活跃用户的 ARPU（Average Revenue Per User，每个用户的平均收入）值。ARPU 值最早应用在通信行业，比如电信和移动，是以每个月客户的电话费消费来判别客户价值的计算方法。现在的含义是一段时间（比如一个月）平均每个活跃客户在网站上产生的总收入。

方法之一就是通过邮件告诉客户一些与其所购买产品相关的信息以及吸引其继续购买的促销广告。这里可以借鉴的想法和

创意是很多的，比如每周一句或者每周特惠的邮件等都可以帮助你保持客户的活跃度。我们也可以从下面 UTC 的案例中学到一些独特的做法。

#### 8.2.1.4 短期流失客户

当客户在较短的时间内未产生任何购买行为则称之为短期流失客户，具体的时间长短可能会因产品种类、购买周期等因素的不同需要企业自己来设定。比如女装类产品我们可以设定为两个月，游戏类产品设定为一个月等。由于这类客户与公司之间的关系较为密切，也产生过购买行为，对品牌有一定的认知度。如果之前有过多次购买，该客户可能还对公司有过一定的忠诚度，所以可能只要小小的一个优惠就可以将他们留住。

此时可以做的事情很多，我们通过电子邮件可以做的是选择合适的时刻给这些客户发送一封邮件将其留住。这时发送的邮件不一定是促销信息，比如我们可以选择在节日发送个性化的祝福邮件，其中可以重申一下之前所购买产品的一些信息，比如可以说一下之前购买的护肤品需要长期使用才会最有效果等。通过电子邮件发送这些信息的成本很低，如果仅此就可以引导一定比例客户重新购买，性价比是相当高的。

#### 8.2.1.5 长期流失客户

挽留一个老客户比挖掘一个新客户的成本低得多，所以对于那些很长时间未购买公司产品的客户，我们挽回所花的成本应当还是低于重新挖掘一个新的客户。

对于长期流失客户的失效挽救只是发送电子邮件可能是不够的。但是我们也应在一定时间内发送一些提醒邮件，比如可以通过邮件给他们一个无法拒绝的优惠等。

除了对客户进行分析之外，在邮件营销上还有一些简单的技巧也需要注意。比如：

① 避免将一封邮件发给大量的接收者。确信一封邮件每次仅发给一个人。如有可能，在邮件的开始加上客户的姓名。

② 选择一个能够使人信任的主题。避免使用垃圾邮件常用的单词和符号，例如“免费”、“派送”、“优惠”等。

③ 响应电子邮件营销的要求。不仅要在电子邮件中含有“退订”的超级链接。订阅者可能会要求你将他的地址从你数据库中删除。马上去做，并将处理的结果通知他。响应他们的要求有助于避免潜在的垃圾邮件抱怨，避免他们将你列入 ISP 的黑名单中。

④ 在做 EDM 时，你需要监视所有的退回邮件。当一个特殊的邮件多次被退回，你可以亲自联系接收人，并请他们将你的地址加入他们的白名单。如果得不到他们的响应，就要从数据库中将这些电子邮件地址删除。如果你使用的是电子邮件管理软件，找出无效的邮件地址并删除它们。过多的退回邮件可能被垃圾邮件过滤器过滤。

我们下面来通过几个 EDM 的实际案例来看如何把数据分析和数据挖掘与 EDM 相结合，通过个性化的 EDM，为企业产生最高的商业价值。

### 8.2.2　邮件营销案例分享之一

电子商务品牌 UTC 行家作为国际箱包品牌运营商，旗下拥有瑞士军刀威戈、全球旅行家第一选择 EagleCreek、精英商务新经典演绎 BRIGGS&RILEY 等十余个国际顶级箱包品牌在大中华地区的总代理权，产品线涉及时尚、商务、休闲、户外运动等多个系列。而作为传统企业成功转型进军电子商务领域的代表，UTC 行家在行业被誉为“中国箱包隐形冠军”。

不以营销为目的市场活动都是耍流氓！人性都是懒惰的，“免费、好玩、简单”是 UTC 行家进行针对客户的个性化邮件营销过程中的关键词。下面，我们以 UTC 行家为例，来看一下他们是如何结合不同生命周期的客户进行“行家 EDM 营销”。如图 8-6 所示。


图 8-6　UTC 营销邮件示意图

UTC 行家自成立以来始终秉持“everything for travel”的团队运营理念，坚持“以人为本”的客户维系原则，为广大旅行爱好者提供国际顶级旅行装备解决方案。反映在邮件营销上，就是结合客户生命周期，围绕“客户互动体验”关键词，开展个性化 EDM 营销，如下：

- 潜在客户——品牌注意力营销

互联网信息时代，电子商务领域从未缺少产品，更从未缺乏品牌，如何对企业潜在客户进行后续网购行为的刺激是企业营销人员的重要工作内容，此谓之“客户转化率”。

UTC 行家拥有瑞士军刀威戈、环球旅行家第一选择 eaglecreek 等国际顶级优质的产品基因，针对平台潜在客户，UTC 行家更加注重提升用户体验：凡首次注册成为行家网用户，即可获得 20 元优惠券，同步后续其结合邮件推送平台品牌产品的活动促销——以活动、折扣等方式开展行家品牌注意力邮件营销，此也是企业 EDM 营销中一种常用的营销手段。

- 首次购买客户——优惠体验+数据挖掘

一般而言，用户注册成为企业的用户之后，即为企业的潜在客户或准客户。不论成交金额大小，凡完成一笔订单，即为企业的首次购买客户。然而，就当前客户邮件营销角度而言，实际能够针对目标人群，执行精准化邮件营销的企业所占比例极小。

与大多数企业不同的是，UTC 行家对平台用户采取了较好的客户维系，其中之一为“优惠体验”，如凡是 UTC 行家平台用户，均可优先享受平台的各种活动优惠体验活动，并可凭借其购物所产生的积分以代金券等方式进行消费、体验互动。对于消费者而言，优惠、好玩是其愿意花更长时间停留页面、参与互动的主要两个因素，当然，此举是在保证产品质量的情况之下。UTC 行家正是结合了消费者的行为习惯，并通过分析合理刺激老客户的时间，利用活动、游戏、优惠等方式，以邮件形式对其进行对老客户的开发，提升客户的活跃度、二次购买、停留时间等。

根据平台上客户的访问和首次消费行为，UTC 对首次购买客户采用了聚类数据挖掘的算法，根据客户的消费金额和是否使用优惠券，定期利用邮件发送活动信息。对于消费金额高而且使用了优惠券的客户，优先发送关于优惠体验的邮件；对于消费金额高但是没有使用优惠券的客户，根据客户的具体信息，优先发送关于活动或者游戏信息的邮件；对于消费金额较低但是使用了优惠券的客户，交替发送关于抽奖活动和优惠体验的邮件；而对于消费金额较低但是没有使用优惠券的客户，发送关于活动和游戏的邮件。如图 8-7 所示。


图 8-7　UTC 营销邮件示意图之二

- 活跃客户——微博+数据挖掘

针对平台活跃客户群体，UTC 行家则采取新的邮件营销策略。比如，众所周知，当前如火如荼的微博是一种新媒体方式，也是一种营销工具。而对从事客户营销人士而言，微博则俨然成为邮件营销的一种新玩法。

笔者在调研中发现，UTC 行家在开展邮件营销过程中较好的结合了微博营销的功能特点，以植入的方式将微博营销与邮件营销优势予以结合：将企业微博活动（其中涉及新品发布、企业新闻、活动公示等）进行植入。据数据结果分析显示：在邮件营销中增加微博活动，如互动交流、抽奖活动等，邮件点击转化率整体提升 10%～25%，如图 8-8 所示为 UTC 行家部分邮件营销中涉及植入的微博营销活动——“赢 UTC 拉杆箱，游最美滩涂霞浦”、“去你的亚马逊”。

对于活跃客户，UTC 行家做数据挖掘分析，采用的模式类似于 8.2.4 节中的 RFM 模型。采用这一模型的最主要目的是为了提高整体的客单价和利润。


图 8-8　UTC 营销邮件示意图之三

- 短期流失客户——老客户开发+数据挖掘

对于短期流失客户，UTC 行家邮件营销旨在通过对老客户

的关系维系，建设 UTC 行家产品口碑形象。通过对老客户的二次开发，提升其对 UTC 行家品牌产品的可信度及使用粘性，进而在市场营销层面上进行以老客户为基础的客户营销，最终促进销售。采取的方式除了常规的节假日、折扣促销邮件外，还同步采取游戏互动方式进行“刺激”营销，如图 8-9 所示。

对于短期失效客户的二次开发，UTC 行家也是采取了聚类数据挖掘算法，而考虑的关键属性有：失效前重复购买总次数、重复购买频率、平均客单价、购买总金额、是否晒过产品、是否进行过积分兑换、是否做过分享推荐等。通过这些属性的聚类，UTC 行家每个月会选择出约 40%的客户进行重点二次开发。


图 8-9　UTC 营销邮件示意图之四

- 长期流失客户——失效挽救+数据挖掘

该客户群已经较长时间没有上平台购买 UTC 行家的产品，较于前几类人群而言，相对较难二次开发成活跃客户。但若客户维系好，也可重新培养成为企业忠实的用户。因为之前已经尝试过短期流失客户上的各种邮件运营方式未果，这里做失效挽救采用的优惠力度要大很多，但为避免引起客户反感而退订，发送邮件的频次需要降低。

对于长期流失客户的二次开发，UTC 行家采取了和短期失效客户相同的聚类数据挖掘算法，而考虑的关键属性虽然与其类似，但是给如下的属性增加了权重：重复购买频率、平均客单价、是否进行过积分兑换、是否做过分享推荐等。

通过这些属性的聚类，每个月选择出约 20%的客户进行重点

二次开发。这时我们需要通过分析老客户的用户属性，合理安排个性化营销，以“转发推荐有奖、登录即免费赢取千元豪礼”等方式进行内容营销，同步配以电话回访。

### 8.2.3 邮件营销案例分享之二

我们先来了解一下这位客户的背景。沙×公司是一家知名的鲜花礼品服务商，通过网上销售和线下实体店铺相结合的方式为客户提供产品和服务，顾客不仅可以在实体店购买产品和预订服务，还可以通过公司网站了解产品信息，在线咨询客服，下订单并用网上支付购买所需产品，还可用网上的订单查询系统查看在线上和线下设定的商品订单。随着公司网上销售规模的不断扩大和在线营销效果的初步显现，公司已经把经营重点放到网上，互联网整合营销方案中的一个组成部分是以EDM邮件营销进行营销和产品推广。

我们来看一下解决方案是怎样的。

这次营销是通过国内一家知名的EDM邮件营销咨询公司操作的。该EDM邮件营销公司的资料库中已收集五千万个有效邮箱地址。至于该EDM营销公司所发送的邮件是否符合我们之前所说的正规邮件营销条件，我们暂时不加评论。

根据要求，对于该邮件营销公司的KPI分为三项，邮件发送至少需要100万封，到达率需要在99%以上，对于客户网站的引流需要在20000个独立访客以上。而这些营销邮件的发送可以在三个月内完成。其中引流到客户网站的统计是可以由客户自己来监测的。假设这次客户的总费用是20000元，而带来的独立访客是20132人，那么平均带来一个独立访客的费用约是0.993元，不到1元。作为对比，在百度搜索上做关键词排名，“鲜花”关键词的每次消费约在2.5元。通过EDM营销，客户可以节省60%的费用。

在第一周，该EDM营销咨询公司通过三种方式来发送营销邮件。我们来看下这三种方式进行邮件营销的效果差别。

（1）对邮箱均不加筛选，直接投递广告邮件。

（2）对邮箱地址进行初步过滤，筛选出曾打开任何一个广告营销邮件的邮箱地址投递与（1）相同创意的广告邮件。

（3）对邮箱地址进行进一步过滤，筛选出曾经打开过某个相关行业公司（礼品服务行业）广告邮件的邮箱地址投递与（1）相同创意的广告邮件。

邮件发送后，利用 EDM 营销系统对以上三种方式投递的邮件进行为期一个月的跟踪，通过相关指标来评估邮件营销的效果。具体来说，跟踪得到邮件的发送数、到达数、打开数、独立打开数、点击数、独立点击数进行系统评估。这里独立打开数不同于打开数是在于同一个用户打开多次邮件只算一次。同样的，独立点击数不同于点击数是在于同一个用户点击多次邮件中的链接也只算一次。

这里所有的邮箱地址都是有效的，所以到达率（邮件打开数除以邮件发送总数）都接近 100%，所以客户 KPI 的第二项是可以得到满足的。而三种方式的对比采用了同样的策划创意，所以在视觉效果上的差别也是不存在的。

我们来对效果做一下分析。在对三种方式进行一个月跟踪后，监测到邮件营销的效果如下：

在方式（1）下，不经任何筛选直接进行邮件营销时，营销的效果不明显。在随机发送了 154518 封邮件后，邮件独立打开仅有 1674 封，占比 1.8%，而邮件独立点击仅 375 次，占比 0.2%，邮件打开点击率相对比较低，如图 8-10 所示。

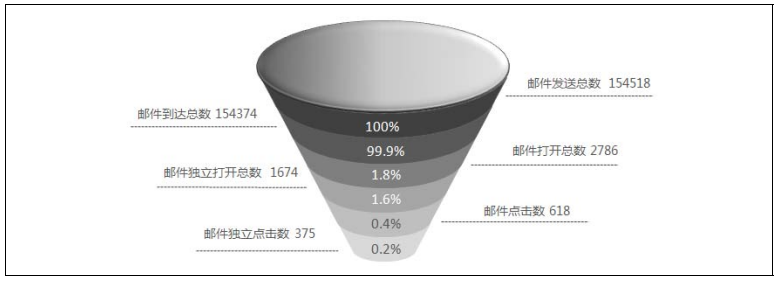



图 8-10　营销邮件漏斗示意图

采用方式（2），我们筛选客户，对曾经打开过广告邮件的客户进行邮件营销。这次每一个发送对象都曾经打开过至少一个广告营销邮件。营销效果和（1）相比出现了飞跃，邮件总计发送量 63470 封，而邮件独立打开数为 3024 封，点比 9.1%，邮件独立点击 651 次，占比 1%，邮件的打开点击率都提升了 5 倍。如图 8-11 所示。

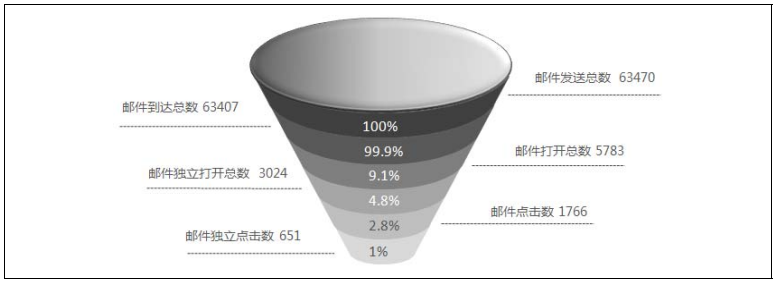



图 8-11　营销邮件漏斗示意图之二

而采用方式（3），对目标人群进行进一步筛选，这里的每一个客户都不仅打开过广告营销邮件，而且曾经打开过相关公司（礼品行业）的邮件。邮件总计发送量 34665 封，邮件独立打开数为 9621 封，占比达到惊人的 27.8%，邮件独立点击 1139 次，占比 3.3%，从结果来看，邮件打开点击率提升了 205%，而点击率，比方式（2）提高了 220%，是方式（1）的 15 倍，如图 8-12 所示。

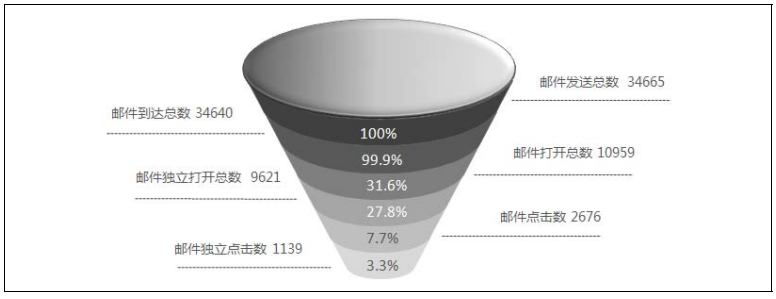



图 8-12　营销邮件漏斗示意图之三

我们来总结一下。

从这个案例中我们看到合理应用数据是多么重要。在 EDM

邮件营销中，在对邮箱地址进行一系列过滤操作后，营销效果可以有明显提升，点击率和转化率都能大幅提高。同时，过滤掉不必要的邮箱地址后邮件营销成本也可以大大降低。因为虽然邮件发送成本很低廉，但是如果要发送海量的邮件，占用的服务器和带宽资源还是需要一定费用的。如果说发送 1 万封邮件的成本是 1 元，那么少发送 1000 万封邮件就可节省 1000 元。

而这次活动也给该 EDM 营销咨询公司带来了额外的好处。这次给沙×公司的 EDM 营销执行结束之后，相关的数据会加入到邮件营销数据库中，会使下一轮的 EDM 效果会更加有效果和针对性。

### 8.2.4　运用数据挖掘 RFM 模型提高邮件营销效果

在 8.2.1 节中我们曾根据客户生命周期将客户分成潜在客户、首次购买客户、活跃客户、短期流失客户、长期流失客户，而且在8.2.2节中也以UTC为案例讲述了如何用此客户生命周期为模型做 EDM。这种分类算法只考虑一个因素——上次购买时间。比如可以把一定时间段购买过商品的客户都作为活跃客户，将连续五个月未发生购买行为的客户都作为长期流失客户等。

现实中客户的分类往往没有这么简单，我们需要考虑的因素有上次购买时间、购买频次、客单价等。假设我们需要对两个客户 A 和 B 发促销邮件，A 客户在一周内购买了 5 次，均价在 30 元，B 客户在一周内只购买了 2 次，均价在 80 元，如果从单一的购买频次分析，A 客户无疑应该是优先 B 客户发送的，也就是 A 的客户价值大于 B。但是如果结合了购买金额再看，B 的客户价值就会得到提升，B 虽然在一周内的购买频次低，但是他每次购买商品的金额都高，也就是 B 客户都挑贵的买。如此看来，相对于客户 A，B 能为公司带来的利润会更高。

当我们面对海量数据，考虑的因素多样时，仅考虑单一因素的分类算法就显得很不全面，而数据挖掘技术在此时就可以充分发挥它的作用。例如，我们可以用聚类的方式将客户进行细分，

邮件营销人员可以根据客户细分的结果对不同的客户采取不同的促销策略。而用分类的方法则可以根据海量的历史数据进行建模，理想的模型可以对促销邮件列表中的数据进行打分，分值较高的客户说明他们响应促销活动的可能性大一些。此时只要将分值进行降序排列，邮件营销人员从列表中选择排名靠前的客户发送促销邮件即可。

这里我们介绍邮件营销中最常用的一种数据模型，RFM 模型。

数据挖掘中的 RFM（Recency，Frequency，Monetary Value，即消费间隔时间、消费频率和消费金额）模型技术是根据消费者交易数据库中三个核心指标构建并计算的消费者细分或销售得分进行有针对性营销的一种市场研究技术。RFM 既是传统的数据库营销手段，也是数据挖掘技术关注的模型技术，RFM 在客户细分模型、客户响应模型、客户价值模型、客户促销模型等模型中都是重要的变量和分析模块。RFM 模型也是建构客户关系管理的核心分析技术。RFM 模型最早来自于美国数据库营销研究所 Arthur Hughes 的一项研究，Arthur Hughes 指出在客户的数据库中，有三个信息是最重要的，分别是消费间隔时间、消费频率和消费金额。其中消费间隔时间即参数 R（Recency）表示客户最近一次购买商品的时间与今天的间隔。消费频次即 F（Frequency）表示客户在一定时期内购买的次数。消费金额即 M（Monetary Value）表示客户在一定时期内每次购买的平均金额。

作为一种对客户分类的方法，RFM 分析模型起初主要用于直营行销（Direct Marketing）领域，目的是为了提高老客户交易的次数。下面具体介绍这 3 个参数的实际含义及其对邮件营销效果的作用。

- 消费间隔时间

消费间隔时间反映的是上一次购买与现在间隔时间的长短。一般来讲，上一次消费时间越近，顾客的客户价值越大，因为距离时间越近，消费者对公司产品的印象越深，也就越能对公司发

送的广告邮件中的描述产生共鸣，而且对公司的促销活动做出响应的可能性也越大。

- 消费频次

消费频次是顾客在一定时期内所购买商品的次数。顾客大量重复的购买公司产品，说明这些顾客对公司的产品认同度很高，也就是通常所说的品牌忠诚度很高。这类客户自然是邮件营销所重点关注的。对此类相对忠诚的客户，我们需要通过定时定期的广告邮件发送，让客户感觉到公司对他们的重视和诚意。这样至少可以保持或加深客户对公司的印象，不至于被竞争对手抢走。当然，也需要考虑一个度的问题，如果相同或者很类似的邮件发送过于频繁，可能会适得其反。

- 消费金额

衡量一个客户的价值最终还是要看他给公司带来了多少收入，所以消费金额是客户价值最直接的体现。帕累托规则告诉我们，一个公司 80%的收入都是由消费最多的 20%的客户贡献的，在任何的行业几乎都是这样。所以消费金额大的客户自然应该更加引起我们的关注。对于高价值客户，我们要尽量通过邮件营销把他们留住，所发送的邮件可能除了给予足够的优惠之外，还要考虑个性化。

消费间隔时间、消费频次、消费金额是测算客户价值最重要也是最容易的方法，这充分的表现了这三个参数对邮件营销活动的指导意义。在这三者之中，相对而言，我们通常认为最近一次消费是最有力的预测参数。

下面我们用一个案例来描述将 RFM 模型运用到邮件营销的具体做法。下面的数据来自一家欲进行邮件营销的服装公司。他们打算做一次折扣比较大的促销活动以唤醒一些高价值的长期流失客户。他们这次活动的目标不是提高销售额，而是找回这些曾经的忠实客户，所以他们需要选择发送的客户对象。他们对于长期流失客户的定义是 4 个月没有任何消费的客户。从该公司我们获取了他们全部的消费数据。在客户列表中详细记载了客户一

个月内购买次数、一个月内平均每次购买消费的金额以及距离上次购买产品的时间的间隔（月）。我们从长期失效客户中选取了1000名客户。这里列举前20位客户的样表，如表8-1所示。

表8-1　RFM示意表

| 消费频次 | 平均消费金额 | 消费间隔时间 |
|---|---|---|
| 9 | 59.9 | 4 |
| 10 | 157.99 | 4 |
| 10 | 168 | 4 |
| 10 | 217.89 | 4 |
| 9 | 54.9 | 5 |
| 10 | 157.99 | 5 |
| 10 | 157.99 | 5 |
| 10 | 168 | 5 |
| 10 | 168 | 5 |
| 10 | 217.89 | 5 |
| 10 | 217.89 | 5 |
| 9 | 348 | 5 |
| 7 | 39.92 | 6 |
| 7 | 40.92 | 6 |
| 7 | 41.92 | 6 |
| 9 | 59.9 | 6 |
| 9 | 59.9 | 6 |
| 8 | 59.9 | 6 |
| 8 | 59.9 | 6 |
| 8 | 59.9 | 6 |

表8-1中列出的客户应该来说都不错，因为他们的消费频次都比较高。当然，在一段时间内消费频次比较高但是最近几个月都没有购买也可能有多种原因，比如该服装公司有可能在几个月前一直都采用免邮费的促销方式来吸引用户多次购买，而最近的几个月取消了这一政策。在RFM模型的实际应用中，由于具体

的场景和产品不同，往往会事先对这三个参数的权重进行调整，比如某公司会认为消费间隔时间比消费频次更重要，此时就需要把消费间隔时间的权重设高一点。权重的具体设定方法可以采用由专家参与的层次分析法。在此案例中没有对参数权重做调整，所以三个参数的权重是相同的。

我们先通过 K-means 聚类算法对原表中 1000 位客户进行聚类。K-means 聚类从空间距离的角度出发同时考虑三个参数，将距离最近的客户划分为同类。关于 K-means 聚类算法请参照 4.2 节。划分结果如表 8-2 所示。

表 8-2　对客户进行 K-means 聚类结果表

| 类别 | 消费频次 | 平均消费金额 | 消费间隔时间 | 该类别用户数 |
|---|---|---|---|---|
| 1 | 6.377778 | 654.92286 | 8.666667 | 180 |
| 2 | 2.714286 | 42.18667 | 15.928571 | 212 |
| 3 | 5.391892 | 139.62784 | 9.905405 | 198 |
| 4 | 4.508475 | 56.30254 | 9.677966 | 209 |
| 5 | 4.892857 | 292.10762 | 8.309524 | 201 |

从表 8-2 中可以看出，K-means 聚类将 1000 个客户分成了 5 类，表中的具体数据表示各类别相应参数的均值，比如第一行第一列的 6.377778 就表示第一个类别的客户在一个月内消费频次的平均值是 6.377778 次，而平均消费金额是 654.92286 元。

从表中可以看出第一个类别无论是消费频次还是平均消费金额都是最大的，消费间隔时间都在 8 个月左右，这类客户无疑是重点关注的，因此在给这类客户发送广告邮件时应适当的采取一些相对最优惠的措施，因为他们对于公司的长期效益会很有帮助。仅从平均消费金额来看，第 3 类和第 5 类客户的表现是很接近的，而且也比较高，所以这两类客户也应当是我们关注的客户，如果能够从长期流失客户群中重新转化成活跃客户，提供的价值也比较高，所以我们可以给他们也发送相对优惠的广告邮件。

剩下的第 2 类和第 4 类客户平均消费金额偏低，虽然消费频次不算低，但是和其他三类客户相比频次也是偏低的。特别是第

2 类客户，消费频次是最低的，而且消费间隔时间也最长，所以在促销资源有限的情况下，这类客户应当从我们的邮件客户列表中舍去。

综上所述，我们应该将这次优惠活动的邮件重点发送给第 1、3、5 类客户，这类客户总共有 579 位，而其余的 421 位客户将不会接收到这次的促销邮件。这样的选择势必会比一股脑的全部发送 1000 份促销邮件或只考虑单一因素的方式能更大的提高邮件营销效果。

我们在 2012 年春节前还帮多家电子商务企业做了类似基于如上的 RFM 数据挖掘。用同样的方法对于这些公司的 CRM 数据进行处理，我们从中找出两类老客户：

- 给在一段时间内没有购买行为的高价值客户（类似表 8-2 中的第 3 和第 5 类客户）发送唤醒邮件，推荐他们之前购买商品的关联产品，并在这个产品上送出大约五折的优惠券。
- 罗列出了近期消费最频繁、消费额最高的客户（类似表 8-2 中的第 1 类客户），然后给他们发送节日问候邮件和价值 50 到 500 元不等的优惠券。

最后的效果我们在这里就不赘述了。

## 8.3 数据挖掘和垃圾邮件过滤

当我们介绍完邮件营销后，下面来看下邮件营销这把双刃剑的另一面，过滤垃圾邮件的几种方法，以及数据挖掘技术在这方面的运用。每天打开邮箱，看到的都是些让我们头痛的垃圾邮件。作为一种日益盛行的网络营销方式，垃圾邮件让我们苦不堪言。有关数据显示，中国网民每年收到的电子邮件中有 60%以上为垃圾邮件，严重影响到正常的通信活动。

## 8.3.1　垃圾邮件

自从《中国互联网协会反垃圾邮件规范》出台以来，国内对反垃圾邮件的研究和运用进入一个高峰时期，政府、高校、企业、电信运营商等纷纷部署反垃圾邮件产品。反垃圾邮件产品已经成为网络安全防护必不可少的重要一环。

图 8-13 是中国互联网协会反垃圾信息中心发布的 2011 年第一季度至 2012 年第一季度用户每周收到的垃圾邮件状况，从图中可以看出用户在 2012 年第一季度收到的垃圾邮件封数和比例与上季度相比没有明显变化（分别增加 0.3 封和 0.1%），但垃圾邮件的比例同比 2011 年第一季度的 40.1%下降了 4.5%，由此可以看出垃圾邮件的治理在一定程度上取得了一定的效果。但是 35.6%的垃圾邮件占比说明反垃圾邮件这一使命依然是任重道远的。

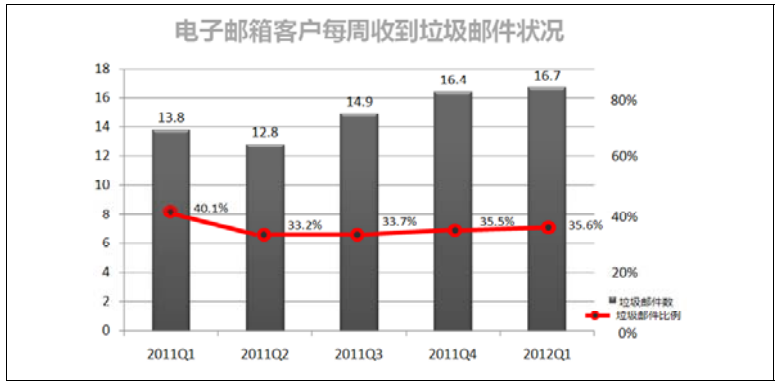



图 8-13　垃圾邮件变化示意图

当然另外一个问题是垃圾邮件误判。比如你的签名中如果有网址，或者内容中不小心某些关键词出现的频次高了些，那么被误判成垃圾邮件的概率会大大提升。

## 8.3.2　垃圾邮件过滤技术

在垃圾邮件过滤上有很多技术可以使用，有不少是基于数据

挖掘算法的。我们先来看一个基于贝叶斯过滤法的垃圾邮件过滤技术。

#### 8.3.2.1 基于贝叶斯过滤法的分类技术

我们在第 4 章数据挖掘基本原理和算法一章中提到过朴素贝叶斯算法，而在这里用到的过滤法也是基于贝叶斯算法的，因为过滤本身就是把事件分类成“该过滤”和“不该过滤”两类。被广泛运用的贝叶斯过滤法来源于 18 世纪著名数学家托马斯贝叶斯创建的贝叶斯理论，该理论的核心是通过对过去事件的分析，对未来将发生的事件做一个概率性的推断。应用到垃圾邮件过滤上则是通过对大量已经判定的垃圾邮件和正常邮件进行学习，根据两种邮件中相同词语出现的概率对比来确定垃圾邮件的可能性。贝叶斯过滤法的优点是可以通过自主地学习来适应垃圾邮件的新规则。该方法是目前过滤垃圾邮件最为精确也是最为普遍的技术之一，准确率可以达到 99%，缺点则是此方法的成功实施需要大量的历史数据。

贝叶斯过滤法其实运用的就是数据挖掘中分类算法的一种，当然也有很多实际应用围绕诸如支持向量机、KNN 等其他分类算法展开，但由于朴素贝叶斯算法的众多优点，该方法目前仍然是首选的。我们简要介绍一些涉及的理论。

我们先来看一个贝叶斯定理的公式：

$$p(A \mid B) = \frac{p(B \mid A)}{p(B)} \times p(A)$$

现实生活中常常会涉及对某个事件发生的可能性作出判断，如果我们在判断之前，有一些辅助信息的存在，必定会使我们的判断精确性有所提高。反映到这个公式中，$p(A \mid B)$ 表示在事件 B 发生的条件下事件 A 发生的概率，常被称作是后验概率，而 $p(A)$ 则被称为先验概率，即不考虑事件 B，事件 A 单独发生的概率。由于加上了 $\frac{p(B \mid A)}{p(B)}$，后验概率相当于是对先验概率的一

个修正，所以 $\frac{p(B|A)}{p(B)}$ 又可称之为修正因子。由于贝叶斯定理增加了其他事件的发生对预测事件影响的辅助信息，所以预测事件的预测效果会得到显著提高。

整个朴素贝叶斯分类法都是基于贝叶斯定理进行的，假设现在有 n 个邮件用于训练，每个邮件有 n 个特征项（词）$(w_1, w_2, \cdots, w_n)$，给定的类 $c_k$ (k=1, 2, $\cdots$, m)，那么当一个邮件由特征项集 d 构成时，该邮件被判为类 $c_k$ 的概率就为：

$$p(c_k|d) = \frac{p(d|c_k)p(c_k)}{p(d)} \text{ (k=1, 2, } \cdots \text{, m)}$$

其中：

$$p(d|c_k) = p(w_1, w_2, \cdots, w_n|c_k)$$

我们现在要比较的是邮件属于哪个类别的概率大，也就是 $p(c_k|d)$ 的大小，在公式当中，分母 p(d)与邮件类别无关，也就是说要比较邮件属于哪个类别的概率，只需计算概率 $p(c_k)$ 和 $p(d|c_k)$。$p(c_k)$ 为先验概率也就是各个邮件类别出现的概率，这个很容易计算，难点是 $p(d|c_k)$ 的计算。为了简化计算，假定各特征项之间是相互独立的，这也是朴素贝叶斯算法之所以“朴素”的原因。有了这个假设，就可以通过下列公式求出 $p(d|c_k)$。

$$p(d|c_k) = p(w_1, w_2, \cdots, w_n|c_k) = \prod_{i=1}^{n} p(w_i|c_k)$$

我们在下一节的垃圾邮件案例中会看到贝叶斯算法的一个应用实例。

#### 8.3.2.2　机器学习领域的 Winnow 算法

除了贝叶斯过滤法之外，机器学习领域中的 Winnow 算法也是比较常用的垃圾邮件过滤方法。Winnow 算法是一个简单的用来根据标记样本学习分类器的算法，对于垃圾邮件过滤有一定的效果。我们来看下 Winnow 算法的执行过程。

在算法执行之前，我们根据经验设定一个阈值 $\theta$，然后对于

测试集中的每一个数据，做以下的训练：

第一步，获取特征空间大小 $n$，并对 $i$ 在 1 到 $n$ 之间所有的权重初始化权重向量 $w_i = 1$。

第二步，训练调整权重向量 $w_i$，具体如下：

- 如果分类正确，那么不做任何动作；
- 若邮件不属于垃圾邮件的类别，但 $\sum_{i=1}^{n} w_i x_i > \theta$，则降低权重 $w_i$，可重置 $x_i \neq 0$ 对应的 $w_i = \alpha w_i \ (0 < \alpha < 1)$，反复调整直到满足 $\sum_{i=1}^{n} w_i x_i < \theta$。这个过程称为降权过程。
- 若邮件属于垃圾邮件的类别，但 $\sum_{i=1}^{n} w_i x_i < \theta$，则提高权重 $w_i$，可重置 $x_i \neq 0$ 对应的 $w_i = \beta w_i \ (\beta > 1)$，反复调整直到 $\sum_{i=1}^{n} w_i x_i > \theta$。这个过程称为升权过程。

第三步，测试数据集，将数据集中得到的 $x_i = (x_1, x_2, x_3, \cdots, x_n)$ 与训练得到的 $w_i (= 1, 2, \cdots, n)$ 求加权和 $\sum_{i=1}^{n} w_i x_i$，若 $> \theta$ 则 $x_i$ 为垃圾邮件。

基于 Winnow 算法还衍生出 Balanced Winnow 算法，在一定程度上可以进一步提高分类效率。

#### 8.3.2.3 基于搜索引擎的算法

另外我们可以借助搜索引擎算法中的 Hilltop 算法结合第 4 章中的 PageRank 算法，帮助我们筛选出更接近的垃圾邮件判别结果。用这个算法我们基本假设和出发点是邮件内容中的相关链接所指向的页面或站点，相对于网络中的权威（Expert）页面和权威站点的价值权重度的高低，越高则为垃圾邮件的可能越低。

与此类似的 SpamRank 方法与之思路接近，但是算法正好相反。基本假设和出发点是邮件内容中的相关链接所指向的页面或站点与已知 Spam（垃圾）站点或页面的相关度权重，权重越高

则为垃圾邮件的可能性也越高。

Hilltop 算法和 SpamRank 算法的思路很直观，是根据邮件内链接的质量而来。可想而知，如果某个邮件中的链接指向的是一个 PageRank 为 7 的网站，那么该邮件是垃圾邮件的概率比较低，而如果其中的链接指向的是一个 PageRank 为 0 的网站，那么它是垃圾邮件的概率相对会比较高一些。

我们再来看一下其他一些常用的垃圾邮件过滤方法。

#### 8.3.2.4　黑白名单技术

黑名单（Black List）和白名单（White List）分别指已知的垃圾邮件发送者和可信任发送者的 IP 地址或邮件地址。黑名单技术是早期出现的垃圾邮件过滤技术，它的原理是先确定垃圾邮件制造者及其 ISP 的域名或 IP 地址、电子邮件地址，将这些资料整理成黑名单，再将黑名单部署在处理网关处，凡是列表中的邮件地址发来的邮件一律拒收。而其他邮件地址发来的邮件则被正常接收。与此相对应，白名单的原理则是只有列表中的邮件地址才被正常接收，而其他地址发来的邮件则会被拒收。

该技术的优点是不占用计算机资源，易于实施；缺点是需要不断维护黑白名单。另外由于现在的垃圾邮件发送者很容易就可以修改和伪造 IP 地址和邮件地址，因此这种技术在总体的垃圾邮件解决方案中只能起到辅助作用。

CBL（China Black List，中国垃圾邮件黑名单）是采集并分析整理的当前的垃圾邮件源，该地址属于恶意或无意的垃圾邮件来源，来自它的邮件属于垃圾邮件的可能性极大。CBL 主要面向中国国内的垃圾邮件情况，所甄选的黑名单地址也以国内的垃圾邮件反馈情况为主。向 CBL 提交邮箱或者申诉的网址是 http://anti-spam.org.cn/CID/1。

#### 8.3.2.5　其他垃圾邮件过滤技术

还有其他一些垃圾邮件过滤技术，比如：

- 反向域名验证

域名的反向验证，就是将 IP 地址转换为对应的域名，是域名解析的逆过程。这种方法已经成为目前大多数 E-mail 服务器识别垃圾邮件的重要标准之一，如果经过解析的域与邮件上的来源 IP 地址相符合，该邮件被接受。如果不符合，该邮件将被定位为垃圾邮件而抛弃。

反向域名解析的缺点是对于实际存在的域名，或者是通过跳板、SMTP 劫持等方式产生的垃圾邮件，反向解析无法产生效果，从而造成很高的误判率。

- 关键词过滤

关键词过滤则是一种基于内容检查的过滤技术，通常会创建一些与垃圾邮件关联的关键词表来识别垃圾邮件，比如类似“免费”、“促销”、“色情”等词语经常会出现在垃圾邮件中，通过对发送邮件的内容进行排查，如果邮件中包含这些单词或频繁出现这些单词，就会被判定为垃圾邮件。这是一种相对简单的过滤方式来处理垃圾邮件，它的前提是必须要创建一个庞大的且时时更新的垃圾邮件关键词列表。

该方法在现实中运用的效果越来越差，原因是过滤的能力与关键词列表有直接联系，光是关键词列表本身所具有的局限性就会造成高的漏报率，同时垃圾邮件发送者经常会有意躲避关键词的运用，例如通过拆词、组词、避免敏感词汇、增加图的方式绕过词语过滤器，这又会造成过滤器大量错报的发生。

- 基于规则评分的过滤技术

该技术的核心思想是通过对海量数据的挖掘，定义一系列符合垃圾邮件的规则，每一条规则对应一个评分。当判定一封邮件是否是垃圾邮件时，就将该邮件与规则库进行比较并设定一个最大评分值，邮件每符合一条规则就加上该规则评分，获得的分数越高，该邮件是垃圾邮件的可能性就越高。如果一封邮件超过最大评分值，该邮件将被分类为垃圾邮件。

该方法的缺点是要持续追踪垃圾邮件的规则变化，不断更新规则库。在现实中，这项技术可以清除 90%的垃圾邮件。

我们下面通过一个实例来看应如何进行垃圾邮件过滤的实际操作。

### 8.3.3　垃圾邮件过滤案例

下面结合一个公用垃圾邮件数据集简要探讨如何运用朴素贝叶斯算法进行垃圾邮件过滤，该垃圾邮件数据集的下载地址为：http://spamassassin.apache.org/publiccorpus/，需要说明的是本章在进行垃圾邮件分类时对这个数据集进行了简化并只采用了部分数据文件。

这里一共分成 4 步来操作。

（1）邮件预处理

对垃圾邮件的分类不同于常规的数据挖掘分类，传统的数据挖掘所处理的特征一般也就在几十或上百个，但是邮件分类的特征可能高达上千甚至上万。所以这项技术首先要解决的就是如何提取出文本特征，通俗地说就是从众多邮件文本内容中提取出能够反映邮件类别的词。无论是对中文还是对英文邮件，首先要做的就是消除一些无关紧要的标点和一些没有实际意义的词（停词），例如汉语中的“的”、“是”等。然后就是提取特征词，这对于英文来说相对容易一些，因为英文的书写习惯是每个单词之间都有空格，也就是很容易就能区分出单个词语。对于中文则需要通过分词来对词语进行切分，再组成特征项集合。

下面以垃圾邮件数据集中的一个垃圾邮件作参考，该垃圾邮件的原文如下：

```
"Lowest rates available for term life insurance!
Take a moment and fill out our online form to see the low rate you qualify for.
Save up to 70% from regular rates! Smokers accepted,
 Representing quality nationwide carriers. Act now!"
```

按照邮件预处理的规则我们去掉邮件中的标点、停词、数字，并将大写转化为小写，得出的特征项集合如下：

```
lowest rates available term life insurance moment fill
online form low rate qualify save regular rates smokers
accepted representing quality nationwide carriers act
```

（2）构建布尔或词频矩阵

当对邮件数据进行预处理以后，下一步需要做的就是构建布尔或词频矩阵，布尔矩阵是数学中很常见的矩阵，矩阵中所有的值都只由简单的 0 和 1 组成，衍生到邮件分类，则矩阵的行表示的是邮件名，列表示的是特征项，如果邮件中包含了这个词，取值为 1，反之则为 0。而词频矩阵的含义与布尔矩阵一样，所不同的在于矩阵具体的值是词在邮件中出现的频数。

如表 8-3 所示是该数据集中垃圾邮件的词频矩阵（部分）。

表 8-3　垃圾邮件的词频矩阵（部分）

| 文档序号 / 特片词 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| absolutely | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| access | 1 | 0 | 0 | 2 | 0 | 0 | 0 |
| accessories | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| account | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| adult | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| advertisement | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| advertisers | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| advice | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| affiliated | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| affordable | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| again | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

在该矩阵中，行表示的是出现的特征词，列则是相应的文档，从这个矩阵中可以看出某个特征词在相应的文档中出现的次数，例如从表中可以看出“absolutely”在第 4 个文档出现了 2 次，“access”在第一个文档出现了 1 次，在第 4 个文档出现了 2 次。

（3）计算词在各类邮件中出现的概率

上面曾提到，要想求出邮件属于何种类别时，必须假设各个特征项之间是独立的，再利用下列公式求出 $p(d \mid c_k)$。

$$p(d \mid c_k) = p(w_1, w_2, \cdots, w_n \mid c_k) = \prod_{i=1}^{n} p(w_i \mid c_k)$$

从公式中可以看出，只要将邮件中各个特征词在各类别邮件出现的概率相乘即可，这点通过布尔矩阵可以很容易得到。这里的特征词出现的概率并不是指特征词在文档出现的频数与该特征词出现的总频数之间的比值而是该特征词在该类邮件中的出现率，例如单就上面出现的词频矩阵来说，“absolutely”在垃圾邮件出现的概率应该是 1/7，而不是 1，“access”的概率是 2/7。

表 8-4 为该数据集垃圾邮件中出现概率最大排名前十的特征词。

表 8-4　排名前十的特征词

| 词 | 发生概率 |
|---|---|
| Html | 0.338 |
| Body | 0.298 |
| Table | 0.284 |
| E-mail | 0.262 |
| Font | 0.262 |
| Head | 0.246 |
| Free | 0.202 |
| List | 0.202 |
| Please | 0.188 |
| Receive | 0.188 |

（4）最终分类

当我们做到这一步时，对于一个新的未分类邮件，只需计算该邮件包含的特征项（词）在各类别出现的概率并将其相乘即可。最终由贝叶斯定理，我们可以比较这个邮件是垃圾邮件或正常邮件的概率，若垃圾邮件的概率大，就将邮件判为垃圾邮件。

在现实应用中，垃圾邮件过滤技术依然是个比较复杂的领域，尤其是当涉及中文邮件时，本章仅仅阐述朴素贝叶斯算法以及通过一个简单案例来阐述方法运用的大体过程，以求能达到让

读者了解的目的。

谷歌、雅虎和网易的企业邮箱都采用了不同的反垃圾邮箱技术。而盘石公司自主开发的企业邮箱，磐邮 http://www.upanshi.com 的垃圾邮件过滤技术采用了 8.3.2 节中的多种不同方法，包括朴素贝叶斯算法。

## 8.4 本章相关资源

- 本章相关参考文献：

[1] Nick Littlestone. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. Machine Learning 285–318(2), 1988.

[2] Nick Littlestone. *Mistake bounds an d l ogarithmic linear-threshold l earning algorithms*. Technical report UCSC-CRL-89-11, University of California, Santa Cruz, 1989.

[3] Andras A. Benczur, Karoly Csalogany, Tamas Sarlos, Mate Uher, Máté Uher，SpamRank - Fully Automatic Link Spam Detection, In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[4] *K-means Clustering Versus Validation Mea sures: A Da ta Distribution Perspective* . IEEE Transactions on Systems, Man, and Cybernetics --- Part B (TSMCB), Vol. 39, No. 2. (2009), pp. 318-331, 2009.

- 本章相关网站：

[1] http://spamassassin.apache.org/publiccorpus/

[2] http://anti-spam.org.cn/

# 第 9 章
# 数据挖掘和互联网广告

本书中提及的“大数据”，指的是收集、整理互联网中某一领域的海量相关数据。大数据需要通过数据共享、模式分析、数据挖掘来获取最大的数据价值。面对大数据不能仅仅停留在表面，我们需要提出解决问题的方法，并对其进行分析挖掘，进而从中获得有价值的信息，最终产生商业价值。我们在本章中要看“大数据”在互联网广告领域的应用。一方面讲述互联网广告如何利用大数据提升广告效果，另一方面会介绍如何用数据挖掘方法抓出广告作弊行为。

在 9.4 节中，将以一个网站联盟广告公司为实例，展示在网络广告上如何做数据分析和数据挖掘，同时也会介绍如何应对广告作弊行为。

## 9.1 互联网广告

作为广告的一种新形式，互联网广告在过去的 10 年发展远超其他的广告形式。2010 年在美国，互联网广告市场就以超过 250 亿美元的规模超越报纸等其他传统媒体，成为继电视类媒体之后的第二大广告载体。并且发展趋于成熟，越来越稳定。而相对来看，其他的广告形式，户外、电视、广播、纸媒（杂志和报纸）等或是持平，或是有一定程度的下滑。在美国和欧洲比较发达的地区，由于经济发展不好，广告市场的整体盘子没有增加，而纸媒下降的趋势非常明显，下降的这部分都加在互联网广告上

了。即使涵盖 2008 年美国金融危机，互联网广告的每年增长率也在 15%以上。

在发展中国家我们也看到类似的趋势。图 9-1 是 MarketingCharts.com 制作的关于全球各种媒体上广告投放占比的示意图。我们可以看到互联网广告在全球广告市场中的占比从 2010 年的 14.4%到 2014 年会增长到约 21%，而对应的报纸作为广告媒体所占的份额正好相反，会从 2010 年 21%的份额下降到 2014 年的 17%。

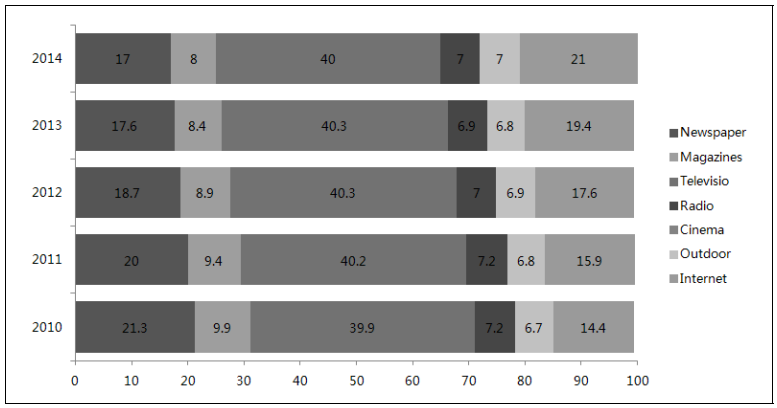



图 9-1　来自 Zenith Optimedia 的全球广告支出变化图

大部分互联网公司的收入有相当比例来自于互联网广告，比如巨无霸 Google，2012 年 97%的收入来自互联网广告，主要来自搜索页面上的 AdWords 关键字竞价和谷歌网盟的 AdSense。而中国的百度，2011 年 99.6%的收入来自于互联网广告。所以我们说 Google 和百度也可以算是互联网广告公司。互联网上的新贵们，脸谱网 Facebook，推特网 Twitter 和 2012 年最热的针趣网 Pinterest，绝大部分收入也是来自于互联网广告，特别是 Google 在 2011 年广告收入高达 365 亿美元。

我们来看两个在报纸和杂志上投放广告费用的大概数据。

在图 9-2 中我们可以看到，假设每一个收到报纸和周刊的人都看到客户投放的广告，每千次展现成本大概分别在 5.78 元和 28.01 元，而且并不一定能保证用户一定会看到这一广告。而在

互联网上投放广告，费用只是这一数字的几分之一、几十分之一甚至几百分之一。

| 媒体 | 发行量（册） | 每周广告投入（元） | 全年广告投入（元） | 千次展现成本（元/千次） |
|---|---|---|---|---|
| XX晚报 | 660000 | 8000 | 38400 | 5.78 |
| XX周刊 | 25000 | 7000 | 336000 | 28.01 |

图 9-2　纸媒广告支出计算示意图

图 9-3 是艾瑞咨询公司根据中国企业公开财务报告、行业访谈预测出的中国互联网广告数据。在中国，过去的几年，每个季度都有超过环比 40%以上的增长。

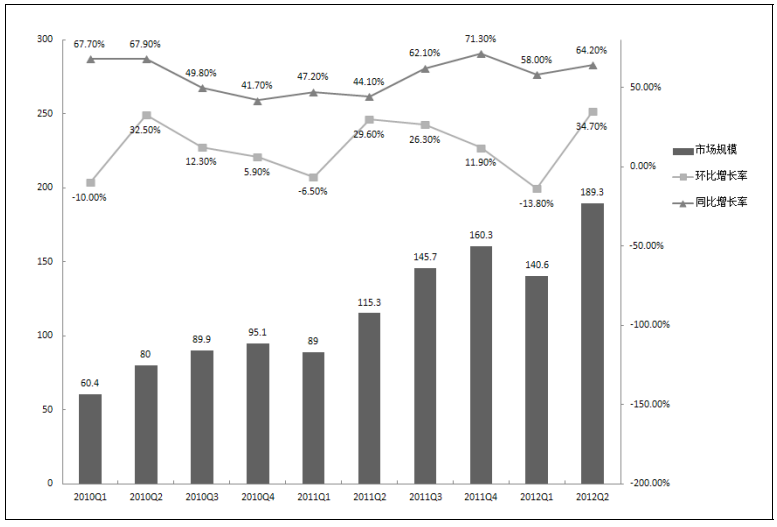



图 9-3　艾瑞咨询公司提供的中国网络广告市场规模

和传统广告相比，网络广告有着无可比拟的优势，下面详细讲解：

（1）覆盖面广，传播速度快

网络广告的传播范围广泛，可以通过互联网络的渠道把广告信息全天候、24 小时不间断地传播到世界各地。目前全球网民已接近 20 亿，中国也超过了 5 亿，并且这些数字还每年成几何倍数的速度不断发展壮大。这些网民是网络广告的受众，他们可以在互联网上随时随意浏览广告信息 ，其传播的范围和速度是

显而易见的。这些效果，传统媒体是无法达到的。

（2）形式多样，交互性好

网络广告的载体有文字、图片、视频等各种形式，只要受众对某样产品、某个企业感兴趣，仅需轻按鼠标就能进一步了解更多、更详细、更生动的信息，从而使消费者能亲身“体验”产品、服务与品牌。如能将虚拟现实等新技术应用到网络广告，让顾客如身临其境般感受商品或服务，将大大增强网络广告的实效。在传统媒体上做广告，发版后较难更改，即使可改动往往也需付出很大的经济代价。而在互联网上做广告能按照需要及时变更广告内容、及时改正错误。这样，经营决策的变化也能及时实施和推广。

（3）性价比高

各个媒体的优势各有所长，但是作为网络媒体却可以给所有媒体一个强有力的互补。比如说企业在电视媒体投放了广告，但是仅有短短的几秒钟，用户可能只是了解了这家企业的品牌名称，却无从了解更多的信息，一个意向客户很可能就这样流失了。但如果这家企业同时也做了网络广告，那么这个意向客户就可以通过网络广告得到更多的信息，从而真正意义上的成为这家企业的合作伙伴。

（4）效果可评估

网络广告可通过人群定向、地域定向和行为定向达到真正的精准投放，同时可以通过广告后台准确分析广告了解到有多少人看了广告，有多少人通过广告访问了企业网站，有多少人留言，有多少人有购买意向，让企业真正的掌握了主动权。

在人体的感觉器官中，眼睛接受信息的比例占 70%，而在互联网上通过眼睛传递的信息高达 95%。所以广告的文字和图片做的质量好坏，会直接影响到广告的效果。广告的主题是否明确，图片选择布局是否美观，文案是否有吸引力以及色彩搭配是否恰当，在一定程度上决定了广告的效果，如图 9-4 所示。但是什么样的广告形式和投放方式是最优的，这些就要靠数据来说话了。


图 9-4　网络广告示意图

互联网广告的精准性是毋庸置疑的，特别是如果我们能够通过社会化媒体上的信息和用户行为分析找到客户的具体信息和兴趣喜好等，与此相关的隐私问题我们会在 11.3 节中专门提及。如果我们发现一个用户把他在 LinkedIn 主页上在某个公司的职位从主管改成经理，那么可以推断他刚被提升，这也意味着他的薪水可能有一定程度的提升。这时给他推送新车的折扣券或者某种奢侈品的广告，效果可能会不错。如果我们发现一个用户把他在 Facebook 的主页上的个人状态由单身转变成在恋爱中，那么我们可以给他发送双人旅游的优惠券或者是他附近地区买一送一的餐馆折扣，被点击的概率会比较高。如果我们发现一个用户在论坛上抨击苹果公司 iPhone5 上的地图功能，我们可以适时发送广告推荐三星公司最新款手机。这样的例子还可以举出很多。总而言之，互联网广告是所有广告形式里唯一能够精准定位客户的广告形式，而到底能有多精准，则是依赖于我们手里的数据。

## 9.2 广告作弊行为

互联网广告最大的敌人就是作弊。从 Facebook 走上市流程开始，关于网络广告的负面消息就接连不断。继 GM（美国通用汽车公司）结束和 Facebook 的广告合作之后，在网上发布音乐和推广歌手的数字媒体公司 Limited Press 在 2012 年 7 月宣布将要结

束与Facebook的广告合作关系。Limited Press，现在更名为Limited Run，宣称他们公司在Facebook上80%的广告点击均来自于外挂程序，使得他们获取真实用户的广告费用远超预算。他们在Facebook的公司官方网页上披露了公司计划删除在Facebook上的品牌页面，因为公司发现只有20%的广告点击来自于真实的Facebook用户，剩下的全部来自于外挂程序，也就是我们熟知的机器人程序。这家公司使用了6款分析工具再加上自己开发的分析工具对广告来源进行追踪，结果发现外挂程序点击占了多数。公司发言人表示，“在我们进行广告系统测试时，我们发现了一些非常奇怪的事，Facebook会向我们收取广告点击费，但是我们确定只有约20%的广告是来自于正常的用户点击。”在这之后，Facebook做出了反应，并表示通过他们的系统分析没有发现Limited Run所说的机器人程序的迹象。Limited Run公司在网上说，他们的分析软件需要浏览器开放JavaScript来分析客户来源及属性，但是发现80%的广告点击是来自于那些在客户端关闭了JavaScript的用户，使点击来源无从得知。公司又称，按照公司员工的经验，正常情况下，约1%～2%的点击会来自于关闭了JavaScript的用户，而现在超过80%的点击用户，此行为一定是不正常的。

我们且不说Limited Run做的判断是否合理，也不推断是谁一定想要以此得利，不过在网络上点击作弊的行为确实是一个普遍现象，这使广告效果大打折扣。除了Facebook，在互联网广告领域的几大商业巨头，包括谷歌、雅虎、百度都也曾因为类似事情而被质疑。

特别对于广告联盟这种新兴的互联网广告模式，广告主通常是按点击收费（CPC），而不是按展现收费（CPM）。点击次数越多，广告主花费越多，而展示广告的网站主也收益更多。在互联网广告联盟PPC（Pay Per Click，按点击付费）的商业模式下，点击作弊是指为了谋取自身利益，采取不正当的手段对广告主投放的广告进行恶意点击。如果存在大量的虚假点击，效果会大打折扣。

对于广告主来说，如果有 80%的量是作弊，那么没有作弊的理想效果应该是现在他所看到的实际效果的 5 倍，也就是说他的投资回报率可以高 5 倍。可以说，点击作弊已经成为悬挂在整个互联网广告行业上的“达摩克利斯之剑”。由于利益争夺相对更为激烈和该行业本身的脆弱性，使识别广告联盟上的点击作弊更为重要。

## 9.3 网站联盟广告

我们来看一下互联网广告联盟 PPC（Pay Per Click，按点击付费）的商业模式。在网站联盟中，有三个主体：需要做广告的广告主、提供广告位的网站主和承载广告运作的广告联盟平台。而网站联盟的大致运营过程有以下四个步骤：

（1）广告主登录联盟平台，将自己的广告代码、广告创意、需投放的广告类型以及对于网站类型的要求和对广告位的要求写入联盟平台，在这个过程中需预先支付投放的佣金。

（2）联盟平台对广告主的广告进行审核，将审核通过的广告挂在符合要求的前台页面上。

（3）网站主登录联盟平台，在平台上通过自动过滤和行业内容以及形式匹配，登记合适的广告位，并将广告投放代码放置在自己的网站上。

（4）正式投放开始，这时如果有用户点击网站主上广告主的某一个广告，就有费用产生了。

收费方式有 CPD（按时间段付费）、CPM（按千次展现付费）、CPC（按点击付费）和 CPA（按最终效果付费）等多种。不过在众多广告联盟中，CPC 是最常用的方式，谷歌的 Google AdSense，百度网盟和盘石网盟采用的都是 CPC 方式收费。在 CPC 的模式下，一旦用户点击了广告主的广告，广告主就需支付报酬，联盟平台按照事先制定的提成比例把这份报酬的一部分分给网站主。而如果只有展示，没有点击，那么是没有任何费用发生的。

点击作弊是指为了谋取自身利益，采取不正当的手段对广告主投放的广告进行恶意点击。由上述的网盟运营过程我们可以看出，作为投放广告的广告主本身，大量的欺诈点击既不能带来最终的转化又会引起额外的成本，不可能参与作弊。联盟平台作为整个广告联盟的构建者，点击作弊现象泛滥会使联盟的信用大幅下滑、联盟参与者锐减，而且整个作弊识别的任务都是联盟平台在做，作弊可能性也不大。因此最有可能发生点击作弊的就是网站主和广告主的竞争者了。

对于网站主来说，点击作弊可以带来丰厚的额外收益，如果没有有效的作弊识别方法和严厉的惩罚措施，网站主可以肆无忌惮地进行作弊。而广告主的竞争者可以通过点击作弊来达到虚耗广告主的广告费用，减弱对方的竞争力的目的，也有一定的作弊的动机。不过，由于网站联盟的海量广告位和随机展现，广告主的竞争对手很难找到对应的广告，所以相对有一定难度。

我们在下一节会以网站联盟广告为例，讲述怎样通过数据挖掘方式提高广告的转化率和应对网站联盟上的广告作弊行为。

## 9.4 网站联盟广告上的数据挖掘

在网站联盟广告上存在大量数据，再加上联盟网站上用户的访问信息，每天都会产生海量的数据。

通过类似于第 7 章中提及的网站日志分析，我们可以掌握到很多与网站和访客相关的信息。再进一步分析访客在网站主和访客点击广告的后续行为，我们可以对访客的属性，包括年龄、性别、学历、收入、籍贯和兴趣爱好等各种信息作出大致的判断。

访客属性的判断对于每个人不是 100%准确，但是我们做数据挖掘本来就是在统计学的范畴之上的。如果一个判断的准确度在 75%，那么我们可以认为这个判断做的还是比较准的。如果在 90%的情况下是正确的，那么我们可以认为这个判断是相当精准的。

### 9.4.1　数据助力网盟广告

网站联盟广告本身包含了大量的数据，包括所有的网站内容信息、行业、领域、每天的平均访问量、Alexa 排名、展示的广告内容、广告整体展示次数、广告点击次数、访客信息等。而对于点击之后的用户行为分析，我们还要有更多的信息，包括跳出率、二跳率、活跃时间、停留时间、转化率等。

#### 9.4.1.1　通过数据分析广告投放质量

在本节中我们主要是看如何通过数据信息来分析广告投放质量。我们首先来看跳出率和二跳率。

- 跳出率（Bounce Rate）是互联网上的一个常用指标，指的是进入某一个网站之后不再继续浏览，而直接离开网站的访客比例。通常来说，跳出率越高，网站的粘性就越低。
- 当网站页面展开后，用户在页面上产生的首次点击被称为“二跳”，二跳的次数即为“二跳量”。二跳量与浏览量的比值称为页面的二跳率。

跳出率和二跳率是用来衡量外部流量质量的重要指标。简单来说，跳出率越低越好，而二跳率是越高越好的。0%的跳出率和 100%的二跳率当然是最好的，但是这样的数字只是在理论中存在。在实际应用中，50%的跳出率和 50%的二跳率就已经很值得庆幸了。

如图 9-5 是一个网站某个时间段的浏览量和跳出率列表，为说明简单，这里并没有列出包括来源、二跳率和停留时间等其他信息。我们可以从图中看到，跳出率平均在 30%到 50%左右，高于普通的企业网站，说明页面的优化和内容做得还是可以的。其中跳出率最高的页面是告诉客户联络方式的页面：http://www.adyun.com/contact/，而跳出率最低的两个页面都是临时性的优惠促销信息。

| 页面标题 | 受访页面 | 浏览量 | 跳出率 |
| --- | --- | --- | --- |
| 盘石-全球最大的中文网站联盟 | http://www.adyun.com/ | 7726 | 47.50% |
| 盘石-全球最大的中文网站联盟 | http://www.adyun.com/midpromotion/proadv/ | 289 | 30.55% |
| 盘石-全球最大的中文网站联盟 | http://www.adyun.com/midpromotion/ | 180 | 31.11% |
| 盘石-全球最大的中文网站联盟 | http://www.adyun.com/about/ | 200 | 40.00% |
| 盘石-全球最大的中文网站联盟 | http://www.adyun.com/contact/ | 154 | 51.43% |

图 9-5　页面跳出率示意图

我们之前提到过的 Google 分析（Google Analytics）工具是在国外使用比较广泛的一个网站分析工具。当网站主在他们的网站上布置了 Google 分析的代码之后，下面这些信息会很直观显现在你面前：

- 多少访客在什么时间段访问你的网站；
- 访客访问网站的频率是怎样的；
- 网站中哪些页面是吸引最多用户的；
- 用户采用哪些搜索关键词（组合）来到网站；
- 用户的来源主要来自哪些地方。

在中国，因为 Google 网站访问不稳定，这个工具的使用率被大大降低了。如果你的公司里需要做网站分析，而网站的服务器主要是在中国，那么笔者建议还是选取其他类似的站长工具，虽然功能没有 Google 分析这么强大。

Google 分析除了访问的稳定性之外，还有一些其他的限制。以下信息你可以从 Google 的官方网站中获得 http://support.google.com/analytics/。

- 最关键的问题是 Google 不保证在什么时间点把数据放到报告中。一般来说在 2 小时内访客数据能在网站报告中体现，但有时会延迟至 48 小时。如果你对网站数据的实时性要求很高，那么这个延迟是无法接受的。
- 如果网站平均每个月的访问量超过 1000 万 PV，那么 Google 不保证超出部分会被处理。

- 因为 Google 分析是免费的，所以 Google 不提供任何形式的客户服务热线。如果你的网站分析系统或者数据出了什么问题，那么只能自求多福了。

关于访客的信息包括访客的年龄、性别、学历等可以从大量的网页浏览记录和网络行为中识别出来。如图 9-6 至图 9-8 是我们根据一个月的数据统计的某一个联盟网站的访客信息。图 9-6 中显示的是网站访客性别比例；图 9-7 显示的是网站访客的年龄分布；图 9-8 显示的是网站访客的学历分布。

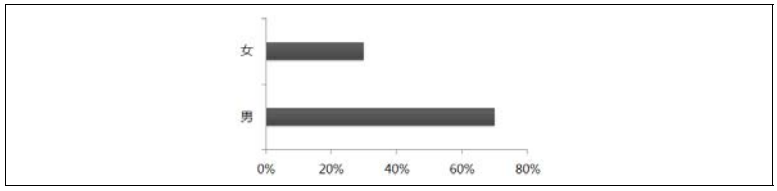



图 9-6　性别比例示意图

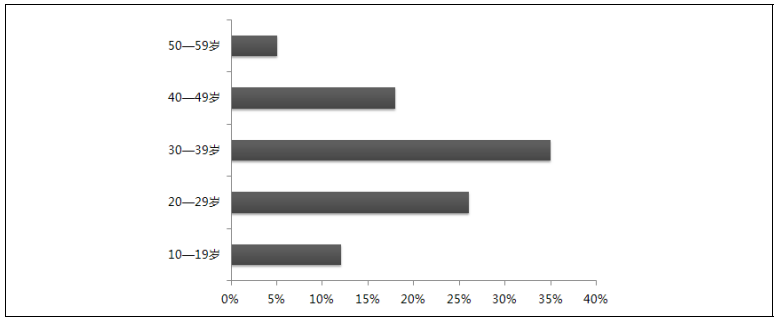



图 9-7　年龄分布示意图

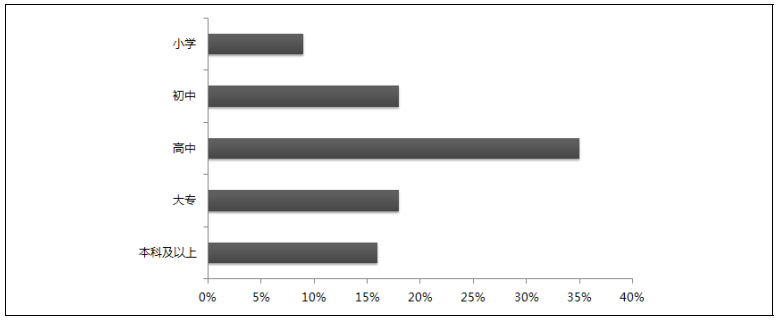



图 9-8　学历分布示意图

上面这些图中的数据对于广告商来说是非常有价值的。如某一款针对男性的产品在这个网站上投放广告的价值会比较高，因为访客中有60%是男性；但是如果一款产品是针对高端人群的，就不太适合在这个网站上做投放，因为只有约16%的人群具有本科或者以上的学历。

#### 9.4.1.2 通过定向和优化提高广告投放质量

除了对人群进行分析之外，我们还可以根据时间段、地区和访问来源区分，使广告投放更加精准。而这样的区分又被称为定向，所以我们对于访问端可以做人群定向、时间定向和区域定向。另外，针对投放广告的网站本身和网站内容我们也可以做选择，这样的选择称为内容定向。下面我们来看一个定向广告投放的实例。

这是我们操作过的某个针对上班族的广告，我们对于客户的网盟广告投放做以下的限制：

- 主要投放在中国经济最发达的地区：北京、上海以及沿海的经济发达地区。
- 只在上班的黄金时间（早上10点到下午6点）投放。
- 不接受网吧或者游戏网站流量的广告投放。

当然，这样的限制会导致一部分潜在用户的流失，我们也可以视广告主的预算和效果要求而调整投放计划。如果在上面这个例子中的广告主有充分的预算，那么我们可以把有上述限制的投放做成一个广告计划，设定每天一定的广告投入预算，而另外开设一个全网全时间段的广告计划来接受辅助流量，设置较少的预算作为前一个广告投放计划的补充。

综合该广告主一周的流量，我们得到如图9-9所示的地域分布图。主要统计广告被显示抓取到的这部分访客的地域来源。即分析比较分布在不同地域的访客行为。

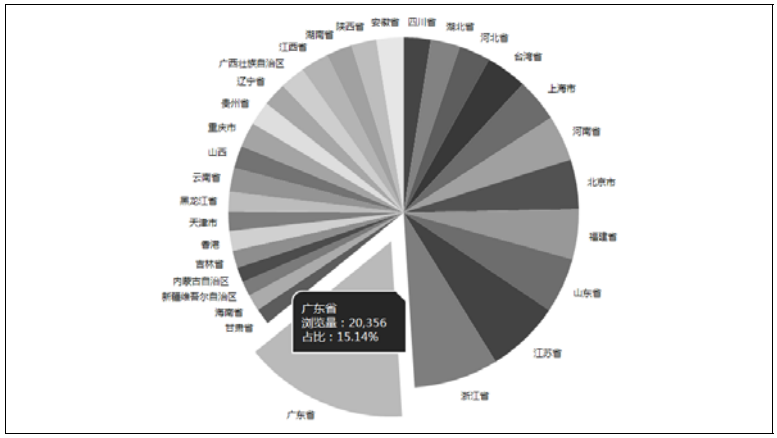



图 9-9　地域分布示意图

从图 9-9 中我们可以看出，该广告的浏览量来源广东省约占 15%，浙江、江苏和山东其次，约各占 7%～8%左右。来自中国经济发达的沿海地区的流量占据整张流量图的 50%以上，证明我们的投放计划设置还是比较合理的。

互联网上网站的种类繁多，大致的种类有门户、IT 类网站、新闻网站、财经网站、房地产网站、游戏网站、汽车网站、生活服务、地方网站、社区网站、视频网站、女性网站、医疗健康和亲子母婴等。图 9-10 是该广告主这一周投放的媒体分布图。我们可以看到在垂直类网站上的投放占据最高的比例，其次是新闻媒体类网站、生活与服务类网站和音乐影视类网站。这个流量分布也可以说明我们针对上班族的投放策略大致是正确的。

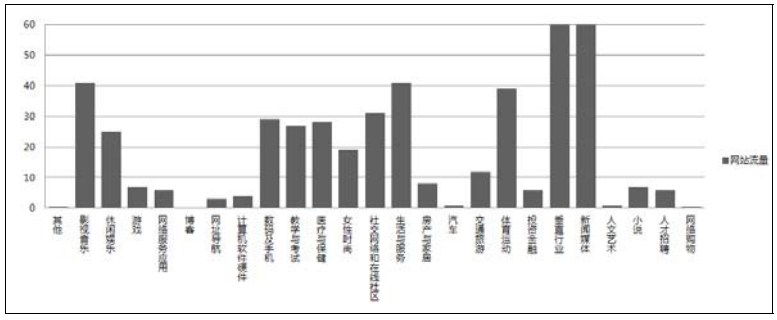



图 9-10　媒体种类分布示意图

我们再来看一个高端母婴类产品的广告主。该广告主是从访客的兴趣点入手，如图 9-11 就展示了他们一个典型客户对于网站内容的兴趣特征。而每个网站也都有一张类似于图 9-9 的表格标识出该网站的普通访客的兴趣特征。通过典型客户的兴趣特征和网站平均访客的兴趣特征之间做的相似比较算法，我们就可以得出该网站的平均访客是否和该广告主的典型客户兴趣一致，从而得出是否要在该网站上投放广告的结论。

我们再来看该广告主某一天的广告浏览情况。如图 9-12 所示。

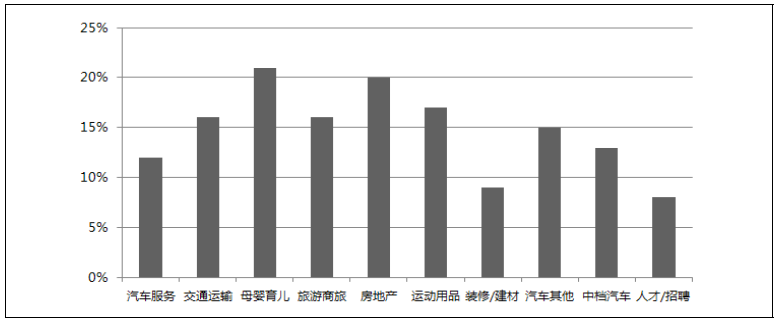



图 9-11　兴趣爱好分布示意图

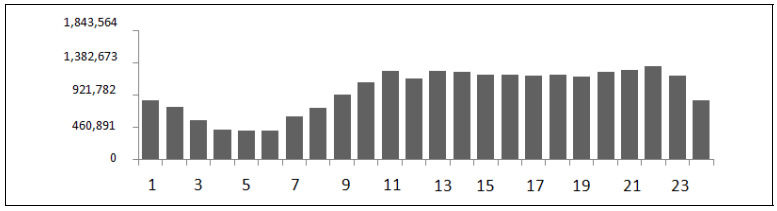



图 9-12　时段分布示意图

网站联盟上的这些数据对于广告商和网站主都是很有价值的。一方面对于广告主来说，他们可以选择针对他们目标人群的网站群来做投放；另一方面对于网站主，他们可以针对广告主做优化，尽量提高点击率以提高总体收入。

我们来看一个广告主在网站联盟上一个阶段投放广告的数据分析，如图 9-13 所示。

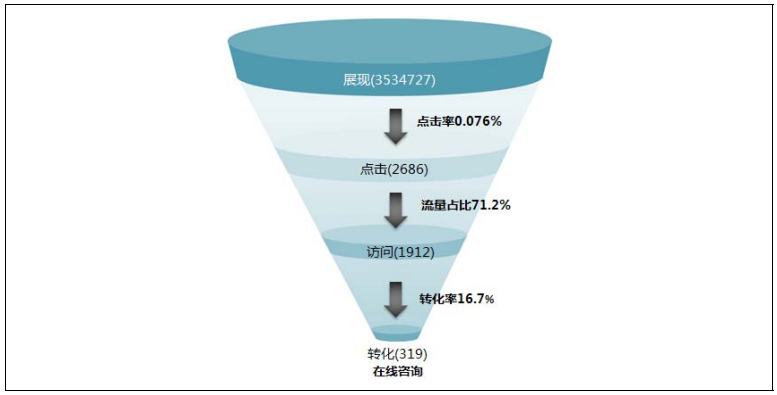



图 9-13　网盟广告投放转化漏斗示意图

这个广告主所有的广告在网站联盟各个位置以各种形式一共展示了 3,534,727 次，被点击了 2686 次，对应的点击率是 0.076%。而这些点击为它的网站一共带来 1912 次访问。这些访问的结果是 319 次在线咨询。这次投放的效果总结如表 9-1 所示。

表 9-1　广告投放效果总结

| 展现量 | 点击量 | ACP | 转化次数 | 转化成本 | 平均展示价格 | 停留时间 | 活跃时间 | 跳出率 |
|---|---|---|---|---|---|---|---|---|
| 3534727 | 2686 | 1.13 | 319 | 9.515 | 0.000859 | 00:00:29 | 00:00:39 | 53.05% |

从表格中可以看出，这次投放整体的效果还是不错的。在网站联盟这种广告形式下，展现量本身是不收费的。这里的 ACP（Average Click Price）是平均点击价格。

广告成本=ACP×点击量

所以该客户的总体费用是 3035.18。

转化成本=广告成本/转化次数

平均转化成本，也就是获取每一个客户的成本是 9.515 人民币。

请读者注意的是，刚才我们列出的点击量乃至 9.4 节中所有关于网站联盟的访客数据都是独立访客的点击量和独立访客的统计信息。对网站信息统计来说，独立访客指的是在一天之内（00:00～24:00）访问网站的上网计算机数量（以 Cookie 为依据）。

一天内同一台计算机多次点击网站联盟的加盟网站的同一广告只被计算 1 次。

我们再来看下这次投放中在小说阅读网站投放广告的效果，如图 9-14 所示。

图 9-13 和图 9-14 展示的是同一次投放中广告出现在全部网站和其中在小说阅读网站上的相应点击率、访问量和转化率的对比。这里我们可以看到，点击率 0.195%，要比平均值高出两倍，而转化率 3.5%只有平均值的五分之一左右。

再分析原因，可能是因为该广告主的目标人群和小说阅读网站的浏览人群不一致造成的。为了尽量提高投资回报率，作为调整的一个步骤，该广告主下一个阶段的广告投放会把小说阅读类网站排除在投放媒体之外。

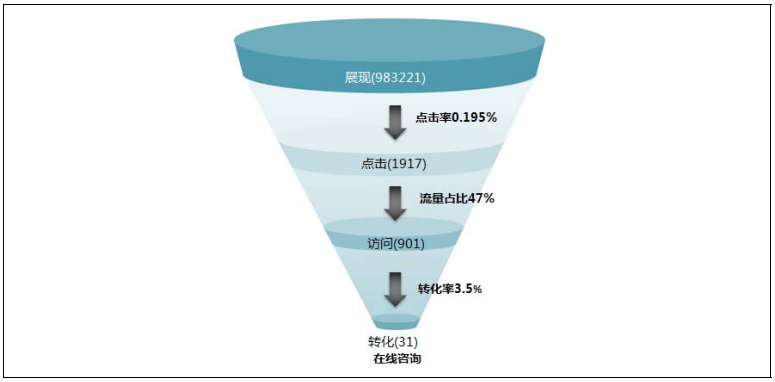



图 9-14　网盟广告投放小说阅读网站转化漏斗示意图

除了上面这些信息以外，还有一些数据分析报表可以用来分析广告主和网站主的具体广告投放数据信息。比如有以下这些报表。

- 时段报表：以常规分析的数据为基础，根据用户自行选取的时间划分方式，进行时间切片式的统计。这样的统计有利于统计数据的定向分析，帮助用户更精确地分析流量数据在时间轴上的纵向分布。统计广告主网站按月、按周、按日或者按小时段的流量分析情况。

- 频次报表：频次是指广告在特定时间内被显示的次数。比如说一个广告在一天中，5 个独立访客观看，每个人观看了广告 2 次，其中每人产生了一次点击，那么这则广告今日 2 频次显示数为 10，2 频次点击数为 5，2 频次点击率为：5/10=50%。
- 点击决策报表：点击决策时间指广告从展现到受众点击广告之间的时间差。
- 搜索引擎流量分析：在流量来源分类统计数据的基础上，进一步地对从搜索引擎而来的流量进行分析，给出指定时间范围内流量趋势、各大搜索引擎的流量数据对比，并可选择查看时间范围内的每日明细或对单个搜索引擎的流量按来源关键字查看数据。
- 广告效果分析报表：统计由各媒体广告投放带到目标网站的整体流量情况。可以通过不同媒体数据的比较从而区分出媒体的优劣度。
- 页面转化：统计由各媒体广告投放带到网站目标页面的流量情况及转化效果。通过页面转化能了解到网站目标页面的转化率以及广告显示点击的转化情况。
- 目标渠道分析：“渠道”是指访客在达到目标转换之前必须通过的一系列页面（只针对广告主网站内的转化）。我们跟踪导向目标的各网页的访客流失率，而此报表名称来源于到达每个页面的访客图表。第一页显示的访客数量最多，在后续页面上，由于访客在到达最终目标之前会不断离开，因此人数也逐渐减少。
- 覆盖度报表：覆盖度是在特定排期和时间段内所覆盖的绝对唯一访客。覆盖度报表统计的是根据 Cookie 识别，统计在一定时间段内观看广告的唯一绝对访客（另外也可统计广告主网站的唯一绝对访客）。
- 覆盖度报表：统计不同排期和媒体在选择时段内拥有的重复访客或者是相同排期不同频道在选择时间内拥有的重复访客。根据 Cookie 来判定重复访客。

- 影响度报表：广告影响度是指广告投放结束后一段时间内广告的显示、点击以及后续行为分析的数据追踪。根据 Cookie 追踪那些访客的后续行为，也可以判断这些 Cookie 的广告投放结束后是通过何种途径过来的。此报表只显示广告投放结束到所选时间点的数据，比如说广告投放是 10 月 5 日结束，所选时间点 10 月 10 日，那么我们只统计 10 月 5 日至 10 日之间的广告影响度。

充分利用这些报表可以使我们的广告投放更有针对性，更有效果，因而广告投放的最终性价比可以达到最高。

## 9.4.2 如何应对网盟广告作弊

在网站联盟上大规模的点击作弊手段五花八门，但是基本上可以分成两类，一种是通过点击机器人，另一种是雇佣廉价劳动力的人为点击。道高一尺魔高一丈，应该说现今的作弊技术比以前的形式更加复杂，而侦查的难度也有所增加。

我们随便在网上搜一下，就可以看到类似图 9-15 的信息。网站主只需要花很少的钱，就可以用作弊软件在他们放置谷歌、百度网盟、腾讯搜搜的页面上自动点击广告来增加收入。


图 9-15 网盟作弊示意图

如图 9-15 所示，点击作弊的方式多种多样。而网站联盟识别点击作弊的方法也随着作弊手段的变化而不断发展，已经有几类行之有效的成熟方法。各家网站联盟都积累了大量的相关数据，但是因为数据涉及多个概念层次的维度，所以人工探测基本不可行。应该来说各家网站联盟公司的作弊识别方法并不相同，而且各网盟也不会把自己防作弊方法的具体细节公布出来。然而，主要的防作弊方法无外乎以下三类：基于异常组分析的方法；基于规则的识别方法；基于分类的方法。

#### 9.4.2.1　基于异常值分析的方法

异常值（Anomaly）的定义是基于某种度量，异常值是指样本中的个别值，其数值明显偏离它（或它们）所属样本的其余观测值。网络作弊行为即使行为再隐蔽（Cloaking），和普通网民的人工行为还是有相当不同的。在网站联盟上用来识别网站的基于异常值分析的方法，根据不同理论的异常值检测方法，可以分成以下几种：

- 基于统计学的异常值检测

在统计学中，假设数据集服从正态分布，那些与均值之间的偏差达到或超过 3 倍标准差的数据对象就可称之为异常值。根据这个定律，可以衍生出一套点击欺诈检测方案。我们对点击率、转化率、对话时间差这些单个指标都进行分析，根据不同行业类型的网站和广告做了统计分析，如果某个网站一定时间段内的数据超出标准，即可怀疑点击欺诈。

- 基于距离和密度的异常值检测

基于统计分布的方法有一个缺陷，它只能检测单个变量，即每次检测只能局限于单个指标，此时若采用基于距离和基于密度的方法，就可结合多指标进行分析。我们目前主要是针对点击率、转化率、对话时间差这些单个指标做基于统计学的分析，但是也可以把这三个指标综合起来用基于距离的方法做分析。

- 基于偏差的异常值检测

该方法的基本思想是通过检查数据的主要特征来确定异常

对象。如果一个对象的特征过分偏离给定的数据特征，则该对象被认为是异常对象。在广告作弊算法中我们主要关注的是 OLAP 数据立方体方法。我们可以利用在大规模的多维数据中采用数据立方体（Data Cube）确定反常区域，如果一个立方体的单元值明显不同于根据统计模型得到的期望值，该单元值被认为是一个孤立点。结合点击欺诈识别分析，基于偏差的方法最主要的是点击流分析，通过点击流分析，我们可以发现那些不规则的点击过程，这些自然可以作为点击欺诈的怀疑对象。

#### 9.4.2.2 基于规则的识别方法

一个对行业熟悉的联盟平台商对各种作弊手段必然了如指掌，通常能够根据经验设定一些作弊防范规则，比如：

- 同一 IP 的用户单日点击次数超过多少即可作为作弊；
- 如果某个广告位的点击率突然大幅增加也可能存在作弊。

制定防作弊规则的优点是方便，在一定程度上也能起到防范作弊的作用，然而这种方法显得比较片面也不能与时俱进，必须要随时间变化而不断更改。

这种基于规则的识别方法相对于其他识别方法来说执行起来要简单很多，而其实这种方法从某种程度上来说也是一种简化了的决策树算法。

#### 9.4.2.3 基于分类的方法

这种方法主要是根据数据挖掘分类算法对历史数据进行模拟，通过构建分类器来对点击行为进行预测。这种方法的缺点在于需要事先对历史点击行为进行分类，即标注出作弊的数据。另外，该方法对数据的完整性和质量要求很高，在我国目前的情况下，大多数网盟平台还不具备满足条件。例如访客在广告主网站的转化数据是识别点击作弊的一个非常重要的因素，但是广告主一般不会将真实数据反馈给联盟平台，造成了这一数据的缺失，而且点击数据一般也都很稀疏，这些因素都会对分类器的实际效果造成影响。

这里列出的第一和第二种方法在很多条件上会存在一定的相通性，因为很多规则也是根据异常值分析得出的。

我们介绍了三种作弊识别方法，那么在现实中，应该采用哪种方法呢。初学者在接触数据挖掘时都会对高级挖掘算法盲目崇拜，觉得方法越复杂，它的实际效果就越好。但实际情况并非如此。现实中很多成功的数据挖掘项目之所以成功往往并不是因为它采用了多么复杂多么先进的理论，当然，这里并不是说高级算法不实用，而是希望告诫每一位数据挖掘工作者，所有的数据挖掘工作都应该紧紧围绕业务为目的来展开，什么方法能在保证最低成本的要求下最大程度的解决问题，那它就是好方法。

纵观各大广告联盟，无论是 Google、百度这样的大型联盟平台还是一些中小联盟平台，在点击作弊识别上几乎主要采用的都是基于异常值分析和基于规则的识别方法。这些方法看起来非常简单，但实际效果却很好。美国纽约大学的 Alexander Tuzhilin 教授在对 Google 的防作弊措施进行研究后，曾经结合长尾分布对这个现象进行解释。Alexander Tuzhilin 教授惊讶于 Google 的简单的基于规则的方法的巨大作用，所做出的解释是大量的点击作弊行为其实都是那些最常用的作弊方法，所以只要不断对点击作弊的表现形式进行分析就能够识别出大部分作弊的规则。这其实很好理解，比如说无论学生用什么作弊方式，一个有经验的老师总能察觉，即使这个老师并不了解学生的那些先进的作弊工具。因为老师要看的是学生作弊时的表现。

采用数据挖掘的分类算法，对于联盟平台在数据质量和数据完善上的要求是比较高的。通常来说，有 Cookie 的情况下作弊可能性会比较少，而无 Cookie 的比例高，作弊的可能性也会比较大；跳出率极高的情况下，作弊的概率会比较高，而跳出率越低，作弊的概率也越低；点击之后在网页上的停留时间极短，作弊的概率会比较高，而停留时间越长，那么是正常流量的概率会越大。

如果跳出率（Bounce Rate）较高，那么一个访客进入网站之后不再继续浏览，直接离开网站的比例就越高。通常来说，跳

出率越高，网站的粘性就越低。而对于网站联盟来说，如果从联盟网站上点击广告到达的广告主页面跳出率比较高，那么说明引流的效果不好，特别是无论什么广告，点击之后的跳出率都比较高，那么我们就需要考虑该联盟网站是否有作弊嫌疑还是本身就是低质网站。例如说国内的有些阅读和视频网站，在你打开每个页面时，都会自动有窗口弹出，正式说法叫做“弹窗广告”。这些广告往往在弹出的瞬间您就会把它关闭，但是对于广告主来说，这已经产生了一次点击，是要收费的。这样的引流方式，虽然不一定算是作弊，但至少是低质的流量。

我们来看一个国内一家网站联盟公司用决策树判断作弊流量的案例。

这家网站联盟公司之前积累了大量关于作弊网站的数据。通过决策树生成算法对于这些数据进行学习，最后发现和网站作弊最相关的数据包含 Cookie、网页停留时间、跳出率、二跳率等。我们来看一下生成的决策树。如图 9-16 所示。

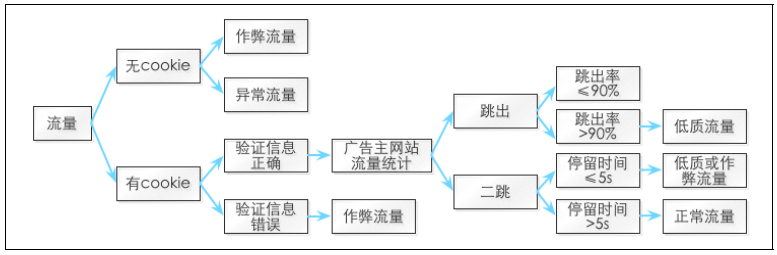



图 9-16　网盟作弊分析决策树示意图

从图 9-16 中我们可以看到决策树模型示意图中第一层是 Cookie 的有无。如果有来自该网站较高比例的流量没有 Cookie，那么我们判断为作弊流量的概率是比较高的。在 9.2 节中我们讲述的 Facebook 案例其实就是因为 80%的流量没有 Cookie 就被认为是作弊的。在图 9-16 的第三层，对于流量的统计，如果跳出率比较高，那么在跳出率到达令人恐怖的 90%时，我们就不需要证明该网站是否是作弊网站了。即使该网站并没有作弊，如此高的跳出率也使我们做出排除该网站的低质流量的决定。同样，

如果二跳率比较高，但是平均停留时间在 5s 以下的，该网站的流量或者是低质或者是作弊流量，也是不可取的。

## 9.5 本章相关资源

- 本章相关参考文献：

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. *Anomaly De tection*: *A Survey* (*2009*) *ACM Computing Surveys*. Vol. 41(3), Article 15, July 2009.

[2] Ar Lazarevic, Aysel Ozgur, Levent Ertoz, Jaideep Srivastava, Vipin Kumar, A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA.

[3] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan, Data Mining for Network Intrusion Detection (2002). In Proc. NSF Workshop on Next Generation Data Mining, Baltimore, MD.

[4] Aleksandar Lazarevic, Paul Dokas, Levent Ertoz, Vipin Kumar, Jaideep Srivastava, Pang-Ning Tan, Cyber Threat Analysis - A Key Enabling Technology for the Objective Force (A Case Study in Network Intrusion Detection) (2002). Proceedings 23rd Army Science Conference, Orlando, FL.

- 本章相关网址：

[1] http://eguan.cn

[2] http://www.iresearch.cn/

[3] http://www.itongji.cn

[4] http://support.google.com/analytics

# 第 10 章
# 数据挖掘和电子商务

有人说过数据挖掘就是一个商业过程（Data Mining is a Business Process），不能产生商业价值的数据挖掘是没有意义的。只有把数据挖掘充分应用到商业上，我们之前做的所有准备和算法研究工作才有意义。

在本章中我们来看如何把数据分析技术和数据挖掘过程应用到电子商务领域以产生商业价值。我们先阐述电子商务企业的需求，总结出电子商务面临的一些问题以及需要数据挖掘可以起到作用的 6 大应用，然后再通过实际案例来看怎样通过数据挖掘实现这些应用以帮助电子商务企业解决问题。

## 10.1 中国电子商务现状

电子商务通常是指是在全球各地广泛的商业贸易活动中，在开放的网络环境下，买卖双方不谋面地进行各种商贸活动，实现消费者的网上购物、商户之间的网上交易和在线电子支付以及各种商务活动、交易活动、金融活动和相关的综合服务活动的一种新型的商业运营模式。其实质就是以互联网为平台进行的一种贸易活动。

在中国，电子商务是从 B2B（Business to Business，企业对企业）发展而来的，而后有了 C2C（Consumer to Consumer，消费者对消费者）和 B2C（Business to Consumer，企业对消费者）两种商业模式。应该说，在我国目前的经济环境下，B2C 商业

模式的全面发展受居民生活水平、物流行业发展、在线支付的完善度、网络普及率等因素的制约，这决定了 B2C 这种商业模式更适合一级城市和一些较为发达的二级城市。目前，做得较好的一些 B2C 网站基本是以北京、上海、广州、深圳、杭州等经济发达的城市为据点，在这些城市中，由于工作、交通等因素，导致时间成本高，生活节奏快，使得这种方便、快捷的网络购物方式更容易被上班族所接受。

B2C 是英文 Business to Consumer 的缩写，其中文含义为企业对消费者。B2C 是电子商务的一种模式，也就是通常意义上的网络零售业，即商家借助于互联网直接面向消费者销售产品和服务。在今天，B2C 电子商务以其灵活的交易手段、低成本高效益的营销模式、快捷的物流配送等优势成为电子商务的几种业态中发展最迅速也是最具有生命力的商业模式。这种形式的电子商务一般以网络零售业为主，主要借助于互联网开展在线的销售活动。

如今，B2C 电子商务正在深刻的改变着经济、市场和产业结构，改变着产品、服务及竞争模式，同样也改变着消费者的价值和行为以及就业和劳动力市场结构。

在快速腾飞的中国经济的带动下，人们对电子商务的接受程度不断提升，除原有的互联网电子商务企业外，传统行业的企业互联网化趋势明显，纷纷将电子商务平台作为线下产品推广渠道的线上补充，通过在大型电子商务平台上面建立店铺或专区或自建电子商务平台的形式，积极推行企业电子商务的全线布局。同时，外贸行业的不景气也是触动传统企业转型的一个重要原因。传统企业对于电子商务的认知与应用成为 2011 年以来中国电子商务市场蓬勃发展的主要原动力之一。

电子商务行业的快速发展和激烈竞争，使这个行业的变化越来越快，对于电子商务企业的老板们而言，要想在瞬息万变的行业趋势中作出最快的反应，就必须能够作出准确地预判，那么学习运用数据分析就是必然的了。所以从另外一个角度来说，在所

有的企业中，电子商务对于数据分析最重视，需求也是最大的。特别传统企业，如果想要在网络上胜出，需要接入互联网的基因，而数据分析和数据挖掘的应用是富含互联网基因的。

中国电子商务新贵凡客诚品 VANCL 的 CEO 陈年就曾经说过，“凡客越来越像是一家数学公司，需要对大量的订单和用户信息进行分析，进而更好地指导生产工作，减少高库存”。这已显现出数据分析对电子商务企业的重要性，而订单管理只是企业管理中可以用到数据分析的过程中的一个。

阿里巴巴集团董事长马云在 2012 年网商大会上做了讲话，有好事者对他的全篇演讲做了分词研究，发现在其中“数据”作为关键词出现了 15 次，仅次于“经济”的 38 次。我们可以认为马总认为对于网商来说，赚钱是第一位的，而数据的重要性仅次于赚钱，或者对于网络商家而言，未来是否能够赚更多钱只有靠数据作为基础。

图 10-1 中是对中国 B2C 市场规模的预测，其中不包括 C2C 市场的数字，也就是把淘宝的数字排除在外。如果包含了淘宝的 C2C 数字，在 2012 年电子商务的总体规模已经超过了 1 万亿元人民币。

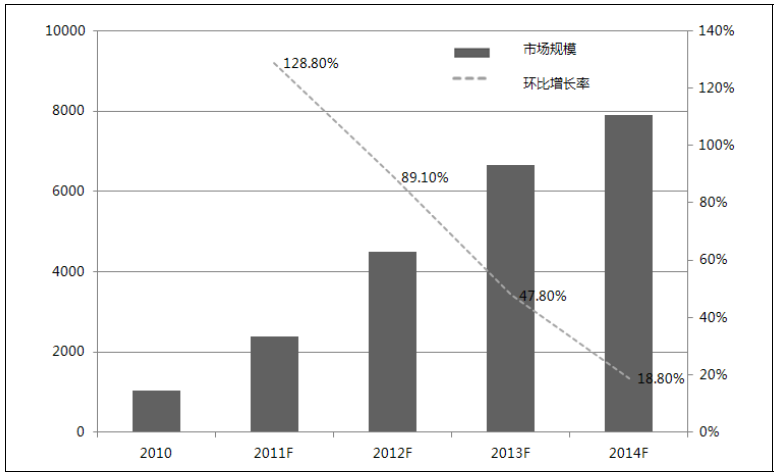



图 10-1　来自中国电子商务协会的中国 B2C 市场规模预测

由于大量的企业涌入电子商务市场，使市场竞争愈演愈烈。在众多行业中，竞争最激烈的莫过于 3C 领域（3C 指的是通信产品 Communication、消费类电子产品 Consumer Electronics 和计算机产品 Computer，三类产品的首字母都是 C，所以称 3C 产品）。在 3C 行业中，我们可以听到各种竞争源源不断的通过各种渠道传到我们的耳朵里，无论是从 2012 年 6 月 18 日京东商城店庆引起的各大电子商务疯狂促销活动或是 8 月份从微博上引发的京东、苏宁、国美这三家电子商务企业的“约架”事件。虽然这些事件真真假假，其中也会有大量的炒作成分存在，但他们还是向我们透漏了这样一个信息“电子商务市场正朝着红海大跨度迈进”。

贝恩公司曾经联合市场调研机构 Kantar Worldpanel，对中国 40000 户家庭购买 26 个快速消费品品类的真实购物行为进行了深入研究，并总结出了对相关企业和品牌具有深远意义的观点。而研究得出的最重要的结论就是：中国消费者没有品牌忠诚度。

事实的确如此，很多时候中国消费者在购买大多数消费品时，通常会在多个不同的品牌中进行选择，往往很少忠于某个特定的品牌。而对于期望赢得中国消费者青睐的品牌方而言，使出各种招数赢得消费者的关注便是重中之重了。其实这些事件同样告诉企业，应当勇敢地参加战斗，因为竞争的结果可能是赢，而不竞争的结果一定是输。就拿 2012 年 8 月份“京东、苏宁、国美大战”来说，闹剧之后，我们看到了多个调查，结论虽不尽相同，但是除了苏宁易购在品牌知名度上增加得比较快之外，京东和国美也都有不小的获益。

在大多数情况下，当消费者购买某一品类产品的频率增加时，通常也倾向于尝试更多品牌，我们称之为“多品牌偏好”行为，即在相同的购买场合或消费需求下，消费者是“三心二意”的，他们往往倾向于在同一品类中选择不同的品牌。这些品类中的高频率购买者（即某品类或品牌购买频率前 20%的消费者）也同样表现出“多品牌偏好”行为。研究还表明，在表现出“多

品牌偏好”行为的品类中，某一品牌的高频率购买者通常也是其竞争对手的高频率购买者。

以中国饼干市场领导品牌奥利奥为例。卡夫集团自 1996 年将奥利奥引进中国后采取了众多举措，成功地吸引了更多新的消费者尝试并购买其产品。首先，卡夫为了迎合中国人的口味，降低了饼干的含糖量，从而促进了销售；其次，推出了小包装产品（如迷你奥利奥），不但符合中国消费者对于饼干大小的偏好，还降低了产品单价，使更多的中国消费者能够负担得起；此外，卡夫还推出了绿茶、冰淇淋等本地风味的奥利奥，以满足消费者的不同需求。但是，奥利奥的高频率购买者在为奥利奥贡献了约六成销售额的同时，也为各主要竞争品牌分别贡献了 25%～35%的销售额。事实上，奥利奥的高频率购买者将高达 3/4 的饼干开支贡献给了其他饼干品牌。

中国消费者的“多品牌偏好”或者说“缺乏品牌忠诚度”，在电子商务领域表现得尤其明显。从 2008 年才初创的淘宝品牌“韩都衣舍”今天已经达到日订单 8000 单，但是在“韩都衣舍”上消费的客户往往在同一时间还是另外一个淘宝品牌“七格格”的大客户。对于电子商务企业或者新加入电子商务领域的传统企业来说，这是一个非常棒的好消息，因为只要你能让客户喜欢你的产品，就会有客户购买。

电子商务既充满挑战，也蕴含机遇，而我们认为成功的关键在于详细了解每个消费者个性化的需求和真实购物行为，而非通过市场调查得到大多数购物者对过往经历的回忆或对未来行为的预判这样的普遍信息。

电子商务覆盖面很广，为使叙述更有针对性，本章主要基于 B2C、C2C 这类的网络零售市场中的卖家来探讨如何利用数据挖掘来提高网络零售商们的业绩。我们首先从客户分析的角度对数据挖掘技术在电子商务卖家中的应用进行阐述，再结合两个案例详细介绍如何在电子商务中做具体的实施。

B2C 电子商务的营销手段一般分为以下四步。如图 10-2 所示。

- 第一步：引进流量。通过各种推广手段，带来尽量多的潜在客户。常用的网站推广方法有：搜索引擎优化、活动推广、新闻推广、关键词广告、联盟推广、友情链接等。
- 第二步：提高转化率。让潜在客户转化成实际客户。提高转化率的常用措施有提高网页打开的速度、美化页面视觉效果、简化页面的排版、优化购物流程等。
- 第三步：提高网站粘度。提供高质量的产品是产生用户“粘性”的最重要因素。如果一个 B2C 网站能够长期提供给消费者价格合理且高质量的商品，或者有一些让访客不得不再次访问的理由，则必然会吸引大量的消费者来购买和分享。
- 第四步：效果检测。通过分析网络营销的效果，B2C 网站就可以对营销状况一目了然。然后再相应地调整营销策略，如此循序渐进，改进营销手段。

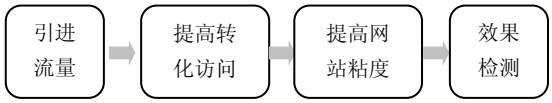



图 10-2　电子商务营销手段

在电子商务市场的残酷竞争中，要想胜出只有看是否能够比别人更深的了解客户的内在需求，在我们看来就是能否比别人更好地做好数据分析和数据挖掘工作。

做好数据工作，需要企业从基础数据就开始规划，在后期才可能有更大扩展性。而在数据建模的过程中，根据业务场景统计汇总基础数据，分别结合用户访问信息、用户消费信息、订单数据、物流供应链数据等，汇总为客户生命周期数据模型、商品销售预测数据模型、生产计划数据模型、物流规划数据模型等面向不同主题的数据模型，这些都需要非常专业的综合知识才能得出。

所以在电子商务中做数据挖掘除了需要数据分析和数据挖

掘的知识之外，还需要对电子商务行业有深入的了解，二者缺一不可。如果只有数据挖掘知识，很可能做出的数据模型在理论上是正确的，但是不符合电子商务的实际情况；反之，如果只是对电子商务的运作有深入了解，而对数据挖掘不理解，那么可能会对数据模型提出不切实际的非分要求，或者做出一个在功能上有缺失的模型。

## 10.2 在互联网上卖米

在网上进行交易的最大优点是电子商务企业（个人）可以在互联网中取得大量的真实数据，包括真实的市场数据、网站流量数据、产品被关注和浏览数据、产品销售数据等，从而使他们可以有效地估计出访客的兴趣和对产品的反应。当我们有明确的且可以量化的目标时，采用数据挖掘技术的效果是最好的。由于我们可以很清晰的得到客户的行为数据，分析客户的各种行为，这就方便了我们通过各种数据分析和数据挖掘方法来了解客户。

销售最简单的需求就是增加尽量多的新客户，留住更多的老客户。而在互联网做销售，也是一样的。如果再细化，电子商务可能会考虑这样一些目标：提高转化率；增加每次会话的平均浏览页数；增加每次结账的平均利润；减少退货；增加总顾客数量；提高商标知名度；提高回头率（在一定时间，比如 30 天内做第二次购买的顾客的数量）；增加每次访问的平均结账次数；提高客单价等。我们要做的就是通过充分了解客户来满足这些需求。

“台湾经营之神”王永庆是台塑集团的创办人，他 16 岁时靠父亲借来的钱开了一家米店。由于居民都已经有了自己熟识的米店，王永庆新开的米店生意很冷清。但王永庆并未放弃，他一家家地走访附近的居民，当他发现买米的主要是家庭主妇时，他提出了当时其他米店所没有的一项服务——送货上门。不仅如此，每当王永庆把米背到主顾家里时，他还会热心地询问家里有多少

人，每天大概会用掉多少米，并细心地连同主顾本次购米的数量、家中米缸大小、家庭住址等信息一起记录在随身携带的小本子上。通过这些信息的收集，王永庆很容易计算出主顾大概什么时候需要新购大米。这样，每到哪家的米快吃完时，他就会主动将米送上门来。如此以往，王永庆的米店立刻就红火了起来。

“王永庆卖米”的故事常会被作为营销学经典案例来讲解，王永庆之所以成功，就在他对于客户的了解和对老客户的深度服务。衍生到电子商务上，如果电子商务卖家能对买家有深入的了解，势必会对业绩的提高有所帮助。可喜的是，作为电子商务卖家，搜集客户资料的方式不需要像王永庆那样一家一家询问，因为客户在互联网上的行为都是有迹可循的。客户只要登录到卖家的页面，他的一举一动都可以变成数据记载下来，但困难的是电子商务企业（个人）所面对的客户和客户访问轨迹的信息量往往是巨大的，如何有效的利用这些客户数据来充分了解每一个客户并挖掘出背后的客户价值则成为在电子商务企业成功运营的关键。

我们来看一下如何通过这些数据帮助网络厂商达成目标。解决问题的第一步是清晰地描述问题。通常，网络厂商需要解决的问题有以下这些：

- 如何投放广告以寻找合适的客户人群；
- 如何找出同一类访客的特征并预测其未来的购买行为；
- 如何组织安排网页内容，以符合访客的个性化需求；
- 如何自动地把商品分类，把同时可能购买的货物放在同一个网页上，以增加单次购买的商品总值；
- 如何估计购物车被放弃的可能性以及如何降低这一数字。

所有这一切都建立在寻找不同的隐含数据模式并正确应对的基础上。

像当年在卖米的王永庆一样，随着电子商务付费推广流量的上涨，越来越多的电子商务企业（个人）开始关注在以往高速发展的背景下潜在水底的隐形冰山——老客户的价值。怎样像“王

永庆卖米”这样最高程度的抓住老客户，也是我们接下来要研究的课题中比较重要的一项。

## 10.3 用数据来掌握客户

传统企业在做营销时，经常通过市场调研、电话访谈和调查问卷等方式来发现潜在的客户。但是客户一般不愿意主动透露自己过多的私人信息而随意作答，市场调查所得到的信息可能不符合事实，从而造成一种假象。此外，这些调研的对象数量不可能太多，未必具有代表性，可能只反映了局部市场信息。总之不能有效发掘出潜在的客户。

身处大数据时代，营销必然会更多地依赖数据，以便更精准地找到目标用户并留住客户。对来自于不同平台的数据进行进一步挖掘和分析，找到这些数据相对应的人群，再将这些群体进行个性化的对比，并以此展开个性化的营销服务，是掌握客户的根本。例如利用针对合适的人群发送 EDM，或者流量对接，或是与传统搜索相结合等方式引入新客户，进而对于这些客户行为进行数据分析，尽量把新客户转化成优质的活跃客户，提高客户的消费。

大数据时代的一个重要趋势就是数据服务革新。拥有的海量数据使我们可以把人分成很多有自己独特属性的群体，针对每个群体给予不同的服务。而数据挖掘主要侧重于如何找到我们关注的客户属性，总结出用户需求，以及将数据转化为对客户有帮助的信息。我们将数据挖掘在电子商务中的主要应用总结成以下 5 条：

（1）发现潜在客户

对一个电子商务网站来说，了解、关注记录在册的客户群体是非常重要的，但从众多的随意访客中发现潜在客户群体也同样非常关键。如果发现某些访客属于潜在客户群体，就可以针对这

类访客实施一定的营销策略，使他们尽快成为我们的新客户。对客户访问记录进行数据挖掘，可以利用分类技术在网络上找到潜在客户。对已经存在的访客进行分类，一般可以分为三种：新来访者、偶然来访者和常客。对于每个来访者，可以通过分类模型识别出这个客户与已经分类的老客户的一些公共属性，从而对这个新客户进行正确的归类。然后根据归类判断，决定是否要把这个来访者作为潜在的客户来对待。对新来访者，我们可以收集的信息比较有限，在没有其他关联网站信息的情况下，只有一些日志信息。而对于偶然来访者，通过两次访问的停留时间和访问深度，我们可以有比较多的信息。

我们也可以总结有价值的客户来源，通过对来源进行数据分析，发现他们的共性，从而加大对于引流最多的来源或者性价比最高的来源的投入。比如，我们通过数据分析可能会发现，在网盟广告中，在某个时间段针对某些类型的网站投放广告性价比是最高的，因此改变我们的广告投放策略，使我们能在有限的预算下找到更多的潜在客户。也可能会发现自然流量占的引流比例比较高，那么我们就需要增加 SEO 的投入。

（2）留住老顾客

二八定律说的是企业 80%的业务收入通常来自于 20%的客户，而向新客户进行推销的花费要数倍甚至数十倍于向现有的客户进行推销的花费。通过 Web 数据挖掘，我们可以发现什么样的顾客群在什么样的时间段内在网站上购买了什么样商品，平均支出是多少，他们最喜欢的商品是什么类型，对于新推出的产品哪些客户可能会购买，哪些是网站最需要留住的客户等，以便对其进行个性化营销和人性化关怀。

（3）针对不同客户提供个性化的产品

电子商务企业（平台）可以获知访客的个人爱好，能够更加充分地了解客户的需要，根据各种信息来细化市场，甚至是为每一个顾客的独特需求提供个性化的产品，这都有利于获取新的客户和提高老客户的满意度。为了使网络信息挖掘技术更好地应用，商家必须记录访客的所有特征及条款特征。当访客持续访问

某网站或者其关联网站时，有关访客的数据便会逐渐积累起来。

（4）建立电子商务推荐系统

推荐系统就是向客户推荐商品或提供信息来引导客户购买商品的系统。推荐系统可以根据其他客户的或该客户的信息，模拟销售人员帮助客户导购的过程，为客户提供个性化服务。推荐的形式包括预测用户对某种商品感兴趣的程度，向客户推荐商品，或是根据用户的兴趣特点和购买行为，提供个性化的商品信息等。

推荐系统可以将浏览者转变为购买者。有时人们只是看看网站的内容而并没有购买的意思，那么推荐系统可以帮客户找到他们感兴趣并且愿意购买的某样商品的兴奋点，以推进消费者形成购买行为。推荐系统可以基于客户已经购买的商品，推荐客户购买一些相关的商品，或者购买一些相关的但是对于商家来说利润更高的商品来增加交叉销售（Cross-Selling）和向上销售（Up-Selling）。推荐系统的作用还在于建立忠诚度，因为客户往往更愿意到那些最能满足自己需求的网站去购物。

（5）改进网站的设计

站点上页面内容的安排就如超级市场中物品在货架上的摆设一样，把具有一定支持度和信任度的相关联的物品摆放在一起可能有助于销售，利用关联规则可以了解如何针对客户动态调整站点的结构，使客户访问的有关联的文件之间的链接能够比较直接，让客户更容易访问到有兴趣的页面。网站如果具有这样的便利性，就能给客户留下较好的印象，同样可以增加下次访问的几率。

我们对 Web 站点链接结构的优化可从两方面来考虑：一是通过对 Web 日志的挖掘，发现用户访问页面的相关性，从而在密切联系的网页之间增加链接，方便用户的使用；二是通过对 Web 日志的挖掘，发现用户最终要到达的目标页面。如果在目标页面的访问频率高于对实际到达页面的访问频率，可考虑把目标页面和实际到达页面互换，从而实现对 Web 站点的优化。

一般来说，用户在一个网站上的平均停留时间和每个用户对网站的平均贡献值是成正比的。那么，为了使客户在自己的网站上驻留更长的时间，我们就应该深入的了解客户的浏览行为，知

道客户的兴趣及需求所在，动态地调整 Web 页面，以满足客户的需要，向客户推荐、提供一些特有的商品信息和广告，从而使客户能继续保持对访问站点的兴趣。如何延长客户的停留时间是我们网站设计改版最重要的任务。

下面我们具体来看这 5 项数据挖掘的应用在电子商务领域一般是怎样实现的。最后在 10.4 节中会通过案例来具体阐述数据挖掘在电子商务中的应用。我们会以两个电子商务企业为案例介绍数据分析和数据挖掘，包括序列算法的应用、K-means 算法的实现和关联规则的分析等。

### 10.3.1　客户何时来，从哪来

要解答客户何时来，从哪来的问题，诉诸于电子商务领域最常听到的一个词就是流量。通常说的流量（Traffic）是指网站的访问量，是用来描述访问一个网站或是网店的用户数量以及用户所浏览的网页数量等一系列指标，这些指标主要包括：独立访客数量（Unique Visitors）、页面浏览数（Page Views）、每个访客的页面浏览数（Page Views Per User）。

查看流量数据可以采用的工具有 Google 分析（Google Analysis）、百度统计、淘宝量子恒道等。利用这些工具，我们可以从多维度来分析流量，例如从时间维度来分析流量，可以得出在什么时间段访问某类商家的客户最多，也就是客户最喜欢在什么时候来到店铺。这对一些中小型卖家非常有用，简单来说淘宝店铺的宝贝上下架时间就直接与此相关。在淘宝规则中快下架的宝贝往往在搜索结果中会靠前，如果能够分析出商家在一周中的某一天某时刻的流量将会达到高峰，再设定宝贝上架时间为 7 天，那么当宝贝下架时正好也处在流量最高峰的时候。也就是我们在客户最高峰的时候上架宝贝，当下个高峰来临时宝贝也正好处在搜索结果的前列，这样的安排下宝贝销量势必会有所提高。

从时间维度上来分析流量，首先需要关注的是流量的高峰时段，这点可以从按时或按天的流量序列图看出，比如可以从淘宝

数据魔方里的“买家什么时候来”的功能中得出买家来访高峰时间段。从图 10-3 中看到，凌晨的流量是比较少的，从早上 8 点开始流量开始增加。流量最高峰的时间段是在晚上 21:00～22:00 之间，最近 30 天有 3800 万人次的访问，其次是 20:00～21:00 之间，最近 30 天有 3670 万人次的访问。

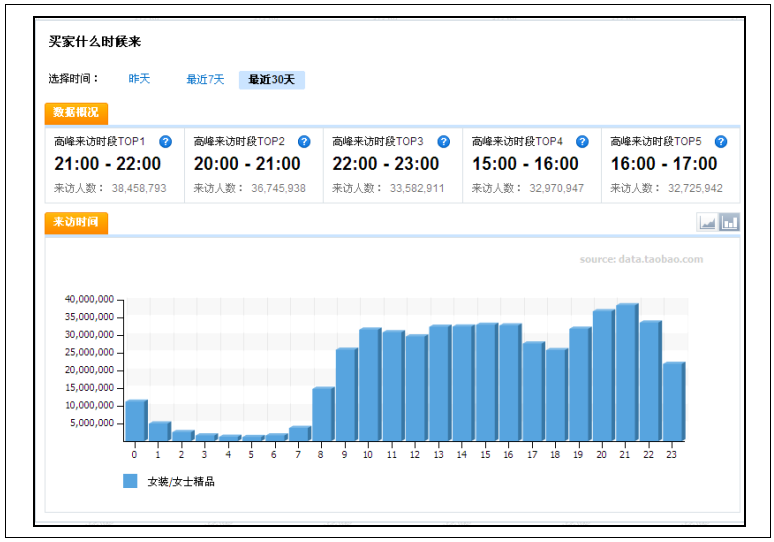



图 10-3　买家来源示意图

除此之外，各种流量分析工具还可以从来源的维度对流量进行分析，对访客来源分析可以让我们了解到客户都来自哪些省份，哪些城市，甚至哪些城市的哪些地区，这样就可以对重点省份或者城市的客户采取促销和优惠活动。访客来源分析还可以得出客户是通过哪种方式找到商家网站的，根据这个信息，商家可以调整广告推广方式，让广告资金集中到最有效的流量来源的推广上。图 10-4 同样以淘宝数据魔方为例。城市的分布比较均匀，其中来自上海的流量最多，占到总流量的 5.84%，其次是北京、杭州和广州。

在做流量分析和访客来源分析时，我们最常使用的数据挖掘的方法是时间序列。时间序列是数据挖掘领域中用来分析一段时间里各项指标的变化情况最常用的方法，通过时间序列我们不光

可以从趋势图中看出网站（店）流量的大体变化情况，更重要的是我们能够预测未来一段时间的网站（店）流量情况。


图 10-4　买家地理位置分布图

除以上两个维度外，网站（店）引进流量的“转化率”也是各电子商务企业（个人）十分重视的一个指标，它同样也是衡量店铺及网站引入流量是否优质的一个重要标准，那么何为“转化率”呢？

转化率（Conversion Rate）指的是产生实际消费的用户和来到用户网页的总用户数量的比值，是将流量转化为实际销售额的一种衡量方式。每一个电子商务企业（个人）的最终目的都是为了赚钱，那么显然提高转化率是电子商务企业（个人）提高销售额最直接的方式。假设某个电子商务企业（个人）每天的销售额在 50000 元，而他们的平均转化率是在 1%，试想在同样流量的情况下，如果把转化率提升到 1.5%，那么他们每天的销售额就可以上升到 75000 元；如果可以把转化率再提升到 2%，那么他们每天的销售额就可以到 100000 元，再加上节日大促和活动营销带来的销量，保守估计这家电子商务企业（个人）就可以达成每年 5000 万元的销售额，一跃而成为一家值得一提的电子商务企业（个人）。

下面，让我们来看一组数据，目前整个电子商务领域的平均订单转化率是 1%～2%。平台的转化率相对较高，而绝大多数B2C 店铺的转化率都在 1%以下。根据当当网自己的数字显示，他们的转化率做得比较好的情况下可以到 3%，而美国最成功的电子商务网站亚马逊（Amazon）则可以达到 4.5%。

其实，影响转化率的因素很多。首先，吸引人的产品图片就是其中一种。下面我们来看两幅晚礼服的商品图片，这两幅图片都是截取自实际的电子商务平台网页。如图 10-5 所示。左图来自 Aliexpress——阿里巴巴在美国的网店，而右图来自美国休斯敦的一家网店。不用去查看实际的消费数据，我们就可以判断在相同流量的情况下，左图的产品转化率一定会高于右图，很显然是因为左图中模特把晚礼服的美感很直接的呈现在消费者面前，而右图只是一个干巴巴的商品。很多网购用户需要能够明确的看见他们在网上花钱购买的产品穿在自己身上可以是什么样。虽然有数据已经证明高质量的产品可以大幅提高商品的转化率，但如果能从各个角度高清度地展示产品的美感和卖点，那么效果必然会更好。


图 10-5　产品展示对比示意图

其次，影响转化率的另外一个因素则是产品描述。虽然说一幅图抵得上十句话，但是很多时候，有些信息是需要通过对产品书面的描述才能够传递给消费者的。那么书面描述需要多大的篇幅呢？是越长越好吗？关于这些问题，我们可以从一些成功的案例上学习，下面为大家提供一个简洁的版本和一个详尽的版本，简洁的版本只需要回答三个问题：产品的目标受众是谁？产品能做什么？产品为什么好？而详尽版本的产品描述中应当包含可以想到的所有内容。详尽产品描述的目的是用户看了这个版本的介绍之后，不会再有任何遗留的问题。

我们来看图 10-6。这张图来自 Sears 的网上商店。这台自动剪草机有两个产品描述，简单描述只有一句话“Honda’s lightest and quietest lawnmower engine delivers 160cc of power”，中文意思是“本田公司最轻和最安静的自动剪草机，能提供 166CC 的力量”。简洁明了，但是如果想要看更多的信息，则可以点击后面的链接来看完整的产品描述，这里面包括消费者想要知道的各种商品参数以及购买评价等。

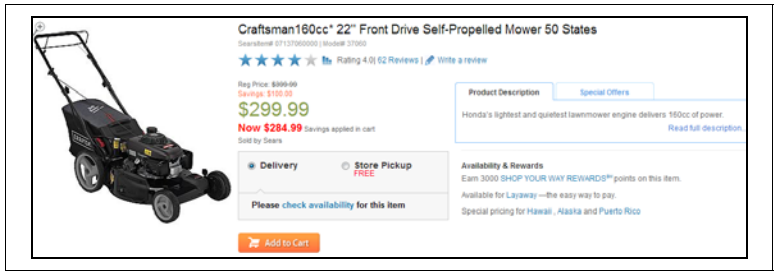



图 10-6　Sears 商家产品介绍示意图

## 10.3.2　客户最喜欢哪种商品

客户最喜欢哪种商品是电子商务企业（个人）经常需要回答的一个问题。很多人可能会觉得要通过数据分析来回答这个问题实在是太简单了，只要对每款商品的销售额或销售量从大到小排序就可以了。这个答案本身无可厚非：销量好的商品自然是客户

喜欢的商品。

但是这个问题在电子商务领域并不那么简单，假设一家网店的一款商品 A 一个月卖出 100 件，而另一款商品 B 一个月的销量只有 10 件，单从销量数字来说商品 B 是远不如 A 的。但我们现在需要多看一个指标，商品 A 的一个月浏览量是 40000 次，而商品 B 一个月浏览量只有 60，也就是说商品 A 的 100 件销量是靠非常高的展现量来支撑的，它有可能占据了很好的展现位置，有很精美的展现页，但产品本身的吸引力可能远不如商品 B。商品 A 的转化率是 0.25%，而商品 B 的转化率高达 16.67%，是商品 A 的 66.7 倍，也就说商品 B 如此高的转化率可能更值得我们去关注。如果我们只把销量作为唯一的维度，就可能因商品 B 的小销量而忽略掉商品 B 的潜力。

单纯从销量出发的思维方式反映出人为识别方法的缺陷，尤其是当商品量及其维度巨大的时候。对于高维度的商品数据，我们可以运用数据挖掘中的聚类来分析并回到本章节的问题。聚类本身就是结合多个角度的属性来对事物进行分类，当我们面对大量的商品时，我们无法从一两个指标中来对商品归类，通过聚类则可以考虑通过多种指标对商品进行分类，最后根据商品的分类结果我们可以分析出各类商品所具有的特征，可以判断客户喜欢和购买每类商品的可能原因，同时针对各类不同的商品采用不同的营销方式。

有数据表明，在美国有 15%的成年人曾经买过被推荐的商品。所以对于电子商务企业来说，产品推荐系统是电子商务网站运营中不可或缺的一个组成部分。还有数据表明，有产品推荐系统的网站平均总收入的 2%～5%是来自推荐系统。美国电子商务前 10 名都在一定程度上采用了产品推荐系统，而做得最好的亚马逊（Amazon.com）甚至把首页的 70%用来做推荐系统。

如果我们可以针对客户分类采取适当的精准定位，分别强调产品的新鲜度、高质量、低价或者安全性，效果会更好。采用了比较优秀的推荐系统之后，即使用户没有购买推荐的商品，对于

网站的整体收入、转化率和客单价都是有一定提升的。

正如图 10-7 所示，促进用户购买一件产品的因素最主要的原因如下：

- 高质量。
- 低价格。
- 全新产品。
- 安全。

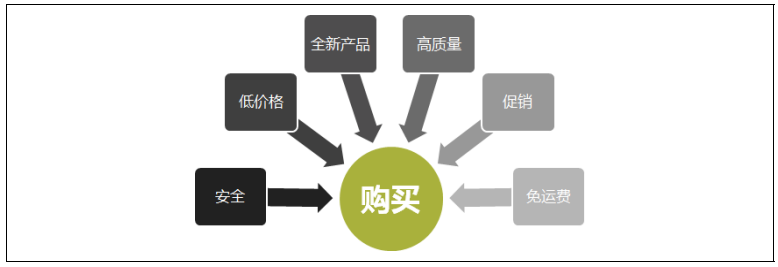



图 10-7　用户购买思考示意图

除了这四个原因之外，还有一些其他的因素。有意思的是，不同网站由于本身定位不同，在这些网站上购买的用户倾向性也不尽相同。比如对于面向高端人群的 A 网站，促进购买的最主要因素可能是高质量和安全，而一个面向折扣消费人群的 B 网站，促进购买的最主要因素可能是低价格和免运费折扣。

拿我们在第 2.1.1 节中提到的美国零售连锁超市 Target 和孕妇的案例来说。从商品数据库的数万类商品中，Target 挖掘出 25 类与怀孕高度相关的商品，制作“怀孕预测”指数，并以此可以推算出预产期，抢先一步将与孕妇相关的产品推送给客户。有一位住在亚特兰大的 23 岁的女性顾客 Susan，在 3 月份购买了按摩乳霜，一个够放置纸尿片的手提包，富含锌钾营养维生素和一块亮蓝色的地毯，那么 Target 的后台程序会显示有超过 80%的可能性是她怀孕了，而且预产期大约在 8 月。因为 Target 有 Susan 的所有消费记录，他们很清楚地知道只需要在周五给她邮寄一张免费星巴克咖啡的消费券，她就会在同一周末来到 Target 实体店中使用这张消费券，那么有很大概率她也会使用她收到的尿不湿

之类的折扣券。有客户的详尽信息，Target 很有把握给客户提供的商品一定是她们喜欢和需要的商品。

Target 的这个案例是基于数据挖掘所做的用户行为分析的结果。当然，这样做是否符合伦理道德标准或者侵犯个人隐私的问题已经在美国各大媒体上讨论得沸沸扬扬，就不在我们评论之列了。读者可以在本书第 11.3 节中看到更多关于互联网隐私方面的讨论。

### 10.3.3 竞争与反竞争分析

竞争情报是指通过合法手段搜集和分析商业竞争中有关商业行为的优势、劣势和机会的信息。互联网为竞争情报工作提供了丰富的信息资源，但是没有一个很好的网络信息挖掘工具便很难获取其中有价值的信息。随着商业竞争的日益激烈，各个企业都纷纷建立了自己的竞争情报系统，以提高自身的竞争力。尤其是在网络环境下，谁忽视了网络信息资源的开发与利用谁就已经失去了领先的机会。

所谓知己知彼百战不殆，在企业竞争情报工作中有两个重要方面：就是获取竞争对手和客户的信息。随着互联网在企业中应用的不断深入，从网上可挖掘的企业信息越来越多，涉及的内容也越来越广泛。从网络信息挖掘技术的实现流程来看，网络信息挖掘不仅仅是像网络信息检索那样只是把符合查询要求的记录返回给用户，这样得到的结果不仅数量庞大，而且包括很多不相关信息。正如前面所提到的，网络信息挖掘不仅能够从互联网上的大量数据中发现信息，而且它还能发现权威站点、有重要价值的“隐藏”信息，并且能够监视和预测用户的访问习惯，这对于企业开展竞争情报工作是非常重要的。

作为防守，反竞争情报子系统是企业竞争情报活动的重要组成部分，忽视竞争对手的竞争情报活动，低估竞争对手搜集竞争情报的能力也会导致企业失去已有的竞争优势。Web 站点是企业与外界进行交流的窗口，同时也是竞争对手获取竞争情报的一个

重要信息源，因此，对它进行监控是企业了解竞争对手的竞争情报的重要途径。在竞争情报计算机系统中，可以充分利用 Web 挖掘技术，通过运用分析访客的 IP 地址、客户端所属域、访问路径分析等 Web 监控技术、统计敏感信息访问率等方法实现对竞争对手的防范，以达到识别竞争对手保护企业敏感信息的目的。

举例而言，有一家电子商务网站，他们通过 IP 地址分析发现竞争对手工作区所在的 IP 段，比如为 172.184.234.1～172.184.234.254。他们准备给网站上的商品做一个大的促销活动，但是不想过早惊动竞争对手，采用的策略是针对不同的访客来源显示不同的页面。对于某个来自 IP 地址段 172.184.234.1～172.184.234.254 中的所有访客，显示旧的产品页面，而对于其他的访客，则采用带有新的促销活动的页面。当然，如果竞争对手通过其他的 IP 地址段访问是可以发现这一问题的，但是至少在一定程度上可以起到对竞争对手的防范。

### 10.3.4　客户还会买什么

哪些商品应该放在一起销售可以提高每次客户购买商品时的客单价呢？电子商务推荐，多是使用关联和协同算法，挖掘不同产品间的关联度。

关联是指确定在一次会话中最可能被购买或浏览的商品，又称市场分析。如果网站在网页中将这些条款放在一起，就可以提醒网站访客购买或浏览可能忘记了的商品。如果在关联的一组商品中有某一项商品是特价，网站很可能会增加同组中其他商品的购买量。

当网站使用静态的目录网页时，也可以使用关联。在这种情况下，网站会依赖厂商选择的且是网站所要查看的第一页目录网页，并提供相关的条款。

我们在网络购物时，经常会在购买完一件商品后看到类似这样的一句话“买了这件商品的顾客还买了”，这句话的下面会展现一系列的商品，这就是最普通的关联商品推荐，通过对所

有客户的消费记录进行分析，找出最常见的商品搭配，并以此来做推荐。

从原理来说关联推荐来源于数据挖掘中的关联规则，从商业角度讲关联规则就是从现实中大量的事物联系中找出既符合实际又能具有一定价值的规则。关联规则挖掘最初多用在大型超市的购物篮分析领域，通过在消费者长期的购买行为所形成的事务数据库中寻找出一些既频繁又可信的商品购买组合，商家利用这些规则来重新安排货物摆放位置。所以在应用领域，关联规则算法又称为购物篮分析（Market Basket Analysis）。随着电子商务的兴起和数据挖掘的广泛应用，关联规则开始在电子商务网站商品推荐系统和其他多个行业进行大量的应用。

关联推荐做得最好的当属电子商务中的大鳄亚马逊（Amazon），他们的个性化推荐机制几乎遍布全站的每一个角落。访问 Amazon 的访客最终下单的比率要高出行业平均比例 50%，这个跟访客进入页面看到的是自己的感兴趣的个性化页面而不是密密麻麻的分类列表不无关系。比如说我们在美国的亚马逊网站上查看《史蒂夫·乔布斯传》的时候，在页面的下方，会出现如图 10-8 所示的推荐。


图 10-8　亚马逊公司产品推介图

一些知名的电子商务网站也从强大的关联规则挖掘中受益。

这些电子购物网站使用关联规则进行挖掘，然后设置用户有意要一起购买的捆绑包。也有一些购物网站使用它们设置相应的交叉销售，也就是购买某种商品的顾客会看到相关的另外一种商品的广告。

关联规则数据挖掘或是购物篮分析转化成算法的要求如下。

- 算法输入：所有的交易数据。
- 算法输出：各个物品之间的常用关联关系。

我们从美国一家办公用品连锁店的案例来看关联关系的数据挖掘和应用。从交易数据中得出大量关联规则，而其中最有价值的一条规则是“同时购买了笔记本电脑和防病毒软件工具的客户中，70%的客户同时购买了该店的 3 年延展服务保障计划（Extended Warranty）”。延展服务保障计划是该店利润率最高的产品。用关联算法表示，结果就是：

{笔记本电脑，防病毒软件}=>{延展服务保障计划}[10%，70%]

那么他们怎么利用这一信息呢？他们根据这次数据分析的结果，做了三件事情：

- 所有的笔记本电脑附近都有防病毒软件的陈列。
- 防病毒软件货架旁边有最新的笔记本电脑宣传册。
- 在店中显眼的位置放置了一个笔记本电脑和防病毒软件的组合，如果用户同时购买这两款特定的商品，会收到该店提供的一个特殊折扣。

互联网和传统商业上的购物篮分析最大的不同在于信息更加丰富。在传统的商家，我们需要知道客户同时购买两件商品的交易信息才能判断该客户对这两件商品都感兴趣。而在互联网上，如果客户访问了某件商品的详细信息，我们可以假设该客户对这件商品 A 是有一定兴趣的，而如果该客户同时也购买了商品 B，那么我们可以建立 A 和 B 之间的关联关系。如果我们需要提高判断的准确性，我们可以增加以下两个条件中的一个：

- 访客在商品 A 上的浏览停留时间大于一个阈值，比如 20s；
- 访客多次访问了商品 A 的页面。

### 10.3.5 哪些客户是我们需要的

怎样的客户是我们最想要保留的呢？从哪些客户身上我们能够取得最大效益？哪些客户是我们能够长期保有的？

我们可以采用目标寻找技术（Targeted Advertising），选择接收特定广告的人群，以增加利润（率），提高商标知名度或增加其他可量化的收入。在网上进行受众目标寻找也必须考虑各种不同的广告费用。如果厂商将广告目标锁定在最可能购买某产品的人群，就可能降低广告费用，并增加投入产出比。比如我们发现某一地区的人群适合我们的产品，就可以根据地理信息确定广告投放目标。如果女性更适合这款产品，就根据性别信息投放广告等。采用数据挖掘技术可以帮助用户选定广告活动的目标标准。

个性化与目标选择相反。目标选择功能是优化查看广告的人的类型，以降低广告费用；个性化的方法选择发给个人的广告，以取得最大成果。通常，在提供的产品或服务有限的情况下我们可以使用基于规则的个性化系统，在面对成千上万的条款时使用自动的系统更加有效。

我们可以通过客户综合价值模型来评估并选出我们最想要保留的客户。客户价值评估模型的搭建，要综合衡量客户五个方面的表现：客户当前贡献度、客户未来贡献度、客户信用度、客户忠诚度以及客户成长潜力。

比如，我们可以构建如下的线性数据模型，从历史数据和结果中找出深层的关系和规律：

```
客户综合价值 = weight_1*客户当前贡献度
     + weight_2*客户未来贡献度
       + weight_3*客户信用度
       + weight_4*客户忠诚度
       + weight_5*及客户成长潜力
```

公式中的权重（weight_1 到 weight_5）是根据历史数据调整的。针对不同的商业目标，我们可以对于公式的各个环节权重做调整。例如我们更加在意短期内的效益，那么可以把和客户当前贡献度相关的权重 weight_1 上调。如果我们需要降低欺诈交易

的比例，那么就可以把和客户信用度相关的权重 weight_3 上调，依此类推。

而客户这五个方面的表现本身也是通过历史数据模型判断出来的，而用以构建模型的数据主要有以下几类：

- 用户最近一次在本商城以及其他竞争商城或者网店购买商品距当前的时间；
- 用户在一段时间内购买本商城商品的频次以及其他网购的频次；
- 用户每次交易平均（Average）客单价和中间值（Mean）客单价；
- 用户单次购买最高支付金额；
- 用户在本商城总购买金额；
- 用户单次购买所覆盖的商品种类；
- 用户平均下单和最终成交比例；
- 用户平均访问商城和下单比例。

我们解释一下平均（Average）客单价和中间值（Mean）客单价的区别。假设用户在商城中购买了三次，客单价分别是 10 元、20 元和 120 元，那么平均客单价是 50 元（数学平均值），而中间值是 20 元（中间的这个数值）。

客户当前贡献度、客户未来贡献度、客户信用度、客户忠诚度以及客户成长潜力这几个方面对于上述数据的要求是不同的，比如客户当前贡献度看重的一定是用户在本商城的总购买金额，其次是用户最近一次在本商城购买距当前的时间，再次是用户单次购买最高支付金额。

## 10.4 电子商务案例

下面我们通过两个案例来看数据挖掘在电子商务中的具体应用。

第一个案例是一家中小电子商务卖家，此类电子商务网店规

模不大，中小电子商务卖家企业通常急需提高数据分析能力，该电子商务企业之前也曾经购买过一些分析工具，但工具的利用效果比较一般。第二个案例的对象是一家海外的大型互联网电子商务企业，我们以它为案例主要来介绍关联推荐的具体应用。

和线下零售不同的是，电子商务网站都有非常丰富的顾客历史数据，包括登录、点击、浏览以及购买等。如果你把数据放在地下室让它们堆满灰尘，这些数据就是一项负资产，它们需要硬件来存储，需要人员来管理，却没有任何使用价值。

## 10.4.1 电子商务企业案例一

天猫就是原来的淘宝商城，是阿里巴巴集团打造的 B2C 式综合性购物网站。天猫通过整合数万家品牌商和生产商，为商家和消费者之间提供一站式解决方案。天猫在整个 B2C 电子商务模式领域一直处于行业领先。在 2012 年，天猫商城的总销售额可能占到整个 B2C 市场的 60%以上。我们的案例正是来自于天猫上的一家品牌商店——Van Doren（范伯伦）网上商城。如图 10-9 所示。


图 10-9 Van Doren（范伯伦）

### 10.4.1.1 问题阐述

范伯伦原来的主要业务是做 OEM 代加工，有一定规模，不过自 2011 年起受到全球经济影响，销售额的增长停滞。他们做线上营销是为了创自己的品牌，希望利用网络的便利性与渗透力迅速扩大销售额。我们接触过很多此类公司，原来是纯粹的厂家，而现在想通过电子商务领域的拓展找出一条提高收入的新途径。

范伯伦电子商城没有一开始就构建自己的网上电子商城，而

挂靠在天猫上的主要原因是想借力于天猫现有的流量和客户资源。在成功塑造了自己的品牌之后再怎么做就是下一步的事情了。我们熟知的电子商务品牌麦包包，就是先利用淘宝把自己的品牌树立了之后，然后双管齐下，一方面继续运营淘宝上的品牌店，另一方面开始经营自己的独立网上商城。

如我们在第 1 章中所述，数据挖掘在直效行销领域是非常成功的，而从一种角度来看，我们可以把厂家直接运营电子商务的公司当做直效行销公司，因为他们也是直接面对消费者。我们拿来做案例的这家电子商务企业，就是其中的一家。不同的是，在电子商务网站上面，我们可以获取比在直效行销领域更多的关于客户的信息，充分利用这些信息可以使我们更加有效的把握客户。

#### 10.4.1.2　转化为业务问题

正如本书一直强调的那样，任何的数据挖掘活动都起始于对业务的理解，只有找出需要解决的具体商业问题，我们才能开始考虑用什么方式来分析挖掘。结合前文在了解客户上的阐述和网店的实际需求，我们需要为该商家回答以下几个问题。

（1）我们的客户来源分布是怎样的？

（2）该如何安排商品的上下架时间？

（3）店铺里哪种或哪些商品是最受欢迎的，即网店应该重点推荐哪些商品？

（4）网店拥有一些会员客户的信息，该如何利用这些信息找出潜在客户？如何利用网络行为的可追踪性与高互动特质，实现一对一行销的理念？

在回答了这些问题的基础上，我们在本节的其余部分给大家讲述如何利用数据，帮助客户提升销售和增加利润。

#### 10.4.1.3　解决问题

我们来看一下如何对前面列出的每个具体问题进行分析，来看如何可以通过数据挖掘给这家电子商务企业创造价值。

而在这之前，我们先来看下如何制定运营指标。

我们从运营指标开始，设立包括用户行为指标、用户价值指标和营销活动指标在内的一系列指标，把大的运营目标分解成阶段性和局部性可以实现的目标。如图 10-10 所示。

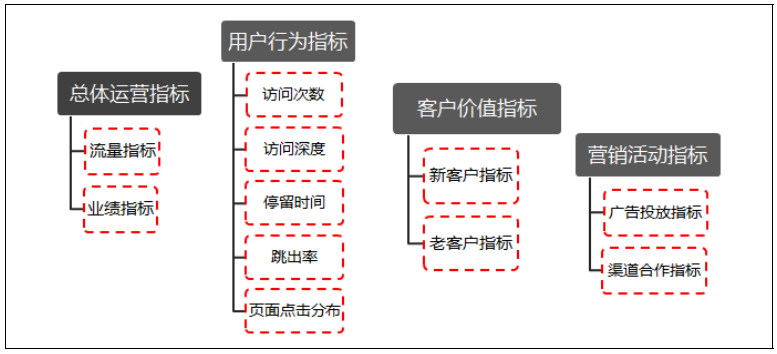



图 10-10　电子商务企业运营指标示意图

图 10-10 中总体运营指标只看两个方面：流量指标和业绩指标。我们可以把大的目标分解成用户行业指标、客户价值指标和营销活动指标三类指标，而这三类指标中分别又有一些各自的子指标。

### 1．客户来源分析

一般的互联网数据分析工具中都有对于网站访客流量来源的分析功能，可以直接得出一定结果。而本案例中的网上商城是构筑在淘宝天猫之上的，所以我们只能采用淘宝本身提供的和淘宝开放平台上的工具来做数据分析。图 10-11 和图 10-13 中的流量来源和访客地理位置分布就是从店铺的淘宝量子恒道工具中直接获取到的。

图 10-11 基本阐明了最近 7 天网店的客户通常采用何种方式进入网店。这里我们可以看到，因为这家网店的店铺优化做的还可以，来自淘宝的免费流量占到了 36.67%。同时因为做了一定时间，有一定的知名度，所以自主访问的比例超过了 20%，占到了 22.41%。通常来说，如果商品的品质和价格吸引人，网站呈良性发展，那么淘宝免费流量和自主访问的所占比例都会稳步提高。

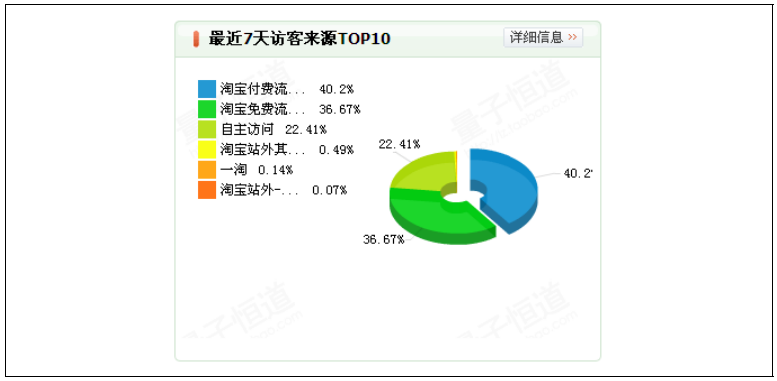



图 10-11　最近 7 天访客来源分布示意图

因为这个网店是在天猫站内，所以来自站外的访问量不是特别多。而对于独立的网上电子商城，基于搜索引擎的流量会占到相对较高的比例。来自搜索的流量同样也要分成自然搜索流量和搜索关键词广告流量。比如凡客诚品每天在百度上要投放数个亿的搜索关键词广告，同时他们也会做搜索引擎优化（SEO），使他们的产品尽量出现在搜索引擎搜索结果的前列。

我们看图 10-12。在图 10-12 中显示的是我们在百度上搜索“凡客诚品衬衫”所出现的搜索结果。其中上面前三条是百度推广链接，每一次用户点击了其中任何一条，百度都会从相应的商家这里扣除一定的广告费，而费用高低与关键词的热门程度有很大关系。普通的是几元人民币，而热门的关键词可以高达几十甚至数百元。在百度上的搜索结果都是动态的，也就是说，你做的每一次搜索，结果都可能会有所不同。

在图 10-12 中，第一条链接和第四条链接都是直接到凡客的官方网站。如果用户点击的是第四条链接，那么恭喜陈年，这次点击是免费的搜索流量。如果用户点击的是第一条链接，那么凡客诚品需要为这次点击付出 1.8 元人民币（大致估算）。

对于网站做 SEO，可以使得网站的主页或站内的其他一些页面在自然搜索结果中尽量靠前，而点击这些链接是不需要计费的。如果网站做 SEM，那么通过付一定的费用，所对应的链接也会出现在搜索结果中。关于搜索引擎优化（SEO，Search Engine

Optimization）和搜索引擎营销（SEM，Search Engine Marketing）又是两个很独立自成一体的专题，不在本书的讨论范围之内。


图 10-12　搜索推广示意图

对于独立的网上商城，也就是说它们不在天猫这类综合电子商城内的，我们可以分析出用户是点击了什么链接进入到商城的。如果是来自于搜索引擎，我们还可以分析出用户是通过搜索什么关键词进入到商城的。

图 10-13 的数据显示了最近 7 天网店的客户分别来自哪个省份。在图中我们看到，访问该网上商城最多的访客来自于广东，约占 19%，而其次是来自于北京和江苏，分别占 11.25%和 8.85%。而值得注意的是来自该品牌其中一个重点目标城市上海的流量并不太多，只占 3.66%。

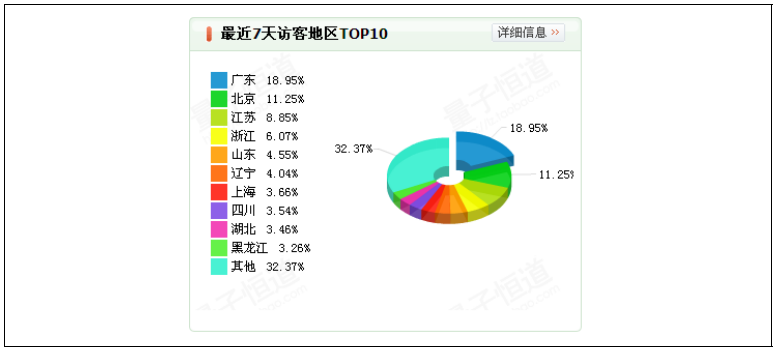



图 10-13　最近 7 天访客来源地理位置分布示意图

发现来自上海的流量占比不高时，我们可以做两种假设：

- 是否上海的受众不喜欢我们推出的产品？
- 是否对于上海的推广力度不够？

为了验证第一种假设，我们可以做客户调研，看是否增加某些关键词的商品描述和图片可以提升客户留存。而对于第二种假设，我们可以针对上海地区投放广告，并监测广告的转化率和效果。

对于单个访客在互联网上的来源分析，可能是没有太大意义的。但是综合一段时间内所有访客的来源信息，我们可以做趋势分析，从而决定在互联网上投放广告和资源的力度及方向。

### 2．用序列算法分析商品上架时间

前文也有提到合理安排商品的上下架时间的重要性，而要了解最佳上架时间，就要知道流量高峰的具体时间段。具体做法是通过时间序列对以往的流量数据进行分析，并用此预测未来一周的流量。时间序列数据最直观的表现方式是通过图。

在图 10-14 中展示了该店铺某一个月每天的浏览量数据。图中的纵轴是访问量，而横轴是日期，并以每周作为分隔，图中横轴上的数字表示的是周。每一个点都表示一天的数据。

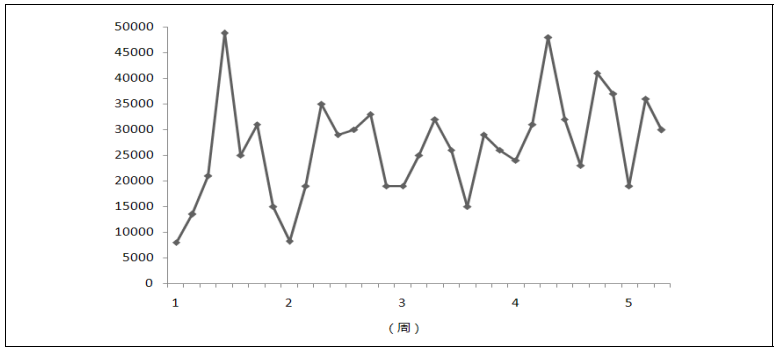



图 10-14　网上商城月流量示意图

图 10-14 中展示的是该网店某个月份浏览量的时间序列数据。图中可以看到在该月 4 号和 23 号分别达到了这段时间流量

的峰值，约 50000 次访问，而最低的几天浏览量只有峰值的 15%，约在 7500 次左右。从这个时间序列图可以看出每周的浏览量都会有一个先上升再下降的过程。

在本书的 4.6 节关于序列算法的描述中我们解释过可以使用时间序列算法做预测分析。利用该店铺全部的历史数据，采用时间序列中的自回归算法，我们可以大体预测接下来一周的浏览量。

图 10-15 是对于某一个月按小时来分隔的周时间序列图，其中的实线部分是已经发生的实际数字，而虚线部分表示的是我们的预测数字。

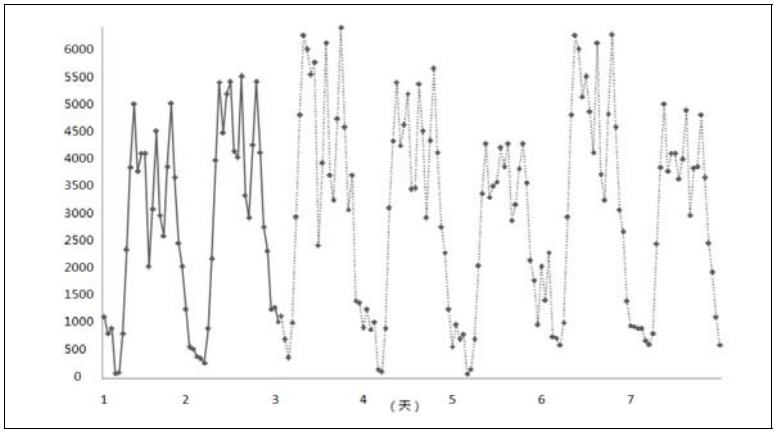



图 10-15　周流量预测示意图

图 10-15 周流量示意图中的预测值是按小时来划分的，每天有 24 小时，而每个小时是一个数据点。周一到周二的 48 个数据点是已经发生的，预测从周三开始。通过时间序列最后得出该店铺从周三开始未来一周的浏览量，周三和周六的浏览量可能是最高的，再结合按时流量数据，可以发现每天的上午 10 点、中午 1 点、下午 4 点、晚上 8 点的浏览量最高，故可以考虑在周三和周六的这几段时间上架宝贝（所谓宝贝，是淘宝和天猫对于网店商品的专门用语，地球人一般都懂的）。

考虑完流量数据之后，我们再看的另一个数据是转化率的问题。转化率越高，浏览者产生消费的可能性就越高。通过进一步

对历史数据的分析之后，我们发现周六晚上 8 点左右的转化率是最高的，其次是周三上午 10 点和周三下午 1 点。所以我们最终提出的建议是在三个时间段进行上架宝贝的工作：

- 50%的商品上架时间是周六晚上 8 点；
- 30%的商品上架时间是周三上午 10 点；
- 20%的商品上架时间是周三下午 1 点。

选择在这些时间上下架的主要原因在前文中提到过，是给宝贝增加被浏览进而产生销售的机会。通过上述的上下架时间调整，一个月之后，我们成功地把该网店的销售额提升了约 45%。而把所有商品都集中在这三个时间段的原因是商城本身的流量并不大，而且商品的品类也是比较少的。在商品丰富和平均流量上升之后，我们宝贝上架的策略也需要进一步调整。

### 3．用聚类算法对商品分类

这里的商品分类不是指商品类别上的分类（图 10-16），而是对于商品在销售上产生价值的深度分类。

图 10-16　网上商城商品示意图

我们采用了该店铺一段时间的数据来说明对商品分类的方法。这里我们采用的数据挖掘方法是聚类算法，而不是分类算法，因为我们事先并不知道会将宝贝分成多少类。作为示例，宝贝用来做聚类的属性，包括以下五个方面：宝贝页浏览量、跳失率、拍下件数、拍下总金额、宝贝页收藏量。在不同案例中实际运用

数据挖掘的情况是，我们通常会采用更多维度的数据，但是聚类的结果可能从表面上不太容易解释。

我们综合这些属性，对商品进行聚类分析。在每个宝贝的多维度数据上，我们采用在第 4 章中讲述过的 K-means 聚类算法，把宝贝分成 3 类，得出了表 10-1。为了做对比，这里采用了一个月的数据。

表 10-1　优化前宝贝分类表

| 类别 | 宝贝页浏览量 | 跳 失 率 | 拍下件数 | 拍下总金额 | 宝贝页收藏量 |
|---|---|---|---|---|---|
| 1 | 1621.00 | 0.6922222 | 122.6666667 | 4058.39 | 118.6667 |
| 2 | 129.98 | 0.5420652 | 1.217391 | 28.47 | 1.6522 |
| 3 | 903.27 | 0.58227273 | 8.85455 | 269.79 | 7.1818 |

表 10-1 中的聚类结果表示所有宝贝被分成了 3 类，表中的数值表示具体类别中宝贝具体属性的均值，例如第一行的 1621 就表示在第一个类别中宝贝页平均被浏览了 1621 次，第 2 行第 3 列的 0.5420652 表示宝贝页的平均跳出率是 54.20652%。从数值看来，第一类宝贝在各个属性的表现都很好，属于店铺应当重点推荐的商品，但是该类宝贝的跳失率为 69.2%，高于其他两类的 54.2%和 58.2%，也就是说这类宝贝的高浏览量并没有完全起到它应有的作用，大量客户进入宝贝页面后就退出来了，所以我们在重点推荐这类宝贝的同时还要注意优化宝贝页或通过促销活动来使客户留下来，使跳失率降低。

第二类宝贝的各个指标偏低，虽然跳失率较低，但是浏览量也不多。这类商品可能并不受客户欢迎，此时我们可以从多方分析不受欢迎的原因，例如可以通过网上的调查问卷或者直接让客服询问客户的方式来了解客户不喜欢此类宝贝的原因并以此来改进或淘汰这类商品。我们发现其中的一个情况是商品的图片制作过于粗糙或者商品说明的介绍性文字过于简单，使商品的品质没能体现出来。

第三类商品各项属性都处在中间，而且这类商品的数量是最

多的，应该说店铺要想增加销量重点得靠这类商品，可以通过打折促销、橱窗推荐等方式来提高销量。

我们给该电子商务企业的意见是对第一类宝贝的描述进行优化，同时重新拍摄多幅产品照片。因为该电子商务企业尚未上架的宝贝比较多，我们建议他们把第二类宝贝中除了个别有特殊情况的宝贝之外全部下架，换上新的宝贝，而且也对新的宝贝做描述优化和产品照片的重新拍摄。

在做了针对性的产品图片和描述修改优化之后，过了一段时间，我们重新对宝贝做 K-means 聚类分析，结果依然是可以把所有的宝贝分为 3 类，如表 10-2 所示。为了和表 10-1 做同质化比较，这里依然采用了一个月的数据。

表 10-2　优化后宝贝分类表

| 类别 | 宝贝页浏览量 | 跳 失 率 | 拍下件数 | 拍下总金额 | 宝贝页收藏量 |
|---|---|---|---|---|---|
| 1 | 2372.29 | 0.5823 | 191.82 | 6188.35 | 159.84 |
| 2 | 188.97 | 0.6567 | 17.57 | 1273.25 | 11.35 |
| 3 | 1072.52 | 0.5432 | 25.99 | 1012.54 | 19.88 |

从表 10-2 中我们可以看出，这次三类宝贝的数据和前面的表 10-1 相比，都有明显的改善。特别是第一类高质量的宝贝，平均浏览量提升到 2372.29 次，平均拍下件数上升到 191.82 次，总金额也上升到 6188.35 元，而跳出率却降低到 58.23%。这里的第二类宝贝，虽然浏览量还是比较少，但是平均被拍下的件数从表 10-1 中的 1.22 次上升到 17.57 次，上升了 15 倍。

### 4．用聚类算法提升会员管理

该电子商务企业在引流方面的统计数据大概如下：

- 一个新客户根据来源不同，引入的成本约 40～80 元人民币；
- 淘宝付费推广流量，主要是淘宝直通车，大概能做到 1UV（独立访客）=1.1 元钱；

- 首焦（淘宝首页上的广告位）和其他硬广贵而且效果一般，ROI（投资回报率）在 1.8～2 之间，也就是说投 10000 元，大概能在网上商城做到 18000～20000 元的销售额；
- 淘宝内部自然搜索 SEO 引流，流量虽然稳定，但是增长缓慢，而且大量是引导到打折较多的爆款上，利润有限；
- 在淘宝修改聚划算规则之后，参与聚划算活动的成本过高，聚划算的活动收入基本不能覆盖活动支出，而且大部分买聚划算爆款的客户是因为价格因素来购买的，除了能够跑量，没有利润可言。

引入一个新客户的成本居高不下，而维系一个老客户复购的成本，是引入一个新客户的几分之一。老客户客单价显然高于新客户，老客户了解店铺产品，对客服压力相对较小，退换货率更低。因此，我们需要建立品牌/店铺的 CRM（客户管理）体系。

RFM 作为客户关系管理领域中一种常见的模型近年来被大量运用于客户价值评价中，模型主要通过客户最近一次购买时间距今有多久（R）、客户在最近一段时间内购买的次数（F）、客户在最近一段时间内购买的金额（M）这三个因素来对客户价值进行评分。分值越高，客户价值越大，这类客户往往是商店应该重点关注的客户。在 8.2.4 节邮件营销中我们已经详细描述了 RFM 模型的细节，在此不再赘述。

在电子商务对于 RFM 模型的实际应用中，我们常常会优先考虑客户最近一次购买时间距今有多久（R）这一因素，一般来说购买时间越近的客户对电子商务商品的体验越清晰，所以会更注意同一公司的商品信息，此时如果加强对客户的沟通往往会增加客户的二次购买机会。比如可以采用交叉销售或追加销售的策略，推荐与客户购买需求相关度高的商品，或者提供额外的重复购买奖励。紧随因素 R 的是因素 F，最近一段时间内购买的次数，购买次数越多的客户对商家的认可度也就越高，这也是重点关注的客户。然而购买金额 M 则主要取决于客户购买商品的价格，与客户价值的相关性是最弱的，不过作为因素之一考虑在提高客户分类精度上也有一定的益处。

结合本案例中网上商城的实际情况，我们采用近半年的客户购买数据来做聚类分析。代表 RFM 模型的三个因素分别是半年内购买次数、半年内消费总金额和上次消费距今时间间隔。

我们对于网上商城中的客户应用 K-means 聚类算法，得到以下 4 个类别，如表 10-3 所示。

表 10-3　RFM 客户聚类表

| 类　　别 | 半年内购买次数 | 消费金额 | 上次消费距今时间间隔（天） |
|---|---|---|---|
| 1 | 2.354962 | 58.00969 | 22.62977 |
| 2 | 1.763158 | 147.14053 | 128.71053 |
| 3 | 2.230769 | 668.40385 | 100.15385 |
| 4 | 2.543210 | 300.41654 | 35.88889 |

从表 10-3 的客户分类结果来看，我们的客户大致分成四类。大概可以看出第一和第四类的客户半年内购买次数最多，上次消费距今间隔最短，此类客户属于重点维护对象，而第二类和第三类客户消费距今时间间隔较长，当促销资源有限时此两类客户可以忽略。

从 CRM 客户管理系统库中的数据表明，随着客户购买次数的增加，客单价与客单件都有逐渐提升。通过促销唤醒一个回头客户比新客户创造更多的盈余，从一次购买到二次购买的回头率，我们是最需要提升的环节。但是二次到三次或是多次就比较高了，多于二次消费的客户我们可以定义为忠诚客户。

我们在 CRM 客户管理系统库中把购买过 2 次以上的客户拉出来，然后计算他们购买第 2 次时距离第 1 次购买的时间间隔。我们发现，40%的客户在首次购买后的一个月内进行复购，而复购发生在一年以上的只占 5%。这说明距离首次购买的时间越长，复购的可能性越少。

按照客户管理原则，我们统计客户购买习惯，按照购买间隔划分成 4 个生命周期（与邮件营销不同，这里不包含潜在用户这个周期）。

这里采用的算法是在第2章和第4章中都做过描述的决策树算法。我们通过决策树算法调整客户生命周期，最后制订针对不同生命周期的营销策略如下。

- 活跃期（1 个月内）：保证接触频次，但不做促销刺激，如果客户回头，我们可以保证利润率；
- 沉默期（2 到 4 个月）：保证接触频次，给予少量的营销折扣；
- 睡眠期（5 到 10 个月）：控制有限接触，通过活动，给予较大折扣挽回客户；
- 流失期（11 个月以上）：只在例如“双 11”之类的大活动时通知。

通过划分生命周期，我们解决了大部分客户细分的问题，当然如果店铺的产品线和价格区间足够宽，还可以根据首次购买客单价进一步对不同生命周期的新客户进行细分。因为产品不同，不是每家店铺的生命周期都是如上所述的，而且根据每个月不同的销售数据，生命周期的划分也会有所不同，但大致的范围是差别不大的。

通过在该电子商城上执行 CRM 生命周期管理，客户的复购率从 11.5%提升到 21.3%，几乎提升了一倍。而且我们发现，首次购买客单价在 250 元以上的客户，复购率达到了 29%。现在客户数量还没有到达一定规模，通过决策树对首次购买客单价进行细分出的结果变化还是有比较多的随机性。

### 5．用数据提升整体营销

我们用图 10-17 来展示该网上商城的营销基本策略，与其他的网上商城相比，基本策略都差不多，唯一不同的是在营销的每个环节，我们都是以数据作为基础。

通过数据挖掘，可以大幅提升转化率，从而使营销中的投资回报率 ROI 变得可以接受。我们可以做一个简单的计算，如果通过数据挖掘和分析，优化在淘宝直通车和硬广中的投放费用，使转化率能达到 1.39%，那么当店铺平均客单价为￥210，流量

UV 达到 135,000 时，总收入则可以达到￥394,000，通过成本核算，我们可以放心大胆地投放广告。


图 10-17　网站营销基本策略示意图

## 10.4.2　电子商务企业案例二

电子商务的第二个案例来自于美国一家大型的主要销售食品的连锁超市，他们既有线下门店，也有网上商店。他们拥有大量包括线上和线下的数据，但是对于这些数据的应用没能够赶上数据的积累。

### 10.4.2.1　通过日志挖掘做网站访客分析

他们每天获取的数据主要分为两类：一类是网站内容和商品基础数据容量；另一类是一些网络上的用户行为偏好数据，包括合作伙伴关联网站的内容，记录的数据包括用户曾经浏览过什么、收藏过什么、购买过什么等数据。这两类基本是离消费者最近的数据，可以说是他们电子商务最核心的数据。这部分数据每天的积累大约在 GB 级的规模，而整体数据仓库中的数据已经在数十个 TB。他们的线下店做得相当不错，但是互联网上的网店，虽然已经投入了大量的资金，但是销售增长没能达到预期。

对该公司的庞大数据资料集合，我们首先做的是最简单的日志分析，而分析所用的方法都是在第 7 章中提到过的，包括对错误代码的数据统计和频繁路径分析等。网站本身在技术上没有问题，服务器资源也足够。

我们先来看一下该网站的基本访问信息，如图 10-18 所示。

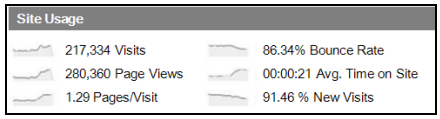



图 10-18　网站访问基本数据示意图

- 访问量（Visit）：217,334 次。
- 跳出率（Bounce Rate）：86.34%。
- 页面访问量（Page Views）：280,360 页。
- 平均停留时间（Average Time on Site）：21 秒。
- 平均每次访问页面深度（Pages/Visit）：1.29。
- 全新访客占比（% New Visits）：91.46%。

这些数字构成了网站访问的基本数据，也就是说我们通常可以用这些数据来衡量一个网站的访客和访问的质量。访问量、页面访问量、平均停留时间、平均每次访问页面深度这四个数字是越大越好。跳出率是越低越好，全新访客占比这个数字在通常情况下也是越低越好。

我们来看某一周网站访问的统计数据，如图 10-19 所示。

| Page Views per Visit | Visit | % |
|---|---|---|
| 1page(s) | 2270823 | 86.1062% |
| 2page(s) | 328849 | 12.4695% |
| 3page(s) | 29833 | 1.1312% |
| 4page(s) | 1812 | 0.0687% |
| 5page(s) | 891 | 0.0338% |
| 6page(s) | 642 | 0.0243% |
| 7page(s) | 429 | 0.0163% |
| 8page(s) | 283 | 0.0107% |
| 9page(s) | 172 | 0.0065% |
| 10page(s) | 108 | 0.0041% |
| subtotal | 2633842 | 99.87% |
| total | 2637234 | 100% |

图 10-19　页面访问量示意图

从图 10-19 这个一周的网站访问数据上直观来看，到网站上只访问 1 个页面的访客占比很高，有 86.1062%，而访问 1 到 2 个页面的访客占比例 98.58%，也就是说只有 1.42%的访客访问了超过 3 个页面以上。这是一个示例，而在实际的操作中我们是

定期统计一个时间段的数据，比如一天、一周、一个月，或者在某次网站改版之后的若干天等。

我们做网站分析的总体诊断结果如下：

- 网站的设计优化，尤其是着陆页的设计上做得不够。所谓着陆页（Landing Page），指的是网站中的一个市场营销专用页面，通常是搜索引擎或是其他广告所指向的页面。着陆页做得吸引力不够，很直接的结果就是相对较高的跳出率。该网站的跳出率达到了 85%以上。
- 用户在网站上平均访问深度不到 1.3，也就是说每四个访问网站的用户，平均只有一个人看到了网站上的第二个页面之后。访问深度指的是访客在网站上依次浏览的网页数量。
- 网站重复访客较少，也就是说广告资金投入吸引新的访客，但是这些新的访客转化成老客户的比例较低。
- 用户在网站上找不到他们想要的商品。我们需要把商城的站内搜索功能提到显著的位置上，让找不到合适商品的客户可以利用搜索功能找到他们想要的商品。

针对以上这些诊断，我们相应地采取一些措施，可以提高用户的停留时间和客户的转化率。对于电子商务网站，用户在一个网站上的平均停留时间和他们对网站的平均贡献值是成正比的。

在我们做了所有的修改之后，网站的基本数据层面有了很大的改观。请看网站改版之后的基本数据，如图 10-20 所示。

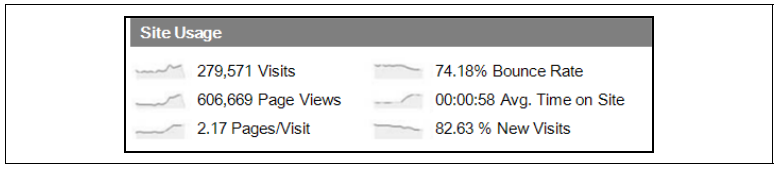



图 10-20　网站访问基本数据示意图

把图 10-20 和图 10-18 中的网站基本数据做对比，我们可以明显发现：

- 访客访问页面的数量增加了。

- 跳出率降低了。
- 停留时间加长了。
- 每次平均访问页面增加了。
- 全新访客比例减少了。

如果做网站优化之后没有太多效果，可能的原因：一是你的网站可能已经比较完美了，提升空间有限；而另外一个原因是你在做网站分析时可能忽略了网站的一些特质，需要回头再做一次分析。

#### 10.4.2.2 决策树数据挖掘的应用

本案例中的公司商品数据信息非常完备，因为它包含了线上和线下所有呈现的商品。对于每一件商品都有上百个属性，比如：商品的名称、厂家、出产地、原料、用途、颜色、年均销售量、月均销售量、周最大销售量、图片的类型、图片的个数、产品描述等。

对于商品库中的商品，我们应用了决策树算法来分析商品在网站上的受欢迎程度，步骤如下：

第一步，我们对商品属性做了数据转换之后保留了所有可以保留的属性。

第二步，把商品分成训练集和测试集。

第三步，统计一定时间内的数据交易集销量，由高到低把商品分为“高销量商品”、“普通销量商品”和“低销量商品”三类。

第四步，采用决策树生成算法来对训练集中的商品分类。

第五步，应用测试集中的商品验证生成的决策树，结果基本一致，那么算法执行完成。

自动生成的决策树相当复杂，我们就不在这里展示了。在叶子节点中产生的关于商品销量的规则，我们需要专业的商场运营人员来解释。不过其中关于“高销量商品”的规则中，我们归纳出以下两条，是很有些意思的：

- 如果“视频信息”一栏中有“是”，那么该商品是“高销量商品”；

- 如果“照片数量”超过 5，那么该商品是“高销量商品”。

如果一张照片等于十句话，那么一个视频可能等于一篇散文。所以视频对于商品有这样的效果是可以理解的。事实上，凡是有视频信息的商品，销量都比没有视频信息的同类商品要高30%以上。仅从这个数据来看，如果我们给某个商品加上视频推荐信息和 5 张以上的照片，商品就有可能成为高销量商品。当然，我们如果给所有的商品都加上视频和足够多的照片，决策树算法的结果可能就不一样了。

最后我们采取的方法是重新整理了商品的相关照片和视频信息，尽量多地把资料库中已有的照片和视频加到网站上。

#### 10.4.2.3　关联规则数据挖掘的应用

我们以下面一组数据来描述关联规则在这家公司的应用。关联规则的原理在 4.5 节中已经详细介绍，这里主要阐述应用过程。

商品推荐是和商品相关的，所以对于每一件商品，系统都会尽量设置它的关联商品。当用户选取了某一个商品，在网页的下方会出现根据关联算法做出的商品推荐。我们来看一下这些关联规则是如何产生的。表 10-4 中列出的原始数据是该公司某一个月的订单数据示意表，每行都包括了订单的部分具体信息，其中 Order Number 是订单号，Product Sku 是产品的编号，Product Type 是产品类型，Qty 是 Quantity 数量的缩写。

表 10-4　食品公司订单表

| Order Number | Product Sku | Product Type | Qty |
|---|---|---|---|
| 12071103756 | E3561 | Yogurt | 9 |
| 12071103756 | F1038 | Soda | 12 |
| 12071103757 | M201 | whole milk | 1 |
| 12071103758 | B1121 | chocolate | 1 |
| 12071103762 | B1129 | Bakery | 1 |
| 12071103762 | E3561 | Yogurt | 4 |
| 12071103762 | I1001 | ice cream | 1 |

表 10-4 是数据库中的展示，所以一个订单已经分解到数据库合适的表达方式。比如订单 12071103756 中包含的是：

{9 个产品标号为 E3561 的 Yogurt，12 个产品标号为 F1038 的 Soda}

而在数据库中，该信息是由表 10-4 中第二行和第三行所表示的：

```
12071103756  E3561    Yogurt   9
12071103756  F1038    Soda     12
```

而订单 12071103762 中包含的是：

{1 个产品标号为 B1129 的 Bakery，4 个产品标号为 E3561 的 Yogurt 和 1 个产品标号为 I1001 的 ice cream}

在数据库中，该信息是由表 10-4 中最后三行所表示的：

```
12071103762  B1129    Bakery   1
12071103762  E3561    Yogurt   4
12071103762  I1001    ice cream    1
```

我们需要挖掘出客户购买的商品之间的关联性，也就是要知道客户在该超市购买了某件商品之外，还买了什么其他的商品。当挖掘出这些关联性之后，如果客户购买了一件符合关联商品条件的商品，就可以据此得出对应的关联商品并对客户进行推荐了。

当然，这里需要注意的是关联关系是有向的，也就是说如果客户购买了商品 A，我们按规则推荐了商品 B，但是反之不一定，如果客户购买商品 B，我们不一定会推荐商品 A。

虽然目前各电子商务网站的推荐系统运用的算法各不同，也有很多是根据协同过滤（根据客户的喜好挖掘出与其最相近的客户购买的商品进行推荐）设计推荐系统，但基于关联规则的推荐方式不失为一个好的策略。而这家超市的实际情况是有大量的线下数据不包含购买方信息，也就是说我们从购买交易数据中不能找到客户的信息。

通过对类似表 10-4 的数据进行关联规则挖掘，我们得到了一批符合条件的关联规则。现在假如有个客户购买了甜圈（Donut），我们应该推荐什么商品呢？通过对关联规则的检索，

我们发现了以下有关甜圈商品的规则，如表 10-5 所示。

表 10-5　关联规则结果表

| 规则号 | 左 规 则 | 右 规 则 | 支 持 度 | 置 信 度 | 提 升 度 |
|---|---|---|---|---|---|
| 1 | Coffee | Donut | 0.03067616 | 0.1711287 | 1.996287 |
| 2 | ice cream | Donut | 0.01457838 | 0.1613063 | 1.988261 |
| 3 | Sausage | Donut | 0.01250635 | 0.1352169 | 1.946236 |
| 4 | tropical fruit | Donut | 0.01321814 | 0.1259687 | 1.617191 |
| 5 | Shrimp Cocktail | Donut | 0.01498119 | 0.1407833 | 1.132388 |
| 6 | Soda | Donut | 0.02146728 | 0.1206997 | 1.358365 |
| 7 | Yogurt | Donut | 0.01769192 | 0.1268222 | 1.885681 |
| 8 | chocolate | Donut | 0.02096560 | 0.1138751 | 1.279956 |
| 9 | Carrot | Donut | 0.02257265 | 0.1183579 | 1.521235 |
| 10 | whole milk | Donut | 0.03326860 | 0.1301236 | 1.902717 |

这个结果显示了所有与甜圈有关的商品及各自的支持度、置信度和提升度，实际应用中常常通过提升度大小来评判规则的可用性。提升度即 Lift 值，是用来衡量关联规则中某条规则的效果的，简而言之，提升度越大越好。

假使我们设定每次为客户推荐 5 种商品。通过对商品的提升度进行排序，从高到低的前 5 名依次是以下这些商品：

Coffee（咖啡）　　1.996287
Ice Cream（冰淇淋）　1.988261
Sausage（香肠）　　1.946236
Whole Milk（全脂牛奶）1.902717
Yogurt（酸奶）　　1.885681

那么当某位客户购买了甜圈（Donut），就可以在“购买了该商品的顾客还买了”这类提示性标语下推荐：Coffee（咖啡）、Ice Cream（冰淇淋）、Sausage（香肠）、Whole Milk（全脂牛奶）、Yogurt（酸奶）这 5 类商品。

在本节的描述中我们其实把实际问题简化了，因为只是简单做好商品类别的推介是不够的，我们需要细化到每一个具体商

品。这时需要的是多层级概念的数据关联算法，而且根据结果的支持度按照商业规则做相应的商品调整。举例来说，如果对于某件商品，Milk（牛奶）的相关支持度非常高，那么我们可以选择推荐利润较高的 Animal Farm organic milk（动物庄园的纯天然牛奶），以促进平均客户整体收益。

通过类似上述的关联规则挖掘，我们针对该公司的全部商品数据做了挖掘，对 15%的商品找出满足最小支持度的关联规则。根据这些规则，该公司修订了这些商品在网络上的关联产品推荐。

## 10.5 本章相关资源

- 本章相关参考文献：

[1] 张朝晖, 陆玉昌, 张钹. 发掘多值属性的关联规则. 软件学报, 1998, 9（11）：801-805.

[2] 铁治欣, 陈奇, 俞瑞钊. 关联规则采掘综述[J]. 计算机应用研究, 2000, 17（1）：1-5.

[3] 潘立武, 王保保, 李绪成. 零售业销售数据关联规则挖掘算法关键思想研究[J]. 信阳师范学院学报（自然科学版）, 2003, 16（1）：90-92.

[4] 冯玉才, 冯剑林. 关联规则的增量式更新算法. 软件学报, 1998, 9（4）：301-306.

[5] 肖劲橙, 林子禹, 毛超. 关联规则在零售商业的应用[J]. 计算机工程, 2004, 30（2）：301-306.

[6] 李宝东, 宋瀚涛. 关联规则增量更新算法研究[J]. 计算机工程与应用, 2002, 38（23）：6-8.

[7] 张伟, 郑涛, 李辉. 一种并行化的分组关联规则算法[J]. 计算机工程, 2004, 30（22）：84-86.

[8] 李逸波, 于吉红, 白晓明. 合理选择数据挖掘工具. 计算机与信息技术, 2005, 16.

[9] 邹丽, 孙辉, 李浩. 分布式系统下挖掘关联规则的两种方案. 计算机应用研究, 2006, 1：77-78.

[10] 蔡伟杰, 张晓辉, 朱建秋等. 关联规则挖掘综述. 计算机工程, 2004, 27（5）.

[11] Brin S，Motwani R，Silverstein C．Beyond market：Generalizing association rules to correlations[A]. Processing of the ACM SIGM OD Conference 1997 [C]．New York：ACM Press，1997．265-2.

[12] Agrawal，Data Mining association rules between sets of items in large databases．In：Proceedings of ACM SIGMOD Conference on Management of Data，Washington，DC May 1993. 207-216.

[13] Sampling large databases for association rules, Hannu Toivonen, In Proceedings of the 22nd international conference on very large databases, Bombay, India, 1996, 134-145.

- 本章相关网址：

[1] http://bbs.paidai.com/

[2] http://eguan.cn

[3] http://www.iresearch.cn/

[4] http://www.itongji.cn

[5] http://www.amazon.com

# 第 11 章
# 数据挖掘和 Web 挖掘

本书讲述的是大数据挖掘，而我们所提到的大数据挖掘中的“大”字，说的就是互联网。

哪些内容、标题、优惠、代言人、广告语是人气最旺的？网站的主要访客是哪些人？什么原因吸引用户前来网站？什么原因能够留住用户？用户对于网站上什么内容最感兴趣？如何从大量网络所得数据中找出让网站运作更有效率的操作因素？以上种种皆属 Web 挖掘分析之范畴。

前文中的很多案例也都是 Web 挖掘。而 Web 挖掘不仅只限于我们在第 7 章中讲述的日志文件分析，第 9 章中讲述的互联网广告和第 10 章中讲述的电子商务，凡网络上的咨询服务、财务服务、通信服务、政府机关、医疗咨询、远程教学，等等，只要通过网络链接的数据信息够大够完整，数据挖掘都可以发挥作用。

在合适的算法基础上，我们可以远程连通处于各个不同地理位置的数据仓库，实施大规模的数据挖掘。我们可以整合外部来源数据让分析功能发挥地更深更广，除了服务器上的日志文件、会员填表数据、线上调查数据、线上交易数据等由网络直接取得的数据之外，结合实体世界累积时间更久、范围更广的资源，比如线下的会员资料，信用卡交易数据等，都将使分析的结果更准确也更深入。

本章中讲述的主要是在社会化媒体 SNS 上的数据挖掘问题。

# 11.1 互联网上的个性化–Like

在网络上的 Web 数据挖掘不仅包括对网页内容本身的挖掘，也包括其链接模式，以及用户访问、存取、浏览、发布、操作等操作行为和访问行为所产生的信息挖掘。有效地研究、挖掘、利用网络信息可以增强网站的吸引力，有的放矢地吸引用户群，更有效地利用网络资源。做好这些挖掘工作才能使用户的个性化需求真正得到满足。

## 11.1.1　Like=像

我们来看一个在网络音乐播放器上做音乐推荐的案例。

在音乐播放器中，第一位听众选择了《北京北京》和《春天里》，那么我们可以在“随便听听”功能里推荐《怒放的生命》和《飞得更高》，因为这些都是摇滚歌手汪峰的成名曲目。

第二位听众选了《High 歌》之后，又选择播放《我的歌声里》和《小情歌》，那么下一首随便听听应该放什么歌曲呢？答案可能是《爱要坦荡荡》和《回到拉萨》。因为这些都是 2012 年《中国好声音》栏目中参选歌手的曲目。

完成前面第一位听众的推荐还是比较容易的，因为数据仓库中会有汪峰完整的音乐库列表。而要完成“像”或者“类似”的歌曲搜索，光靠数据库或今天的搜索引擎是不容易做到的，得靠数据挖掘。传统的在互联网上寻找音乐的方式是只有当你想好某一首歌的名字后，才能去搜索引擎里把它找出来。而今天，数据挖掘的各种应用使我们能够越来越多地提供类似这样的真实服务了。在 5.3.3 节中提到的哥伦比亚大学名为 Million Song（一百万首歌）的数据仓库中存放了来自世界上各种不同的歌曲，而这个数据仓库中也只是存放了与歌曲本身相关的信息。而其他很多关于歌曲的属性是分散在互联网的各个角落，需要我们去挖掘

的。即使有了 Million Song 这样的数据仓库，上述的第二位听众的要求也还是无法满足的。

要完成对第二位听众做歌曲推荐的工作，我们需要做进一步的分析。如果只是采样每个数据的各个特征做数据挖掘类似性的挖掘，我们可能无法做出合理的判断。而如果需要从互联网上找出信息来，就需要分词技术和语义分析技术。我们在互联网上用爬虫抓取网页，做了语义分析和分词之后，发现《High 歌》、《我的歌声里》和《回到拉萨》作为一个词，在一起出现在同一篇文章里，而《High 歌》、《爱要坦荡荡》和《回到拉萨》又同时出现在另一篇文章里等等，这些信息就会以某种格式存入数据存储。在积累了大量的相关歌曲数据之后，我们对于这些数据做聚类分析和关联分析之后，会发现《High 歌》、《我的歌声里》和《爱要坦荡荡》和《回到拉萨》作为歌名，是有很高的关联度的，或者说是很“像”的。那么在播放了《High 歌》、《我的歌声里》和《小情歌》之后，我们有很大的把握再播放《爱要坦荡荡》和《回到拉萨》这两首很“像”的歌会满足客户要求的。

这种找寻项目之间的关联，或者“像”的互联网服务目前越来越受到青睐，因为每个人都希望受到个性化的关注。作为用户，我希望播放器可以一直不断地推荐我喜欢听的歌曲，希望新闻网站只推送我所关注的新闻，电子商务网站一直推荐我所需要的商品。

而以此为基础的个性化搜索引擎在不久的将来也会成为主流。我们所需要的结果会直接出现在搜索结果的第一项。

### 11.1.2 Like=喜欢

互联网上最近两年很火的一个概念是“喜欢”，而有趣的是“像”和“喜欢”是同一个英文词“Like”。每个人的喜好在一段较长的时间内是不太会发生变化的，所以如果某一个新闻“像”之前你看过的多个新闻，那么有比较大的概率是你也会“喜欢”这条新闻。如果某一件商品“像”之前你买过的某件商品，那么

有比较大的概率是你也会“喜欢”这件商品。

在互联网上，如果你喜欢某一件商品，那么你可以“Like”一下这件商品。如果你喜欢某一个活动，你可以“Like”一下这个活动。当然，你也可以表示对某件商品或者活动“Dislike”（不喜欢）。所有这些点滴都会积累起来，为互联网上的推荐引擎所用。我们根据每个用户的相关属性，包括访问习惯和用户喜好，按照分类算法，把该用户归到某一类用户中去，然后通过协同推荐，借助其他人的“Like”的行为来判断该用户是否会“Like”某件商品或活动。

图 11-1 中显示的是美国流行歌手 Justin Bieber 在自己 Facebook 主页上的分享，在英文主页上有“Like”的功能，而在中文版的 Facebook 上，“赞”就是“Like”的意思。从图中我们看到有 20 万人对这张照片发表了“赞”的评论。


图 11-1　Justin Bieber 在 Facebook 主页上的 Like 示意图

图 11-2 截自 Web 2.0 公司 digg.com，图中的“Digg”也是喜欢的意思，源自英语的“dig”。Digg 是一个以互联网内容推荐分享为核心的 SNS 公司，为每个用户寻找偏好相同的其他用户，并把他们喜欢的互联网内容推荐给你。

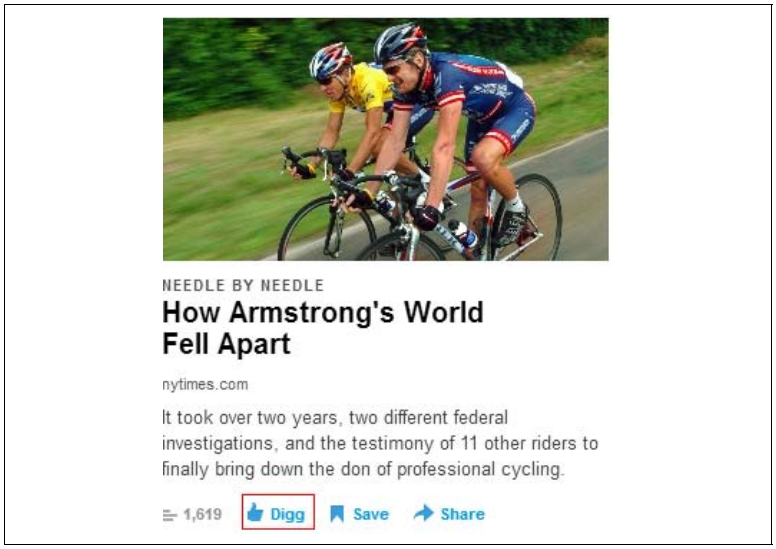



图 11-2　Digg 网页上的 Like 示意图

图中的“1,619”表示在我们看到这篇文章的时候，已经有1619 位用户对这篇文章发表过“Digg”的评论。

如果客户已经明确表示了“Like”或者“Dislike”，那么结果很清晰，说明他们是喜欢这篇文章或相关内容的。我们可以用来判定用户喜好的另一个工具是情感分析（Sentiment Analysis）或者意见挖掘（Opinion Mining），来分析用户对各种事件和产品的看法。实现的方式主要是对文本进行语义分析，简单来说，如果一个产品的描述中，相对正面的形容词比较多，那么作者对于这个产品的看法很有可能是正面的；反之，如果比较多的是相对负面的形容词，那么看法可能是负面的。我们可以通过对于这些信息的挖掘，了解大众对于事件和产品的整体看法。

这些关于用户口碑的看法又被称为舆情信息，而对于这些信息的挖掘称为“舆情监测”。我们团队曾经做过一款名为“网赢口碑”的产品。网赢口碑服务于所有需要关注企业形象和企业网络舆情信息的大中小企业和个人，为他们提供实时的网络舆情信息，如图 11-3 所示。

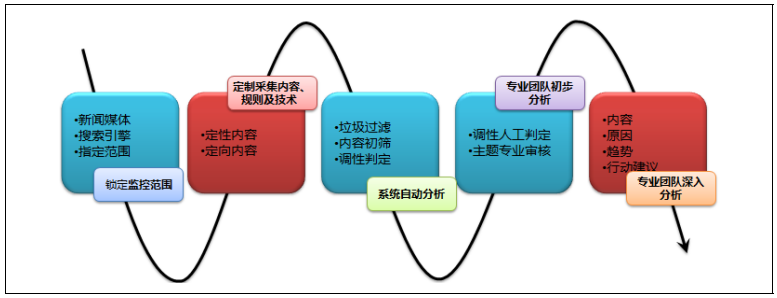



图 11-3　“网赢口碑”产品示意图

在网赢口碑中主要采用的内容挖掘有以下这几个技术。

- 元数据字段抽取：新闻语义分析，在技术上是指分词分析。
- 网页正文抽取：抽取 Html 文件中的正文字段，包含中文字数和标点符号数最多的<table>等节点和其子节点一并提取出来，就可以得到网页正文了。
- 命名实体抽取：自然语言中的统计规律，且简单、高效。规则的方法可以比较好的描述自然语言中的个性特征。命名实体识别（NE）任务是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。
- 话题分析：就是把网络上抓取到的东西根据一定的相关性分词，形成一定的话题，并对话题进行文章检索数量以及媒体转载数量的分析。
- 关键词及其变形形式的匹配算法：从大量数据中快速匹配多个关键字（多个模式）的技术。比如德意电器是一个电器的品牌，但是德意电器经常会被用户误写为“得意电器”或者“德义电器”。如果要找出与“德意电器”相关的全部内容，我们需要使用这些变形形式的关键词，才能匹配我们真正想要的全部关于德意电器的内容。
- 关键词拓展：如对于“外婆家”，我们会相应的拓展出“杭州的外婆家”及其他地点、口碑（正负面）相关词等。
- 潜在语义分析和文本倾向性分析：调性判断，了解内容的舆情信息。

我们来看网赢口碑的分析过程。

从图 11-4 中我们可以看到其实“网赢口碑”产品对于数据的处理过程和我们在本书中描述的数据挖掘过程很一致，也是先收集数据，然后经过预处理和筛选之后进入数据分析过程，最后是分析结果的可视化呈现。

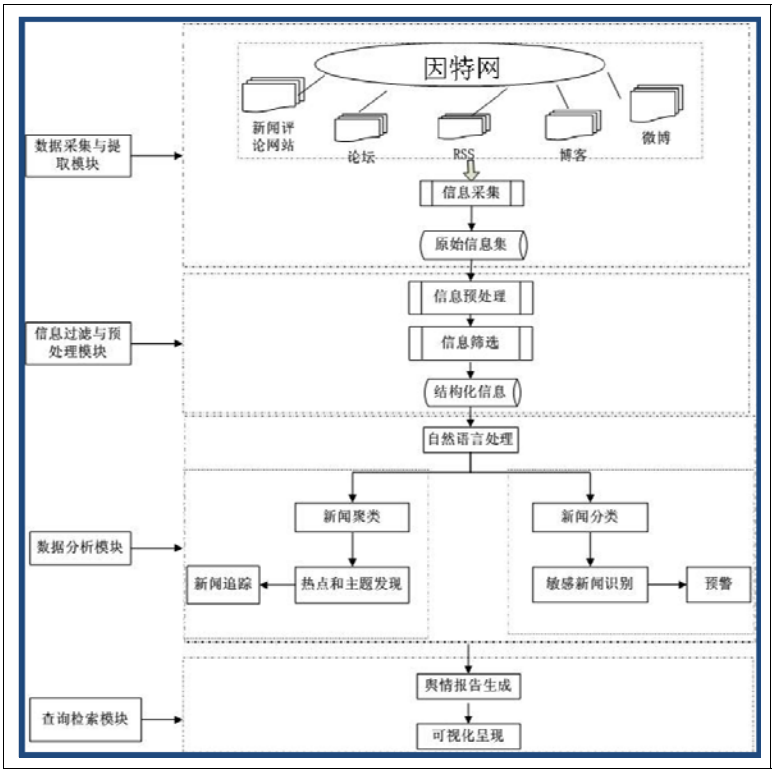



图 11-4 “网赢口碑”数据挖掘过程示意图

由于我们对互联网上信息的重视，也带给很多人新的生财之道。比如在美国，在购物网站亚马逊（Amazon）或者最大的点评网站 Yelp 上发一条五星好评，市场价格大概是在 5 美元，而在亚马逊上，有 100 个五星评论的商品要比没有评论的商品效果好很多。对于商家，花 500 美元就有了 100 个五星级评论，但是对于用户，这种假好评不但没有价值，而且还会影响他们的决策。这种假好评的发布称为垃圾意见（Opinion Spamming）。对于数

据挖掘来说，这是一件非常糟糕的事情，因为大量虚假信息的存在会造成数据挖掘的结果不准确。应运而生的是在 Web 挖掘上的一个新的研究方向：如何识别虚假评论。

## 11.2 Web 挖掘和 SNS

SNS 是社会化服务网络，Social Services Networks 的英文首字母缩写，而我们大数据挖掘中的“大”字从 2011 年起的兴旺发达，要归功于 SNS。SNS 不但是人群在互联网上的聚合，还提供了人与人之间交互的平台和人与人之间关系的集合。

每 60 秒钟，Flicker 上会有 3125 张照片上传，Facebook 上新发布 70 万条信息，YouTube 有 200 万次观赏。图片、声音、文字以及这背后用户的习惯和轨迹构成了互联网上的数据资源，大数据时代迎面袭来。

到了 2012 年，SNS 不再只是用来给人们打发时间，也不只是交友和看电影。在 SNS 上我们可以查询信息、购物、学习新知识，甚至找工作。据专做人力招聘的技术公司 Jobvite 在 2012 年 10 月做的统计，在美国有 52%找工作的人使用 Facebook 来帮助寻找，其次使用 LinkedIn，约有 38%，而三甲的最后一位是 Twitter，约占 34%。

### 11.2.1　SNS 上的数据价值

用户的消费习惯、兴趣爱好、关系网络以及整个互联网的趋势、潮流都将成为互联网从业者关注的热点，而这一切的获取和分析都离不开大数据。一方面，社会化媒体基础上的大数据挖掘和分析将会衍生很多应用；另一方面，基于数据分析的营销咨询服务也正在兴起。

“深刻洞察和理解用户需求”是每一个互联网企业生存和发展的基础，而要达到“洞察”和“理解”就离不开对海量用户进行

数据发掘与行为分析。随着社交网络和社会化媒体的掀起，“社交化”成为了当今互联网的最重要发展趋势之一。在社交时代，对于广大互联网企业来说，有效的数据挖掘和分析算法不仅可以深度分析用户属性和用户关系，并获取用户的真实反馈，从而在此基础上对产品进行针对性的优化和改进，达到真正满足用户的需求和喜好，最终提升用户的使用体验并增强其对产品的使用黏性。

数据背后潜藏着巨大的商业机会。以前只有 Google、微软这样的公司能做大数据的深度挖掘，现在已经有越来越多的创业公司进入这一领域。不同公司在不同维度的数据分析和服务正创造出新的商业模式。而这些专注于数据挖掘和数据服务的公司将成为电子商务乃至互联网第三方服务业中的新兴力量。

一项新的学术发现转化到商业模式通常会涉及很多的因素和很长的时间，比如从社交网络理论的提出到 Facebook 等社交网络兴起，经过了数十年的时间。而大数据领域的商业形态发展也会有这样的滞后性，但资本市场已经先行一步，开始聚焦于具备数据汇聚和挖掘分析能力的公司，并开始投资大数据挖掘的项目。

社交网络产生了海量用户以及实时和完整的数据，同时社交网络也记录了用户群体的情绪，通过深入挖掘这些数据来了解用户，然后将这些分析后的数据信息推给需要的品牌商家。将用户群精准细分，直接找到要找的用户正是社交内容背后的数据挖掘所带来的结果。

对于互联网公司来说，现在我们面临的挑战一方面是需要解决大数据的存储处理和访问，更重要的是让大数据产生价值，解决大数据如何为用户和广大网民服务的问题。

互联网公司一般都记录了所有用户在其网站（尤其是网络游戏和社交网络）上的所有点击、行为路径及相应的时间。如果用户尝试一个新产品，用一两秒钟就退出来了，说明这个产品可能有问题，而不是用户不想用，而其中出问题的很可能就在用户的最后一次点击发生的地方。据说 Facebook 近期采用了北美数据挖掘公司 Datalogix 的服务，分析 Facebook 上用户看的每一篇文章，听的每一首歌和欣赏的每一个视频，通过用户的点击行为和

停留时间来分析用户的喜好。

腾讯曾经就一款网游中的子弹射出后的弹道设置做研究，根据对用户行为的分析数据认为，国外游戏原本设计的逼真效果对中国用户并不合适，而用户对新设计的“比较爽快、节奏快”的弹道设计更加有兴趣。

一方面为了保证新添加用户的真实性与可靠性，而另一方面为了获取更多的关于社交平台的数据，有些新的 Web 2.0 公司已经放弃独立注册，只允许社交账户登录。在国外，互联网公司通常选择的是 Facebook 和 Twitter。而在中国这一趋势在微博兴起之后尤为明显。最经常使用的是新浪和腾讯微博。如果能够确实把握用户的真实性和唯一性，积累数据的时间越长，积累的数据越多，我们对于用户属性和喜好的把握就会越明确。如果我们能够通过一些途径把客户线上线下的行为融合在同一个平台之上，那么这些数据的价值就更加可观。

### 11.2.2　SNS 上的数据关联关系

汽车与咖啡之间存在着什么联系？或者说咖啡与奢侈品牌之间是否会有瓜葛？

答案是一切在冥冥之中皆有关联，只需要存在一个前提，即这三者的背后确定是同一个使用者。SNS 新兴公司车邻会的想法就是基于这一数据概念。作为单个用户，几乎所有的购买行为都取决于他的性格 DNA，这其中包含了他的价值取向与审美取向。而在人的性格表现当中，当做出决定的过程越慎重，性格 DNA 的关键因子就体现得越显著。对于绝大多数人而言，购车是一个大宗购物行为，整个购买决策过程足以完整凸显性格因子。事实上，很多汽车厂商之所以采用同时推广多个子品牌的运营策略，就是为了迎合不同性格的目标人群。在购买汽车时，通常情况下，用户只考虑两方面的因素：第一，价格（购买力）；第二，车辆的性格与人的性格是否匹配。成熟的汽车厂商会把子品牌之间的性格差异区分得尽可能明显，使用户购买时可以快速

定位自己的品牌归属。对照图 11-5 中奔驰公司与宝马公司的子品牌对比，我们可以发现，他们对用户的区分模式完全相同。尊崇、成熟、青春三种性格分别对应三个子品牌。

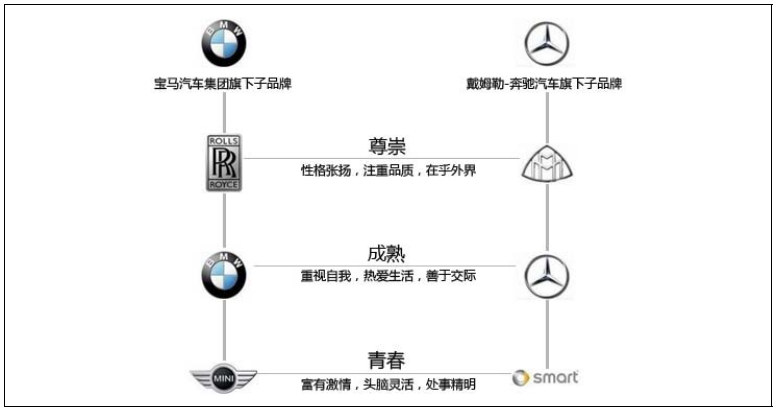



图 11-5　用户喜好分类示意图

中国有句古话叫做：“江山易改本性难移”。用户在购车时体现的性格 DNA 可以平移至其他产品的零售行为。因此，我们回到刚才的问题。汽车与咖啡的关系？宝马公司的子品牌 Mini 曾经做过一个测试，寻找驾车习惯与喜欢的咖啡口味之间是否存在联系？答案是确定的。驾驶习惯雷同的人，他们对咖啡的口味也趋于相近。

同样的，选购汽车时表现出的性格 DNA 因子，也会在选购奢侈品时得以体现。借助汽车作为媒介，咖啡与奢侈品牌之间的关系可以被搭建起来。由此可见，构筑用户的性格模型，找到其中的 DNA 因子成为关键。而这个模型的构建只有在大数据的支持下才能实现，SNS 提供了堆砌所需大数据的最佳途径。

每一款车，都提供了一个性格模型样本，基于用户自行在 SNS 上提交的大量数据，通过特定的算法分析，判断用户对几款车的喜恶程度，然后对车辆所给出的性格模型样本进行叠加，从而大致构建用户的性格 DNA 因子。同样的，用户所拥有的车辆的价格，通过某些算法可以倒推出用户的家庭购买力。当在知道一个用户的购买力和性格特征之后，剩下唯一需要做的事情就

是把适合他的产品 Push（推送）到他的眼前，从而使部分商品的精准化广告推送成为可能。

### 11.2.3　SNS 上的用户关系

谈了用户数据的挖掘，我们再来看一下 SNS 上的用户关系。

Facebook 通过 Kaggle 在 2012 年 6 月发布了一项挑战，网址是：http://www.kaggle.com/c/FacebookRecruiting/，请外界的工程师们用数据挖掘的方法找出在测试集用户中每个测试用户个体接下来可能会选择测试集中的哪些用户来跟随（Follow）。跟随（Follow）在 Twitter 等 SNS 中的意思就是关注，如果关注一个用户你会接收到她所写的内容。Twitter 的数据科学家 Edwin Chen 在他的个人博客中详细讲述了他所做的案例。他把测试集中的用户用图 11-6 的形式标识出来，图中的每一个点都代表一个用户，而每个用户周围的点都是他跟随或者跟随她的用户。这样的图称为社交示意图（Social Graph）。在 Edwin 的案例中，这张图是动态的，被鼠标点击的用户是图中最大的涂黑色实心的点，而这张点直接连接的三个深灰色的点是与用户有跟随关系的用户，或者说是 SNS 中有朋友关系的用户，再小一号的点是该用户朋友的朋友。在这张图中只显示了三层朋友关系。

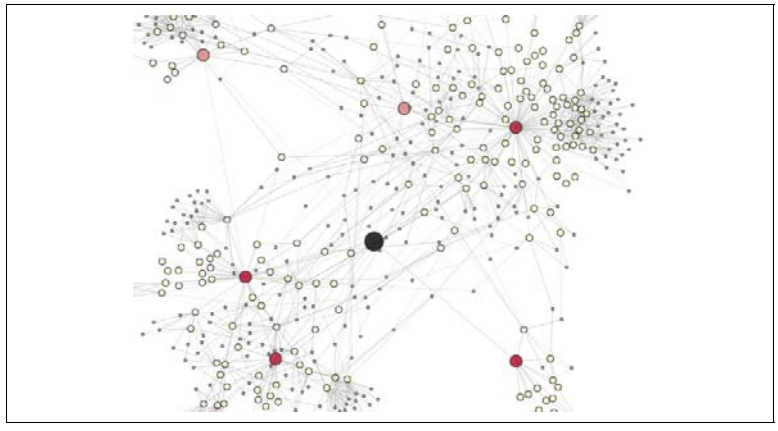


图 11-6　用户社交关系示意图

如果只选取该用户其中的一个朋友的一层朋友关系，我们可以简化图 11-6，得到图 11-7。

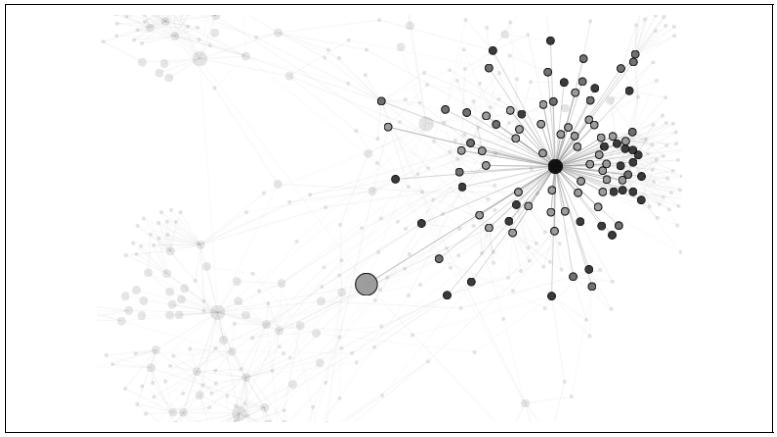


图 11-7　某一个个人的社交关系示意图

对于图 11-7 用三种不同深度的颜色来表示所选中的用户 A 和该用户 B 的不同跟随属性。一共有三种：

- 用户 A 跟随用户 B，但是 B 不跟随 A；
- 用户 B 跟随用户 A，但是 A 不跟随 B；
- 用户 A 和 B 是互相跟随的。

通过如上的方式对测试数据集中每个人的社交关系作分析，我们发现了一些有趣的现象。比如每个人的跟随者数量有以下的分布，如图 11-8 所示。

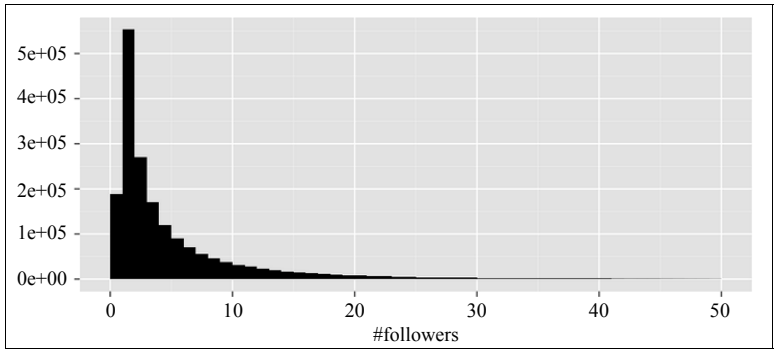



图 11-8　个人跟随者分布示意图

图 11-8 中显示的是每个人的跟随者数量分布，可以看到用户的总数按跟随者数量的上升而呈下降趋势，而绝大部分用户的跟随者在 10 以内。Edwin Chen 继而对这些朋友关系构建了数据模型，从这些已知的朋友关系中找出规律，可以帮助推荐用户好友，使推荐的成功率达到最高。

社交示意图 Social Graph 是一种用来展示和发掘 SNS 中人际关系的展示方式。我所看到做的最早最好的社交示意图是在微软亚洲研究院网络搜索与挖掘组研发的人立方关系搜索中使用的。人立方的官方网址是 http://renlifang.msra.cn/Default.aspx。人立方是一种关系搜索，从超过可以爬取到的十亿量级的中文网页中自动抽取出人名、地名等专有名词，通过算法自动计算出他们之间存在关系的可能性。此外，人立方还自动找出人名之间最可能的关系描述词和与人名最可能相关的称呼等。人立方从这些中文网页中自动地辨别出人名所对应的人物简介文字，并且按照这些文字是人物简介的可能性进行排序。

图 11-9 展示的是以 2012 年要开 32 场演唱会的杨坤为中心的社交示意图，而他周围的这些人名，包括那英、陈琳、张靓颖等都是根据从互联网中挖掘出的和杨坤有关系的人。

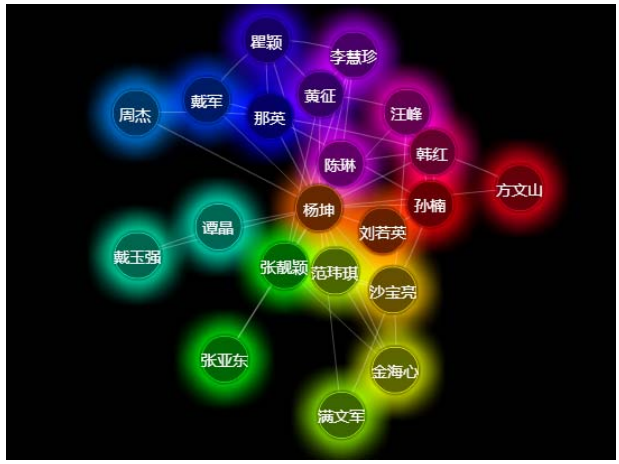



图 11-9　人立方关于杨坤的示意图

进一步查看，图 11-9 中的周杰和杨坤是“好友”的关系，而好友关系是从一个网页中分析出来的，请看人立方中关于杨坤的详细页面的截图，如图 11-10 所示。


图 11-10　人立方关于杨坤和周杰关系的示意图

图 11-10 中展示了把周杰列为杨坤好友的原因是人立方爬虫从网页中获取的信息而至。

## 11.3 数据挖掘和隐私

对个人隐私的威胁产生主要来自于当数据一旦被破译，掌握数据这方或者其他可以接近数据集的人或团体，能够辨别特定的个体，便存在利益被侵犯的可能性。

据传，从 2010 年起，MySpace 通过网络数据交易公司 InfoChimps 将用户的网站信息公开出售给第三方，包括学术研究者、市场调研机构甚至营销人员。他们出售的信息包括用户账户的任何活动内容和信息，涵盖博客日志、用户所在地信息、照片、评论和状态更新等。MySpace 在互联网领域的日薄西山有多方面的原因，但是他们对用户隐私的不重视也是其中的一个原因。

有时，信息泄露并不是互联网公司有意为之的。比如 2011 年 Facebook 的网站就曾经出现过一个安全漏洞，使用户可以通过一个链接看到其他任何在线用户的相册。

在西方国家，隐私问题是为大家所极为看重的。请看 KDNuggets 在 2012 年 7 月做的调查。结果表明，有 50.2%的网民不愿意以任何价格出售他们的 Facebook 信息，而愿意出让的人中，有 37%要以一年 500 美元或者市场价格出让，如表 11-1

所示。

表 11-1　关于 Facebook 上个人隐私调查表

| Would you be willing to sell your Facebook data to advertisers ? | |
|---|---|
| No, at any price（114） |  50.2% |
| Yes, for $10 or less/year（3） |  1.3% |
| Yes, for $50/year（5） |  2.2% |
| Yes, for $100/year（17） |  7.6% |
| Yes, for $200/year（5） |  2.2% |
| Yes, for $500/year（33） |  15% |
| Yes, for Market rate（49） |  22% |

《时代》杂志的主编 Joel Stein 在 2011 年 3 月的一篇文章是这样讲述数据挖掘和个人隐私的：“Google 认为我是一个喜欢政治、明星绯闻、动漫电影并讨厌看书的人；Yahoo 认为我是一个喜欢冰球、菜谱、服装和化妆品的中年男人”。ExeLate，RapLeaf 和 Intellidyn，三家买卖数据的公司也各自通过 Joel 的互联网信息对 Joel 做出各自的描述。而这些关于 Joel 的描述，且不论正确与否，一个商家只需要 2.5 美分就可以从这些数据公司中获得。Joel 认为这些数据的存在超出了他的底线，应当彻底在互联网上禁止通过数据挖掘来获取个人信息。

《信息管理》杂志的主编 Jim Ericson 在他的博客上对于 Google 通过 Cookie 收集用户行为的方式表示很无奈：“我们一方面感谢互联网把我们连在一起，但是我们也成为了这条食物链中的一个环节”。

社交服务网站（SNS）的发展验证了六度分隔理论（Six Degrees of Separation）的假设，也就是说在人际关系脉络方面你可以通过不超出六位中间人与世界上任意一个人认识。把朋友的朋友是朋友的原则应用到互联网世界上，线上社交网络从而得到蓬勃发展。当然，这样的情况带来的后果是个人隐私的唾手可得和不可控。

个人隐私被互联网泄露的后果在 2012 年 3 月美国一家法院

的判案中得以极端体现。一位名为 Clementi 的美国大学一年级新生被室友 Ravi 通过网络搜索确认为同性恋，Clementi 因忍受不了性倾向歧视，以及他的视频被 Ravi 在网上泄露，最终选择了跳桥自杀，Ravi 也因此被判入狱 30 天。Clementi 登录同性恋论坛的频率和网站留言等在网络上记录的痕迹被 Ravi 搜到，成为酿成悲剧的导火索，而 Ravi 做这些搜索的目的只是为了了解一下他的新室友。

这是个体有目的地通过互联网挖掘他人隐私数据（俗称“人肉”）带来的严重后果，而此类个人的事件也屡见不鲜。那么互联网公司的数据挖掘行为呢？

在 10.3.2 节中我们提到的“Target 对于怀孕妇女的营销”案在互联网上也传得沸沸扬扬。一方面大家都惊叹于 Target 能够如此精确地对用户信息进行数据挖掘，另一方面也对用户的隐私表示担忧。这些怀孕的妇女可能并不想大家知道她们怀孕的事情，而其实这些被精准营销的妇女在个人隐私方面是受到严重侵犯的。

新近很火热的移动社交应用 Path（图 11-11）为了帮用户获得和他最密切的 150 个好友，Path 未经用户批准扫描手机通讯录。Path 的创始人出面道歉，并解释说上传用户电话簿的行为是有严格限制的，仅仅用于提高“好友建议”的质量。所有用户在 Path 上分享的消息都经过了加密，这些电话簿的资料同样也是，而在 Path 后期出的版本中，用户可以选择不上传电话簿。


图 11-11　移动社交应用 Path 示意图

“虚拟社交关系”入侵“现实电话簿”的事件不仅仅出现在 Path 身上，大名鼎鼎的 Twitter 近期也承认下载存储用户通讯录长达 18 个月之久，还有 KIK、WhatsApp 以及在中国移动互联

网上很火爆的微信、米聊等一大堆的社交应用程序都会这样或明或暗地从用户手机中吸取数据。而对于这种数据挖掘行为，我们已经听到了不少的担忧甚至反对的声音：用户的个人数据和关系属于用户本人，用户拥有自己数字化数据的知情权、拥有权和绝对控制权，其他公司对这些个人数据的追踪、分析和转让都是不可容忍的。

最近另一起与数据挖掘有关的伦理讨论是由互联网巨头谷歌引起的。2012 年 1 月，谷歌宣布整合包括 Youtube、Gmail、Google+等旗下服务中搜集的用户个人信息，用户将因此有可能从根本上失去在谷歌世界里同时管理和拥有多个不同身份的能力。谷歌的这一行为已经引起了政府的注意，代表欧盟监管机构的法国计算机服务与公民自由国家委员会很快给谷歌写信，称初步调查显示谷歌新的政策不符合欧盟的数据保护指令。

这些争论多数源于对数据源中可能含有的关键信息，例如用户身份、健康状况、家庭情况、个人收入等可能泄露的担忧。另外，互联网公司也可以通过把他们分析的用户与他们所掌握的用户个人数据相结合，而对用户的网络活动进行监控。同时，数据挖掘还有可能通过把原本分散在多个网络系统中的用户数据集成、提炼，从而掌握用户在各个领域的行为，但这些行为所汇集而成的信息极有可能是用户不希望外泄的。但无法避免的是，每一次用户在互联网上注册或者使用一个基于互联网的服务时，必然会留下一些相关的个人信息，而对于这些信息的使用，各个互联网公司往往并没有经过用户的许可。虽然他们将此项授权添加在了大多数用户可能从不会仔细阅读的“用户同意事项”中。

就像保护个人隐私并不阻碍我们愿意在公共场合说话，数据挖掘也不总是站在隐私的对立面。其实我们经常会在公开场合，例如餐馆、车站或地铁上，进行私人性质的谈话。我们明白对话内容会被服务员或者路人听到，但心理学家欧文高夫曼所说的“礼貌性疏忽（Civil Inattention）”会帮到我们。人们一般会选择过滤掉我们的谈话内容，即使听到了也不会加入我们的讨论。当

然我们自己也会通过压低声音来限制传播范围，谈话末了还会加上一句，“千万不要告诉别人”。

在网络这个虚拟公共场合里，用户也可以对一些信息内容采取加设密码、“穿上马甲”等方式来进行“窃窃私语”，效果就像在餐馆谈话时压低声音一样明显。不过用户的大多数数据还是会被记录，然后被使用于商业目的。第 8 章的邮件营销，第 9 章的网盟广告和第 10 章的电子商务其实都运作在用户数据挖掘的基础之上的。而现在流行的精准营销、社会化营销、移动广告等无不是在数据挖掘的支持下产生的。

包括谷歌在内的一些互联网公司认为，这些适度的隐私出让可以让用户受益，并带来社会效率的整体提升。如果没有得到用户足够多的数据并进行分析，Google 的搜索结果可能会像谢耳朵的超严密逻辑分析一样无厘头。某种程度上来说，正是这些用户数据的有心搜集，才让互联网提供的各种服务多了些温情，少了些死板，“您要让它为您更好地服务就不可能不让它更了解您是谁”。在搜索领域，对谷歌搜索威胁最大的将是 Facebook，而其原因正是因为 Facebook 掌握了大量的个人信息和历史数据，使搜索结果更加精准。

其实，数据挖掘本身不存在伦理问题。数据挖掘技术是中立的，大多数用户行为的数据也不会产生伦理问题，从广义上讲，从高速公路上的车流数据，商业用车的碰撞测试数据到股票的历史数据均可被视为数据挖掘的范围，这些类型的数据虽占据可以被数据挖掘方法所分析的很大比例，却很少让人产生道德方面的忧虑。

对个人隐私的威胁的产生主要来自于当数据一旦被破译，导致数据挖掘方或者任何可以接近数据集的人，能够辨别特定的个体，便存在利益侵犯的可能性。例如保险公司可以透过访问医疗记录来筛选出那些有糖尿病或者严重心脏病的人，不允许他们加入保险计划或是大幅提高他们的保费，从而削减保险支出。

面对互联网公司数据挖掘的隐秘性，欧盟正在起草新的数据

保护条例，包括将有可能对违规公司处以其全球收入 2%的罚款。2012 年 2 月，奥巴马政府也定下了新隐私保护规范的架构，此规范能让消费者更好地控制个人信息的使用。

## 11.4 本章相关资源

- 本章相关参考文献：

[1] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Verlag, 2007.

[2] M Hu, B Liu. *Mining and summarizing customer reviews*, In Proceedings of the tenth ACM SIGKDD international Conference, 168-177，2004.

- 本章相关网址：

[1] http://www.kdnuggets.com

[2] http://eguan.cn

[3] http://www.iresearch.cn/

[4] http://blog.echen.me/

[5] http://renlifang.msra.cn/Default.aspx

# 第 12 章
# 数据挖掘和移动互联网

和传统互联网相比，移动互联网有它的特殊性。在本章中我们主要讲述在移动互联网范畴上的数据挖掘问题。和时间信息相对应的地理位置信息是传统互联网所独有的，而 LBS 是具有移动互联网特点的典型应用。在 12.2 节中我们以两个实际案例来介绍 LBS 上的数据挖掘。

对于移动互联网上基于数据的应用，分别从个体和宏观上来看主要有两大类：

- 从个体上来讲，我们可以有针对性地给不同用户推荐他们所需要的产品和服务。
- 从宏观上来看，我们可以通过移动互联网上的数据挖掘，把用户归类到不同的社会群体中，预测这些人群在什么时间出现在哪些地点，预测交通的高峰期以及道路变化对交通的影响等。

作为本书的最后一章，我们在此讲述的移动互联网上的数据挖掘是最新也是最有发展潜力的。笔者一直在关注移动互联网上关于数据的应用和发展。如果读者对这一领域有兴趣，欢迎和我交流。

## 12.1 移动互联网的特殊性

随着各种移动设备、物联网和云存储等技术的发展，人和物的所有轨迹都可以被记录。与互联网不同的是，在移动互联网中

的核心网络节点是人，不再是网页。在数据大爆炸的当下，怎样挖掘这些数据，同样面临着技术与商业的双重挑战。

对于数据挖掘来说，移动互联网的特殊性首先在于它能够锁定一个特定用户，其次在于它能够获取用户地理位置信息，再次是在于移动互联网上的时空信息等多样化的数据种类。而因为这三点，导致移动互联网上的数据数量会比传统互联网更大，形式也比传统互联网更加丰富，从而也有更高的价值。

### 12.1.1　锁定用户的数据价值

归根到底，在移动互联网上的挖掘，或者说 Mobile Mining 的目的和做传统互联网数据挖掘的目的是一样的：都是为了从原始数据上找出有用的信息，进而转化成可用的知识。但是移动互联网有其特殊性，即移动互联网的某一个终端通常是由同一个个体使用的，所以用户在移动终端上的所有行为是具有一定延续性的。试想，如果能直接获取用户移动终端上的信息，那么我们就可以得到包括手机号码和通讯录等在内的一系列的用户个人信息。退一步讲，即使我们没有这些信息，移动终端访问的延续性也使得用户档案（Profile）的建立成为可能。

在第 7 章中我们讲述了互联网企业如何可以通过日志分析来分析访客信息，不过如何判断不同的登录来自同一个访客是一个至今无法全面解决的问题。我们识别一个用户依靠的是用户标识，其次是 Cookie，再次是 IP 地址。

- 首先用户主动去注册一个网站是比较难的，通常用户只会在他经常访问的网站上注册，比如微博、淘宝或者他特别关注的某一个社区等，在其他的网站上可能只是浏览，而不会专门注册和登录。
- 其次各家浏览器公司和安全产品公司都会定期自动或者提醒用户删除 Cookie，给我们在互联网上识别客户造成很大困扰。
- 用户上网环境也会经常发生变化，比如在网吧中是无法

识别该用户的 IP 的。当然，我们也可以通过大量的数据分析出该 IP 在 80%的概率上是一个网吧或者公共上网点，但是除了排除这个 IP，我们并不能够识别出谁是我们的客户。

所以在传统互联网上，在没有 Cookie 的情况下，如果用户不主动登录某个网站，或者是一个偶尔的访客，那么我们很多时候无法判定用户的唯一性，不能确定该用户之前是否访问过该网站，之前的访问行为完全没有延续性。

不过在移动互联网上，这个问题就清楚很多了。我们可以锁定用户，即使因为隐私和用户规则等原因，我们不主动获取用户的个人信息资料，但至少可以知道该用户是否和之前的某个访客是同一个个体。我们能够通过移动互联网应用获取用户当前的位置信息和参加活动的一些信息，并把这些信息记录收集下来，从而积累成关于某个用户的丰富档案信息。

这些关于给定个体的信息积累是一笔很大的财富。从一些位置信息中我们可以分析出用户的大概活动范围，从位置和频率可以发现用户在当前位置是出差还是常住的，甚至可以知道用户在此刻可能需要什么。在一定的概率下，我们可以把用户归类为本地住户、外地暂住的用户或是学生等。

### 12.1.2 移动互联网上数据的形式

移动互联网给我们带来的数据在形式上是多种多样的，在互联网上存在的各种数据类型只是移动互联网上的一个子集。

自苹果公司的 iPhone 和 iPad 问世之后，各种移动设备能够带给使用者的用户体验与互联网几乎是一样的——几乎所有在互联网上存在的信息在移动互联网上也都存在，不过内容可能更加丰满。在苹果公司的这两款产品之前，由于手机和移动端的屏幕限制，在访问网站的浏览器设置中，会以不同的方式来显示网页。iPhone 引领的移动端用户体验使得用户几乎可以以同样的感觉来访问网站。

同样的互联网信息，在移动端访问的方式也会使内容变得更加丰满。同样都是图片，但是在移动设备中存在的与位置相关的图片要比简单存在于互联网上的图片价值高很多，或者说同样都是一句评论，但是有场景的评论和没有场景的评论相比前者更有挖掘的价值。

据 Facebook 上专注于数据分析的 Pagelever 公司报道，如图 12-1 所示，在 2012 年 8 月，Facebook 的新增用户中，接近 20%来自于手机端，而这一数字在三个月前的 2012 年 5 月仅仅是 5%。更多详细信息可以在 Pagelever 公司的官方博客上看到：http://pagelever.com/blog/。

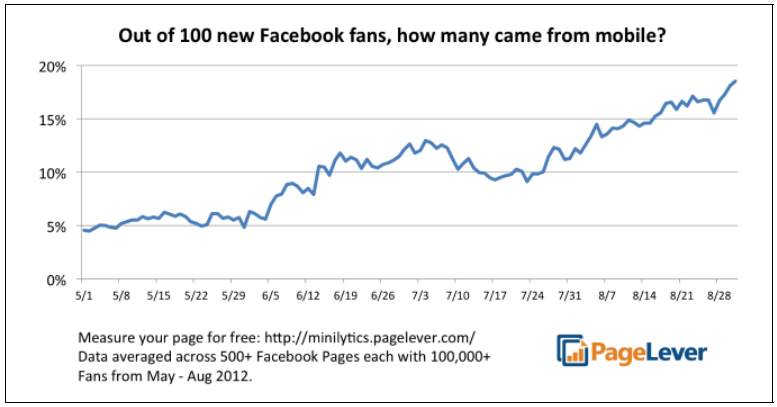



图 12-1　20%的 Facebook 用户来自移动端

图 12-1 中的增长幅度是非常惊人的，在短短三个月中，来自移动端的新增客户比例涨了大约三倍，现在每 5 个新加入 Facebook 的用户，就有一个是来自于移动端的。Facebook 在 2012 年 6 月公布的官方数字显示，在每个月 Facebook 的 9.55 亿活跃用户中，5.43 亿是移动用户，占比达到 56.9%，而且 5.43 亿移动用户中有 20%的用户是纯粹只用移动客户端访问 Facebook。

与传统互联网的数据不同的是，在移动互联网的数据中，文字以外的其他信息占到更加重要的比例。从数据的属性上来讲，移动互联网上的数据比传统互联网更加复杂，其中一个原因是这些数据包含了大量的时间和空间的信息，也就是说我们需要把数

据挖掘延伸到时空数据挖掘的领域（Spatio-temporal Data Mining）。因为多了一个维度，时空数据挖掘的复杂度比一般的数据挖掘又深了一层，虽然说研究方法和算法还是类似的。

自 2010 年开始，Facebook 在网站上添加了用户位置分享的功能，又一次丰富了可以给朋友分享的内容，除了文字、图片和视频之外，又增加了一块——地理位置。在移动端上，地理位置分享很容易做到，只需要手指轻轻标记一下即可。你可以在 Facebook 上很轻松地看到最近都有哪些朋友去各地旅游或者工作，了解到有哪些朋友曾经去过你目前所在的旅游点或者陌生城市以及他们对当地的一些风土人情的了解。

2012 年苹果公司推出的 iPhone5 手机中，第一次采用了自己开发的，可以说并不太成熟的地图服务来取代 Google 的地图服务。这里有很多原因，作者认为最重要的原因其实是在数据上。因为地图应用是收集用户兴趣点和用户地理位置最好的移动互联网应用，苹果公司不甘心把这个最好的入口掌握在竞争对手手中。

### 12.1.3 移动互联网地理位置信息的价值

举例来说，我们通过移动互联网应用可以记录某个用户在过去的一年内的根据时间的地理位置变化。我们可能发现，该用户的移动终端 85%的时间在上海市的一些不同位置登录，10%的时间是在北京，而其他的 5%的时间是在深圳、广州、杭州等各个地方。那么我们可以大概判断这个用户是常住在上海，而由于他的工作或者家庭原因使得他必须一年有一段时间在北京居住，所以以北京为其暂住地，而在一年内的其他时候他可能是去旅游或者工作。又如果他在北京的时间是十一或者春节等节假日期间，那么我们有更高的把握判断他去北京是由于家庭原因。另外，如果数据表明他每次出现在北京之前的 10 个小时内，在上海到北京沿线的不同地点登录，那么我们可以大致判断他每次来回北京上海之间都是坐高铁出行。

有了这些信息，我们可以推荐针对这一人群的服务，比如可以在节假日之前提供北京往返机票和优惠礼品券，在平时提供商务人员需要的个性化产品等。

如果我们拥有的用户基数足够大，就可以用移动互联网的位置信息来做很多更深层次分析。在中国，拥有最多数据的是国内的三大运营商——电信、联通和移动。每个用户进入和退出每一个手机基站都会有相应的数据产生。如果我们需要精确地做旅游景点的人数估计和旅游热点迁移的估测，只要把全国每个旅游景点周围基站每天的数据汇总起来，在之上做统计和数据分析，就可以轻易地回答以下问题：

（1）今年“十一”长假全国最热的一百家景点是哪些？

（2）今年 10 月份大概有多少人次到西湖旅游？

（3）和去年的“十一”相比较，哪些景点的旅游人数是增长最快的？

（4）今年“十一”到乌镇旅游的人，来自各个省份的比例是怎样的？

腾讯 QQ 和微博客户端也有足够的用户基数来回答以上的问题，不过用户基数再小一些的应用做的答案可能误差就会比较大。

对于单个或者多个用户，我们可以构筑热图或热力图（Heat Map）来直观体现用户的位置变化和迁移情况。我们在第 6 章曾经使用过热力图，一个简单的热图可提供的信息概况是即时可见的。不过由于移动互联网的特点在于我们还知道该用户是在什么时间点到达所在的地理位置的，所以使用动态的热力图更加能够表现出位置变化的特点。知道了这些信息，对于移动互联网的应用程序编写或是广告提供商是很有价值的。

有了移动互联网的数据，我们可以真正实现用户的行为定向，通过用户使用各种应用的习惯与场景，还原用户属性，了解用户兴趣和喜好，预测用户消费习惯和消费意图，实现真正的精准定向。

## 12.2 数据挖掘和 LBS

LBS（Location-Based Service）是与位置相关的软件服务的英文缩写，指的是一类利用和控制与位置及时间相关的计算机软件服务。LBS 通常是在移动终端实现的。现在很多原本只是在互联网上的应用都有了 LBS 服务。

与位置相关的数据挖掘是一个看似简单，却非常具有挑战力的工作。举个简单例子，帮用户寻找他所在地附近可能有用的商业地点，并按照一定的规则排序。这句话说起来容易，实现起来却是要费一番功夫。

任何跟与位置相关的数据挖掘的工作必不可少的第一步就是搜集关于地点的可靠数据。在这个过程中，常常会面对多个不同的数据源，有些来自互联网，而有些来自于线下，所以第一步面临的常常就是数据的整合与清理。与位置相关的数据的数据量经常是在 GB 字节上下，对于这个量级的数据频繁的整理、提取、集成和存储都有着一定的难度，但恰恰可以利用我们之前提到的一些框架和应用工具，比如我们在 3.5 节中讲述的 Hadoop 和 Hive 等。Hadoop 有着强大而又灵活的构架，不仅仅提供了容错的分布式存储功能，也支持了各种灵活的数据挖掘工具。而 HBase 是建立在 Hadoop 之上的文件存储系统，它对存储的数据进行索引从而达到对数据的迅速挖掘、查询和更新。以 HBase 为基础的 Hive 所提供的 HQL 查询工具进一步帮助它的用户达到可以快速挖掘数据的目的。这样，我们可以说在 Hadoop 构架上，从最简捷的 Unix Shell Script 到 Pig，到程序员常用的 Java，到 Hive 所提供的 HQL，有着如此强大的挖掘工具的支撑，任何烦琐的数据整理和最终的存储与挖掘工作都变得相对明了简单。我们可以用 5.3 节中展示的方法来对整合的数据做数据清理和转换。

各个地点之间的关联性是需要通过数据挖掘才能完成的任务。每个地点都有多种属性，而地点之间的关联度是根据他们各自的属性匹配所得到的。

不过目前很多 LBS 应用对于关联性的挖掘还是比较简单的。有的只是根据地点的归类、名字、用户设置的标记（Tag）做简单的对比。比如 LBS 中比较有名的公司 Foursquare，网址是 https://foursquare.com/。如图 12-2 所示，Foursquare 提供的 Similar Places 就没有引用系统化的相关性分析，而是主要根据网站上的人气选择来确定分类。根据人气选择来确定分类的方式造成的结果是有大量长尾的地点只有很少或者没有人关注。在 2011 年火过一阵的切客（Check-in）其实就存在这个问题。所谓切客，是用户在她到达的地点上做“到此一游”的标注，而 Check-in 的中文翻译其实就是“到此一游”的意思。光是这样的标识，除了了解地点的火热程度之外，其数据没有太多别的意义。

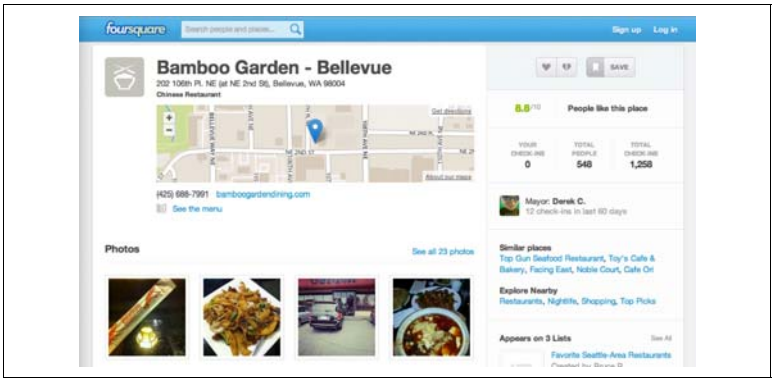



图 12-2　Foursqaure 界面图

## 12.2.1　用 PU 学习算法做文本挖掘

这个方向国内也有类似，甚至可能更前瞻的基于社交化的数据挖掘，提炼出个性化推荐。火花无线，一家 O2O 的无线新秀所推出的美食推荐应用——麻花，就是一个典型案例（图 12-3）。作为一部分基础数据，该应用挖掘了新浪微博上有关餐厅的微博

分享，并汇集成热门餐厅。我们来看一下麻花是怎样在新浪微博等 SNS 上做数据挖掘的。


图 12-3　麻花界面示意图

为了给用户最个性化和最高价值的推荐，该应用进一步通过互粉关系，把互粉用户所推荐的内容提高权重，推荐给用户，从而提高推荐餐馆的相关度和增加搜索结果的可信度，如图 12-4 所示。


图 12-4　麻花用户微博推荐示意图

说起来很简单，但是在这里比较关键的是如何从用户的某条微博中发现地点和判断用户是对该地点做出的评论。不是每条包含地点的微博都是对地点的推荐。比如：

“我在贝塔咖啡吃午饭”

或者

“今天去福地听讲座”

这些微博虽然提到了地点，但只能算是一个“Check-in”，并不是对这些地点的评价。

“在整个个性化推荐过程中，建模是个关键环节”，先前在淘宝搜索任职，而现任火花无线的高级研究员王嵬说到，“建模的目的是将现实的事物映射成计算机能理解的事物，而数据挖掘的相关算法就是进行有效映射的思路、方法和工具。通常我们建立的模型对现实事物的映射的失真越小，那么得到的结果会越好，反之越差。”

以麻花美食的评论过滤系统为例。我们需要通过建模在所有爬取的微博内容中找出对餐厅有评论性质的微博。这里的数据模型使用了 PU 学习文本分类算法。

正例和无标记样本学习（Learning from Positive and Unlabeled examples）一般称为 LPU 或 PU 学习。PU 学习是一种常用的半监督的二元分类模型，它的目的是通过已标注的正例数据和大量的未标注数据训练出一个用于区分正反分类的分类器。

PU 学习的实现方式分为两步：

- 第一步是通过标注好的正例数据在未标注数据里找出反例数据；
- 第二步是通过标注的正例数据和找出的反例数据建立一个二元分类器。

麻花美食最开始得到的分类器是非常糟糕的，查找到的原因有以下几类：

- 在寻找到的反例数据与现实的失真过大，比如正例数据在未标注数据里的比例仅为 20%，而反例数据占 80%，

结果通过第一步找出的反例数据跟已标注的正例数据的比例与现实的比例相差悬殊；

- 找出的反例数据覆盖的并不全面，出现了偏科问题；
- 在 PU 学习的第二步最初采用了一种不能容忍此类失真的算法，结果导致最后训练出的分类器达不到要求。其实这就是在模型建立的过程中某个环节对现实事物的映射出现了严重失真从而导致整个模型对现实事物的映射严重失真。

最后在麻花美食的评论过滤系统里通过对以上三类失真的修复之后让分类器的 $F_1$ 值达到了 82.14%。

这里的 $F_1$ 值是在 Web 信息检索中常用的评价值。Web 信息检索中最重要的两个概念是准确率和召回率。准确率（Precision Rate）是检索出的相关文档中正确的数字和检索出的总相关文档数字的比率；召回率（Recall Rate，也称查全率）是检索出的相关文档数和文档库中所有相关文档数的比率，而 $F_1$ 值是准确率和召回率的调和平均值。下面是关于 $F_\beta$ 的通用公式：

$$F_\beta = (1+\beta^2)\cdot\frac{precision\cdot recall}{\beta^2\cdot precision + recall}$$

当 $\beta$=1 时，所表示的就是 $F_1$ 的公式：

$$F_1 = 2\cdot\frac{precision\cdot recall}{precision + recall}$$

### 12.2.2 用相似匹配算法做地点挖掘

另一个案例是西雅图的一家初创公司做了一个名为 aLike 的 LBS 产品，在 2012 年被评为最优秀的移动互联网应用之一。他们在数据仓库中整合了数百万个有地理位置信息的地点，以地点的空间属性为索引的方式组织起来，并且整理出关于这些地点的多至数百个的其他属性。后台的爬虫程序（Web Spider）不断爬取互联网上的地点，和数据仓库中的地点相匹配。aLike 最有特点的地方就是除了按地理位置空间索引浏览，还可以让用户查看与某个地点相关度最高的其他一些地点。这就结合了在搜索中的

按相关性做的排序和在普通地图应用中按地理位置做的排序。在aLike 应用中，地点之间的相关性是通过每个地点几百个属性构成的空间向量计算得来的。

和普通数据挖掘过程一样，在 aLike 的案例中，预处理是整个数据挖掘过程中最费时的一个步骤，要占据整个过程 60%以上的时间。根据数据挖掘变化定律（Law of Change），数据是一直在变化的，而关于某个地点的信息如何确定和维持它的数据可靠性是有挑战性的工作。特别是同一个位置，往往在不同的数据源上有不同的表示，如何判定网上几个不同的地点表示对应的是同一个地点也是耗费不少工作量的。

在完成了基础的数据预处理和整理工作之后，需要采用聚类算法把数据点聚成不同的类别，而这里比较关键的就是采用什么样的相异度比较公式来计算两个不同数据点之间的差异：

$$d(X,Y) = f(X,Y) \to R$$

基于商业秘密的考虑，这个比较公式以及具体采用的聚类算法我们在此就不说了。在决定了公式之后，确定一个阈值，那么在每个聚类中的各个数据点之间的距离都是小于阈值的。

当我们有了经过整理的与位置相关的数据之后，达成查询数据中与位置相关的信息就只是程序化的工作了。由于每个数据都是与其地理位置相关，我们最终需要把这样的数据根据地理位置，商业名称，地址等索引起来，然后根据用户的地理位置提供他附近的各种商业服务信息。

在 aLike 应用程序中，用户采用浏览方式或者从历史数据中选择了 Dimitriou’s Jazz Alley，一家在美国西雅图的爵士酒吧。我们从 Dimitriou’s Jazz Alley 出发，可以寻找和它相关度较高的地点。如果用户想要在附近区域查找和 Dimitriou’s Jazz Alley 相似的一些地点，那么很遗憾，如图 12-5 所示，相似度最高的也只有 57%，而排在第二位和第三位的位置相似度分别只有 50%和49%。不过如果用户在 aLike 应用程序中查询相似地点的工作中把搜索的地域范围扩大，那么他们可以看见在美国东北部的纽约

有一家类似的地点，其相似度可以达到 70%。

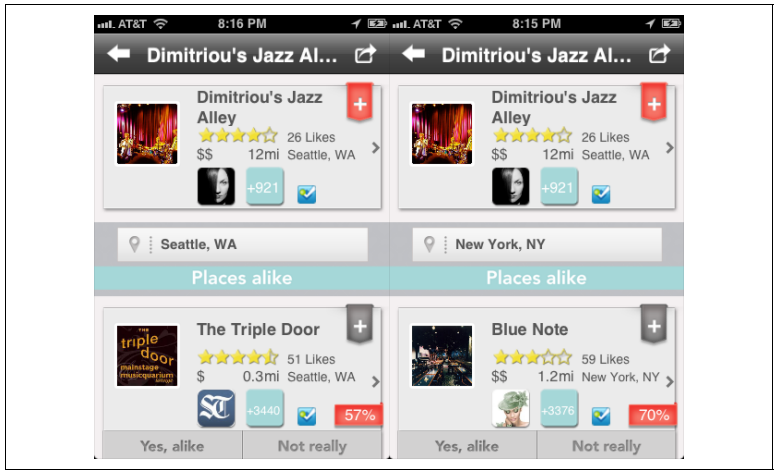



图 12-5　aLike 用户界面图

在 aLike 之后的版本中，在有了大量用户或者从其他来源比如 Facebook 和 Twitter 上获取足够多相关数据之后，可以通过数据挖掘的聚类算法找出相同兴趣的用户，并给他们推荐同类用户所喜欢的地点。比如，当一个用户在查看附近的地点，但是没有给出任何提示时，我们会根据聚类，综合了地理位置由近及远的排列，给出他可能喜欢的地点。

LBS 应用最有价值的地方在于我们藉此能够对用户做精准的地域定向。这样的广告价值相对要高很多。比如一个餐馆可以选择对它周围 1 公里的用户发送折扣券，一个搬家公司可以选择对它周围 10 公里的用户发送广告，等等。

## 12.3 移动互联网数据面临的问题

移动互联网有它的特殊性，而移动互联网上的数据除了它的特殊价值之外，也有和传统互联网不完全相同的问题。

- 数据量

移动互联网所可能产生的数据量是一个我们需要考虑的问题。据统计，在中国，2012 年约有不到 6 亿移动互联网用户，其中有约 1.8 亿是手机应用商店的使用者，而且这个数字正在飞速增长之中。如果每个用户产生的所有数据，包括即时的位置信息、路径信息、访问信息等都需要实时分析，那么需要处理的工作量是目前任何一家互联网公司都望尘莫及的。或许发展中的云计算可能在不久的将来能够做到这一点。而做到这一点的时候就应该是云计算真正能够实现盈利的时候。

- 安全性

而移动互联网上的安全因素是另一个我们需要考虑和解决的问题。因为在移动互联网上有很多恶意的应用程序，而他们的目的就是侵入你的移动设备来窃取个人信息。另外，因为移动终端和个人身份信息密切相关，在移动互联网我们更加要重视在 11.3 节中提到的个人隐私问题。

- 数据质量

移动互联网上的数据价值是大家都很看好的，但是数据质量却令人担忧。移动互联网行业结构目前并不明朗，盈利模式也不清晰。大量的移动应用通过刷量来冲击移动互联网应用排行榜以追求投资人的青睐。大量移动互联网公司付费给水军来给自己的移动应用发五星好评，给竞争对手的应用打一星差评。这些数据所占据的比例过高，已经严重干扰了数据的准确性，而这些行为实际上大大降低了移动互联网数据的整体价值。这样看来如果我们想从移动应用中完全挖掘出真实的用户行为，可能还需要一段时间。

当然，这一切都在往良性的方向发展。2009 年，苹果公司判定 Molinker 涉嫌发布对他们的低质产品进行虚假的正面评价，把他们公司全部的 1011 个应用程序下架。对于着重在移动互联网上开发应用程序的公司来说，这是一个很大的警示作用。

正如我们在互联网刚刚起步的时候所看到的各种问题一样，在移动互联网的萌芽状态呈现出的各类情况和比较难解决的问

题也是相当正常的。在不久的将来，移动互联网上一定会有专注于数据服务的公司出现，他们不但能够解决大数据量的实时处理问题，同时还能够提供安全可靠的服务。我们拭目以待。

## 12.4 本章相关资源

- 本章相关参考文献：

[1] Goh, Jen and Taniar, David. *An Ef ficient Mo bile Da ta Mining Mo del*：*Parallel and D istributed P rocessing and Applications*. Springer Berlin, 2005.

[2] Nafiseh Shabib，John Krogstie. *The use of d ata mining techniques i n loc ation-based r ecommender syst em*. in Proceeding WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics，2011.

- 本章相关网址：

[1] https://foursquare.com/

[2] http://pagelever.com/blog/

# 附录 A
# 技术词汇表

Anomaly：见“异常值”词条。

ARPU（Average revenue per user）：每个用户的平均收入。

Avro：一个在 Hadoop 上的数据序列化系统，设计用于支持大批量数据的交换应用。

宝贝：淘宝和天猫网上商城对于网店商品的专门用语。

贝叶斯分析方法（Bayesian Analysis）：提供了一种计算假设概率的方法，这种方法是基于假设的先验概率、给定假设下观察到不同数据的概率以及观察到的数据本身而得出的。

Bounce Rate：见“跳出率”词条。

B2C：英文 Business to Consumer 的缩写，其中文含义为企业对消费者。

CART：Classification and Regression Trees 的英文首字母缩写，或者称分类与回归树，是一种决策树分类算法。

CBL（China Black List）：中国垃圾邮件黑名单。

Cluster（类或簇的英文）：是一个数据对象的集合。

Cookie：指的是指网站为了辨别用户身份而储存在用户本地终端浏览器上的一类数据。

CRM（用户关系管理，Customer Relationship Management）：指的是公司对客户和潜在客户的管理模式。

Direct Marketing：见“直效行销”词条。

Discriminant analysis：见“判别分析”词条。

DSS（Decision Support System）：决策支持系统的缩写，是辅助决策者通过数据、模型和知识，进行半结构化或非结构化决策的计算机应用系统。

独立访客：指在一天之内（00:00-24:00）访问网站的上网电脑数量（以 Cookie 为依据）。

EB：计算机存储单位，1 EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes（字节），或是 2 的 60 次方字节。

EDM（Email Direct Marketing）：电子邮件营销。

Entropy：见“熵”词条。

二跳率：当网站页面展开后，用户在页面上产生的首次点击被称为“二跳”，二跳的次数即为“二跳量”，而二跳量与浏览量的比值称为页面的二跳率。

ETL：Extract Transform Load 的缩写，是指数据的提取、转换、加载。

分布式数据库（Distributed Database）：用计算机网络将物理上分散的多个数据库单元连接起来组成一个逻辑统一的数据库。

关联规则（Association rules）：是形如 X→Y 的蕴涵式，其中 X 和 Y 分别称为关联规则的先导（antecedent 或 left-hand-side, LHS）和后继（Consequent 或 Right-Hand-Side，RHS）。

根节点：决策树最上面的节点。在它上面没有其他节点，其他所有的属性都是它的后续节点。

购物篮分析（Market Basket Analysis）：就是关联规则算法。在市场上关联规则算法经常作为商品购物车的分析，所以在应

用领域又被称为购物篮分析。

Granularity：见“粒度”词条。

HBase：一个在 HDFS 上搭建的大规模结构化存储集群分布式存储系统，具有高可靠性、高性能、面向列、可伸缩特性。

HDFS：部署在廉价硬件上提供高吞吐量和高容错性的分布式文件系统，适合有超大数据集的应用程序。

Hive：基于 Hadoop 的数据仓库工具，可以将结构化的数据映射成数据表并提供类 SQL 数据库查询管理功能，适合于数据仓库的统计分析。

后验概率（Posterior Probability）：当根据经验及有关材料推测出主观概率后，对其是否准确没有充分把握时，可采用概率论中的贝叶斯公式进行修正，修正前的概率称为先验概率，修正后的概率称为后验概率。

回归分析（Regression Analysis）：是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法。

计量经济学（Econometrics）：是以经济学和数理统计学为方法论作为基础，对于经济问题试图用数量和经验两者进行综合的经济学分支。

基于互联网的挖掘（Web 挖掘）：是利用数据挖掘技术从 Web 文档及 Web 服务中自动发现并提取人们感兴趣的信息。

交叉验证（Cross-validation）：主要用于建模应用中，在给定的建模样本中，拿出大部分样本建模型，留小部分样本用刚建立的模型预报，并求这小部分样本的预报误差，记录它们的平方加和。

机器学习（Machine Learning）：研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

聚类（Clustering）：将物理或抽象对象的集合分成由类似的对象组成的多个类的过程。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。

决策树（Decision Tree）：一般都是自上而下生成的。每个决策或事件（即自然状态）都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干，故称决策树。

决策树剪枝（Decision Tree Pruning）：由于在决策树生成过程中，会过度拟合训练数据，而且易受噪声数据的影响，所以剪枝操作是决策树生成过程中的一个重要步骤。

决策支持系统（Decision Support System）：辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统。

KDD（Knowledge Discovery In Database）：泛指所有从源数据中发掘模式或联系的方法。

K 近邻（K Nearest）：一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 K 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。

LAMP：Linux、Apache、MySQL 和 PHP，四种 Web 技术的缩写，是一些 Web 2.0 公司使用的主要技术组合。

landing page：见“着陆页”词条。

LBS（Location-Based Service）：是与位置相关的软件服务的英文缩写，指的是一类利用和控制与位置及时间相关的计算机软件服务。

粒度（Granularity）：指数据仓库的数据单位中保存数据的细化

或综合程度的级别。

Lift：使用分类器相对于不使用分类器产生的正类的比例。

联机事务处理系统（OLTP）：实时采集处理与事务相连的数据以及共享数据库和其他文件的地位的变化。在联机事务处理中，事务是被立即执行的，这与批处理相反，一批事务被存储一段时间，然后再被执行。

联机分析处理（OLAP）：使分析人员、管理人员或执行人员能够从多角度对信息进行快速一致，交互地存取，从而获得对数据的更深入了解的一类软件技术。

流量（Traffic）：是指网站的访问量，是用来描述访问一个网站或是网店的用户数量以及用户所浏览的网页数量等一系列指标，这些指标主要包括：独立访客数量（Unique Visitors）、页面浏览数（Page Views）、每个访客的页面浏览数（Page Views per user）。

六度分隔理论（Six Degrees of Separation）：是个假设，在人际关系脉络方面你可以通过不超出六位中间人直接与世上任意人认识。

Metadata：见“元数据”词条。

MapReduce：HDFS 上处理大数据集的并行计算框架。

MongoDB：是一个基于分布式文件存储的数据库。

Nginx：开源的高性能 HTTP 服务器。

Outlier：见“异常点”词条。

PAM：见“围绕中心点的划分聚类算法”词条。

判别分析（Discriminant analysis）：是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。

PB：计算机存储单位，1 PB = 1,024 TB = 1,048,576 GB = 1，125，899，906，842，624 Bytes（字节），或是 2 的 50 次方字节。

PU 学习：正例和无标记样本学习（Learning from Positive and Unlabeled examples），一般称为 LPU 或 PU 学习，是一种半监督学习方法。

Pig：在 HDFS 和 MapReduce 上处理大规模数据集的脚本语言，它提供更高层次的抽象并转化为优化处理的 MapReduce 运算。

频繁集（frequent itemset）：是大于最小支持度的项目集。

强关联规则：如果某条规则同时满足最小支持度（min-support）和最小置信度（min-confidence），则称它为强关联规则。

R 语言：R 是属于 GNU 系统的一个自由、免费、源代码开放的软件，是一个用于统计计算和统计制图的工具。

REST（Representational State Transfer，表现状态转移）：是 Roy Fielding 博士在 2000 年的博士论文中提出来的一种软件架构风格，在此风格中，每个资源是由全球唯一的 URI 来指定，资源本身和其表现方式是完全独立的；当一个用户拿到资源的表现方式时，他有足够的信息可以修改或者删除服务器上相应的资源而且每条消息都包含了足够的信息可以描述消息的处理。

热图（heat map）：热图或热力图是数据的一种二维呈现，其中的数值都用颜色表示。一个简单的热图提供信息的即时可见概况。

人工神经网络（Artificial Neural Networks）：一种模范动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。

人工智能（Artificial Intelligence）：研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。

3C 产品：3C 产品指的是通信产品（Communication）、消费类电子产品（Consumer Electronics）和计算机产品（Computer），三类产品的首字母都是 C，所以称 3C。

SEMMA：是数据挖掘过程（Sample, Explore, Modify, Model,and Assess 的英文缩写），意思是抽样、检查、修改、设立模型和评估。

熵（Entropy）：指的是体系的混乱的程度，它在控制论、概率论、数论、天体物理、生命科学等领域都有重要应用，在不同的学科中也有引申出的更为具体的定义，是各领域十分重要的参量。熵由鲁道夫·克劳修斯（Rudolf Clausius）提出，并应用在热力学中。后来，克劳德·艾尔伍德·香农（Claude Elwood Shannon）第一次将熵的概念引入到信息论中来。

商业智能（Business Intelligence）：采用数据库或数据仓库技术进行商业信息的收集、集成、分析和报告以帮助做决策的应用与实践系统。

时间序列（Time Series）：是指将某种现象某一个统计指标在不同时间上的各个数值，按时间先后顺序排列而形成的序列。时间序列法是一种定量预测方法，也称简单外延方法。

事务数据库（Transaction Database）：由文件构成，每条记录代表一个事务。典型的事务包含唯一的事务标记，多个项目组成一个事务。

数据可视化（Data Visualization）：关于数据的视觉表现形式的研究，这种数据的视觉表现形式被定义为一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。

数据挖掘（Data Mining）：从存放在数据库，数据仓库或其他信息库中的大量的数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的过程。

数据可视化（Data Visualization）：多维度数据通过图形的方式来做的展现。

数据仓库：是决策支持系统（DSS）和联机分析应用数据源的结构化数据环境。数据仓库研究和解决从数据库中获取信息的问题。数据仓库的特征在于面向主题、集成性、稳定性和时变性。

数据清洗（data cleaning）：过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。

数据库（Database）：是按照数据结构来组织、存储和管理数据的仓库。

属性（attribute）：属性是实体的描述性性质或特征，具有数据类型、域、默认值三种性质。属性也往往用于对控件特性的描述。对于按钮控件的名称、显示的文字、背景色，背景图片，等等。

SNS：社会化服务网络，Social Services Networks 的英文首字母缩写。

spatio-temporal data mining：时空数据挖掘的领域。

Sqoop：一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具。

特征选择（Feature Selection）：是指从已有的 M 个特征（Feature）中选择 N 个特征使系统的特定指标最优化。

统计学（Statistics）：是应用数学的一个分支，主要通过利用概率论建立数学模型，收集所观察系统的数据，进行量化的分

析、总结，并进而进行推断和预测，为相关决策提供依据和参考。它被广泛的应用在各门学科之上，从物理和社会科学到人文科学，甚至被用在工商业及政府的情报决策之上。

跳出率（Bounce Rate）：是互联网上的一个常用指标，指的是进入某一个网站之后不再继续浏览，而直接离开网站的访客比例。通常来说，跳出率越高，网站的粘性就越低。

Traffic：见“流量”词条。

UGC：User Generated Content 的缩写，即用户生成内容。

Web log 项（日志项）：网络上的服务器记录所有访问该 Web 服务器的数据流的信息。

Web 挖掘（Web Mining）：Web 挖掘是数据挖掘在 Web 上的应用，它利用数据挖掘技术从与 WWW 相关的资源和行为中抽取感兴趣的、有用的模式和隐含信息，涉及 Web 技术、数据挖掘、计算机语言学、信息学等多个领域，是一项综合技术。

围绕中心点的划分聚类算法（PAM）：通过反复地用非代表对象来代替代表对象，提高聚类的质量的算法。

唯一浏览量：是指网站来源是搜索引擎下的广告主网站的唯一浏览量，即在浏览量的基础上，不被记作重复的浏览量，刷新的浏览量不被记作唯一浏览量。

先验概率：见“后验概率”词条。

线性模型（Linear Model）：是一种分析模型，它假定考虑的各变化因素是线性的关系。

协作推荐：是利用用户访问行为的相似性来相互推荐用户可能感兴趣的资源。

文本挖掘（Text Mining）：指从文本数据中抽取有价值的信息和知识的计算机处理技术。即从文本中进行数据挖掘。从这个

意义上讲，文本挖掘是数据挖掘的一个分支，由机器学习、数理统计、自然语言处理等多种学科交叉形成。

信息检索（Information Retrieval）：指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。

信息增益（Information Gain）：是衡量一个属性区分数据样本的能力。信息增益量越大，对信息分类的能力就越强。而用来计算信息增益的公式就需要用到熵（Entropy）。

相关分析（Correlation Analysis）：相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度，是研究随机变量之间的相关关系的一种统计方法。

序列算法：在数据挖掘中的序列算法是对于一个序列（Sequence）中的数据找出统计规律的算法。

异常点（Outlier）：在大规模数据集中，通常存在着不遵循数据模型的普遍行为的样本。这些样本和其他部分数据有很大不同或不一致，叫作异常点（Outlier）。

异常值（Anomaly）的定义：是基于某种度量而言，异常值是指样本中的个别值，其数值明显偏离它（或他们）所属样本的其余观测值。

遗传算法（Genetic Algorithm）：是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。

元数据（Metadata）：是指描述数据仓库内数据的结构和建立方法的数据，是关于数据的数据，是对数据的结构、内容、键码、索引等的一种描述。

ZB：计算机存储单位。1 ZB = 1,024 EB = 1,180,591,620,717,411,303,424 Bytes（字节），或者是 2 的 70 次方字节。

召回率（Recall Rate，也叫查全率）：是检索出的相关文档数和文档库中所有的相关文档数的比率。

直效行销（Direct Marketing）：又名零阶通路，是指制造商或零售商，直接将产品出售给消费者，使通路阶层降至零阶或一阶，减少中间费用，为消费者取得较低价格的销售方式。

知识工程（Knowledge Engineering）：人工智能的原理和方法，对那些需要专家知识才能解决的应用难题提供求解的手段。

知识发现（KDD：Knowledge Discovery in Databases）：从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。

支持度（Support）：描述关联规则的阈值，反映关联规则在数据库中的重要性。

支持向量机（Support Vector Machine，SVM）：Corinna Cortes 和 Vapnik8 等人于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

主成分分析（Principal Component Analysis，PCA）：将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。

转化率（Conversion Rate）：指的是产生实际消费的用户和来到用户网页的总用户数量的比值，是将流量转化为实际的销售额的一种衡量方式。

着陆页（Landing Page）：指的是网站中的一个市场营销专用页面，通常是搜索引擎或是其他广告所指向的页面。

自助法（Bootstrap）：非参数统计中一种重要的估计统计量，采用重抽样技术从原始样本中抽取一定数量（自己给定）的样本。

Zookeeper：一个针对大型分布式系统的可靠协调系统，提供功

能包括配置维护、名字服务、分布式同步、组服务等。

最大频繁项集（Maximal Frequent Itemsets，MFI）：频繁地出现在数据集中的最大子集。

最大似然估计：是用来求一个样本集的相关概率函数的参数的一种统计方法。

# 附录 B
# 英语参考文献表


[1] Mohamed Medhat Gaber (Editor). *Journeys to Data Mining: Experiences from 15 Renowned Researchers [Hardcover]*. Springer, 2012，ISBN-10: 3642280463，with contributions from Dean Abbott, Charu Aggarwal, Michael Berthold et al.

[2] Nancy Lynch and Seth Gilbert. *Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services*. ACM SIGACT News, Volume 33 Issue 2 (2002), pg. 51-59.

[3] Galit Shmueli. *Practical Time Series Forecasting*. July 2012: 2nd Edition.

[4] Bruce Ratner, Ph.D. *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data (4th printing)*. CRC Press, ISBN: 1574443445.

[5] Robert I. Kabacoff. *R in Action : Data Analysis and Graphics with R*. August, 2011, ISBN:9781935182399.

[6] Quinlan, J. R. 1993. *C4.5*: *Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[8] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou,

M. Steinbach, D. J. Hand, D. Steinberg. *Top 10 Algorithms in Data Mining*. Knowl Inf Syst (2008) 141-37.

[9] Hand, D.J., Yu, K., 2001. *Idiot's Bayes*: *Not So Stupid After All*? International Statistics Rev. 69, 385-398.

[10] Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc.

[11] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* J. Wiley, New York.

[12] Tom Mitchell. *Machine Learning.* McGraw Hill, 1996.

[13] Bing Liu. *Web Data Mining：Exploring Hyperlinks, Contents, and Usage Data*. Springer，Second Edition, July 2011.

[14] Brin, S. and Page, L. *The anatomy of a large-scale hypertextual Web search engine.* In Proceedings of the Seventh international Conference on World Wide Web (WWW-7) (Brisbane, Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 107-117. 1998.

[15] Kleinberg, J. M. *Authoritative sources in a hyperlinked environment*. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, California, United States, January 25 - 27, 1998). Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA, 668-677. 1998.

[16] Leland Wilkinson, D. Wills, D. Rope et al. *The Grammar of Graphics* (*Statistics and Computing*). Springer, 2005.

[17] Thuraisingham, B. *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism*. Crc Press，2003.

[18] Freund, Y. and Schapire, R. E. *A deci sion-theoretic generalization of on- line le arning and a n ap plication to boosting*. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

[19] Andras A. Benczur, Karoly Csalogany, Tamas Sarlos, Mate Uher, *SpamRank-Fully Automatic Link Span Detection*, In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[20] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. *PrefixSpan: Mi ning S equential Patterns E fficiently by P refix-Projected Pattern Gr owth*. In Proceedings of the 17th international Conference on Data Engineering (April 02 - 06, 2001). ICDE '01. IEEE Computer Society, Washington, DC.

[21] Liu, B., Hsu, W. and Ma, Y. M. *Integrating c lassification and association rule mining*. KDD-98, 1998, pp. 80-86.

[22] Zdzislaw Pawlak. *Rough Se ts*: *Th eoretical Asp ects of Reasoning ab out Data,* Kluwer Academic Publishers, Norwell, MA, 1992.

[23] Varun Chandola, Arindam Banerjee, and Vipin Kumar, *Anomaly De tection* : *A Survey* (*2009*) ACM Computing Surveys. Vol. 41(3), Article 15, July 2009.

[24] Brin S，Motwani R，Silverstein C. *Beyond market*： *Generalizing association rules to correlations*[*A*]. Processing of the ACM SIGM OD Conference 1997 [C]．New York：ACM Press，1997．265-2.

[25] R. Agrawal, *Data Mining association rules between sets of items in large d atabases*. In： Proceedings of ACM SIGMOD Conference on Management of Data，

Washington, DC May 1993. 207-216.

[26] Hui Xiong, Junjie Wu, Jian Chen. *K-means Clustering Versus Validation Measures: A Data Distribution Perspective.* IEEE Transactions on Systems, Man, and Cybernetics --- Part B (TSMCB), Vol. 39, No. 2. (2009), pp. 318-331, 2009.

[27] R. Agrawal, et al. *Mining association rules between sets of items in large databases.* Proc. ACM SIGMOD int'l conf. management of data, Washington, DC, May 1993, 207-216.

[28] R. Agrawal, R. Srikant. *Fast algorithms for mining association rules.* Proc. 20th int'l conf. very large databases, Santiago, Chile, Sept. 1994, 487-499.

[29] J. S. Park, et al. *Using a hash-based method with transaction trimming for mining association rules.* IEEE Transactions on knowledge and data engineering, 1997, 9(5), 813-825.

[30] L. Page, S. Brin, R. Motwani, T. Winograd. *The Pagerank Citation Ranking: Bringing Order to the Web*. Technical Report, Stanford University, 1999.

[31] Mark Hall，Eibe Frank et al. *The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter，Volume 11 Issue 1，Pages 10-18, June 2009.

[32] D. W. Cheung, et al. *Maintenance of discovered association rules in large databases: an incremental updating technique.* In: Proceedings of the 12th international conference on data engineering, New Orleans Louisiana, 1995, 106-114.

[33] R. Agrawal, et al. *Parallel mining of association rules.* IEEE Transactions on knowledge and data engineering, 1996, 8(6), 962-969.

[34] J. S. Park, et al. *Efficient parallel d ata min ing for association rules.* Proc. Fourth int'l conf. information and Knowledge management, Baltimore, Nov. 1995.

[35] D. W. Cheung, et al. *efficient m ining of association ru les in distributed databases.* IEEE Transactions on knowledge and data engineering, 1996, 8(6), 910-921.

[36] D. W. Cheung, et al. *A f ast d istributed al gorithm f or mining association rules.* Proc. of 1996 Int'l Conf. on Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA, Dec. 1996.

[37] J. Han, Y. Fu. *Discovery of multiple-level association rules from large d atabases. P roc.* of the 21st international conference on very large databases, Zurich, Switzerland, Sept. 1995, 420-431.

[38] R. Srikant, R. Agrawal. *Mining ge neralized association rules.* In: Proceedings of the 21st international conference on very large databases, Zurich, Switzerland, Sept. 1995, 407-419.

[39] Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar. *Enhancing D ata Analysis wi th No ise Rem oval*. IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol. 18, No. 3, pp. 304-319, March 2006.

[40] Varun Chandola and Vipin Kumar. *Summarization - Compressing Da ta in to an Inf ormative Re presentation* (*2006*). Knowledge Discovery and Information Systems (KAIS), Vol. 12(3), 2007.

[41] R. Srikant, R. Agrawal. *Mining association rules with item constrains.* Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997, 67-73.

[42] R. Ng, L. V. S. Lakshmanan, J. Han and A. Pang. *Exploratory Mining a nd Prun ing Optimizations of Constrained Asso ciations Rules,* Proc. of 1998 ACM-SIGMOD Conf. on Management of Data, Seattle, Washington, June 1998, 13-24.

[43] L. V. S. Lakshmanan, R. Ng, J. Han and A. Pang. *Optimization of Co nstrained Fr equent Set Q ueries wi th 2-Variable Constraints,* Proc. 1999 ACM-SIGMOD Conf. on Management of Data, Philadelphia, PA, June 1999.

[44] R. Ng, L. V. S. Lakshmanan, J. Han and T. Mah. *Exploratory Mining via Constrained Frequent Set Queries,* Proc. 1999 ACM-SIGMOD Conf. on Management of Data, Philadelphia, PA, June 1999.

[45] R. J. Bayardo Jr., R. Agrawal, and D. Gunopulos. *Constraint-Based Rule Mining in Large, Dense Databases.* In Proc. of the 15th Int'l Conf. on Data Engineering, 1999, 188-197.

[46] Micheline Kamber, Jiawei Han, Jenny Y. Chiang. *Metarule-Guided Mining of Multi-Dimensional Association Rules Usi ng Da ta Cubes.* Proceeding of the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, Aug. 1997, 207-210.

[47] Y. Fu and J.Han. *Meta-Rule-Guided Mining of Association Rules in Relational Databases,* Proc. 1995 Int'l Workshop. on Knowledge Discovery and Deductive and Object-Oriented Databases(KDOOD'95), Singapore, December 1995, pp.39-46.

[48] Hui Xiong, Shashi Shekhar, Yan Huang, Vipin Kumar, X. Ma, J. Yoo. *A Fram ework for D iscovering Co-l ocation*

*Patterns in Data Sets with Extended Sp atial Objects (2004)*. In Proc. 2004 SIAM International Conf. on Data Mining (SDM'04), Florida, USA, 2004.

[49] Pang-Ning Tan, and Vipin Kumar. *Mining In direct Associations in Web Data* (*2001*). WebKDD 2001: Mining Log Data Across All Customer Touch Points.

[50] R. J. Bayardo Jr. *Efficiently M ining Lo ng Patterns fr om Databases.* Proc. of the ACM SIGMOD Conference on Management of Data, Seattle, Washington, June 1998, 85-93.

[51] S. Brin, R.Motwani, and C. Silverstein. *Beyond market basket: generalizing association rules to correlation.* Proc. 1997 ACM-SIGMOD int. conf. management of data, Tucson, Arizona, May 1997, 265-276.

[52] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. *Scalable t echniques for m ining causal s tructures.* Proc. 1998 int. conf. Very Large Data Bases, New York, NY, August 1998, 594-605.

[53] F. Korn, et al. *Ratio ru les*: *A ne w paradigm for fas t, quantifiable data mining.* Proc. 1998 int. conf. Very Large Data Bases, New York, NY, August 1998, 582—593.

[54] B. Ozden, et al. *Cylic a ssociation ru les*. Proc 1998 int. conf. Data Engineering, Orlando, FL, Feb. 1998, 412-421.

[55] S. Ramaswamy, et al. *On t he discovery of i nteresting patterns in as sociation r ules, Pr oc.* 1998 int. conf. Very Large Data Bases, New York, NY, August 1998, 368-379.

[56] Levent Ertoz, Michael Steinbach, and Vipin Kumar. *Finding Clusters of D ifferent Sizes, Shapes, and D ensities in Noisy, H igh D imensional Data ( 2003)*. SIAM International Conference on Data Mining (SDM '03).

[57] Levent Ertoz, Michael Steinbach, and Vipin Kumar. *Finding Topics in Collections of Documents*: *A Shared Nearest Neighbor Approach*. Clustering and Information Retrieval, forthcoming 2003, Kluwer Academic Publishers, 2003.

[58] Michael Steinbach, Levent Ertoz, and Vipin Kumar. *Challenges of Clustering High Dimensional Data* (*2003*). New Vistas in Statistical Physics -- Applications in Econophysics, Bioinformatics, and Pattern Recognition, forthcoming 2003, Springer-Verlag.

[59] Levent Ertoz, Michael Steinbach, and Vipin Kumar. *A New Shared Nearest Neighbor Clustering Algorithm and its Applications* (*2002*). Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining (2002).

[60] Zhang, T., Ramakrishnan, R., and Livny，BIRCH: an efficient data clustering method for very large databases, In Proceedings of the 1996 ACM SIGMOD international Conference on Management of Data (Montreal, Quebec, Canada, June 04 - 06, 1996). J. Widom, Ed. SIGMOD '96. ACM Press, New York, NY, 103-114.

[61] Yan, X. and Han, J. *gSpan*: *Graph-Based Substructure Pattern Mining*. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02) (December 09 - 12, 2002). IEEE Computer Society, Washington, DC.

[62] R. Srikant, R. Agrawal. *Mining quantitative association rules in large relational tables*. In: Proc. 1996 ACM SIGMOD int'l Conf. Management Data, Montreal, Canada, 1996, 1-12.

[63] Vipin Kumar. High Performance Data Mining, Keynote Presentation at High Performance Computing for Computational Science, VECPAR 2002, June 27, 2002.

[64] Ar Lazarevic , Aysel Ozgur , Levent Ertoz , Jaideep Srivastava, Vipin Kumar. *A C omparative S tudy of Anomaly D etection Schemes in Network I ntrusion Detection.* In Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA.

[65] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan, *Data Mining for Netw ork Intrusion Detection* (2002). Proc. NSF Workshop on Next Generation Data Mining, Baltimore, MD.

[66] Aleksandar Lazarevic, Paul Dokas, Levent Ertoz, Vipin Kumar, Jaideep Srivastava, Pang-Ning Tan, *Cyber Threat Analysis - A Key En abling Technology for the O bjective Force ( A Ca se S tudy in Network I ntrusion De tection)* (2002). Proceedings 23rd Army Science Conference, Orlando, FL.

[67] Pang-Ning Tan, and Vipin Kumar, *Mining Asso ciation Patterns i n W eb U sage D ata* (2002). International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet.

[68] Pang-Ning Tan, and Vipin Kumar, *Discovery of Web Robot Sessions b ased on their N avigational P atterns* (2002). Data Mining and Knowledge Discovery, 6(1):9-35.

[69] Hastie, T. and Tibshirani, R. 1996. *Discriminant Adaptive Nearest Neighbor Classification.* IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.

[70] A. Savasere, E. Omiecinski and S. Navathe. *An efficient*

*algorithm for mining association rules.* Proceedings of the 21st international conference on very large databases, Zurich, Switzerland, Sept. 1995, 432-444.

[71] M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Review, 2003, 45, 167-256.

[72] L. Getoor. *Link Mining*: *A New Data Mining Challenge*. SIGKDD Explorations, 2003, 5(1), 84-89.

[73] Cheung D W et at. *Maintenance of discovered association rules in lar ge databases; an incr emental updati ng technique*, Proceed of the 12th International Conference on Data Engineering, New Orleans Louisana, 1999. 106-114.

[74] Varun Chandola, Shyam Boriah, and Vipin Kumar, *A Framework f or An alyzing Cat egorical Data* (2009). In Proceedings of SIAM Data Mining Conference, April 2009, Sparks, NV.

[75] Wei Fan, Yi-an Huang, Haixun Wang, and Philip S. Yu. *Active Mining of D ata S treams* (*2004*). Proceedings of SIAM International Conference on Data Mining 2004.

[76]（印度）（Soumen Chakrabarti）查凯莱巴蒂.（英文版）Web 数据挖掘. 人民邮电出版社, 图灵原版计算机科学系列, 2009.

[77] Han, J., Pei, J., and Yin, Y. 2000. *Mining frequent patterns without candidate generation.* In Proceedings of the 2000 ACM SIGMOD international Conference on Management of Data (Dallas, Texas, United States, May 15 - 18, 2000). SIGMOD '00. ACM Press, New York, NY, 1-12.

[78] M Hu, B Liu, *Mining and summarizing customer reviews*, In Proceedings of the tenth ACM SIGKDD international Conference, 168-177，2004.

[79] Goh，Jen and Taniar, David. *An Ef ficient Mo bile Da ta*

*Mining Model：Parallel and Distribu ted Processing and Applications*. Springer Berlin，2005.

[80] Nafiseh Shabib, John Krogstie, *The use of data mining techniques in lo cation-based r ecommender sys tem*, Proceeding WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011.

[81] Rakesh Agrawal and Ramakrishnan Srikant, *Fast Algorithms for Mining Ass ociation Rules*, In Proc. of the 20th Int'l Conference on Very Large Databases (VLDB '94), Santiago, Chile, September 1994.

[82] MacQueen, J. B., *Some meth ods fo r classification an d analysis of m ultivariate observ ations*, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281-297.

[83] Nick Littlestone. *Learning Q uickly W hen Irr elevant Attributes Ab ound*: *A N ew Lin ear-threshold Al gorithm*. Machine Learning 285-318(2), 1988.

[84] Nick Littlestone. *Mistake b ounds a nd logarithmic linear-threshold l earning algorithms.* Technical report UCSC-CRL-89-11, University of California, Santa Cruz, 1989.

[85] Roy Thomas Fielding. *Architectural Styles and the Design of N etwork-based S oftware A rchitectures.* Ph. D Dissertation, University of California, Irvine.

[86] Srikant, R. and Agrawal, R. *Mining S equential Pa tterns: Generalizations a nd Per formance Im provements.* In Proceedings of the 5th international Conference on Extending Database Technology: Advances in Database Technology (March 25 - 29, 1996). P. M. Apers, M.

Bouzeghoub, and G. Gardarin, Eds. Lecture Notes In Computer Science, vol. 1057. Springer-Verlag, London, 3-17.

[87] Hannu Toivonen, *Sampling lar ge dat abases for association rules*, In Proceedings of the 22nd international conference on very large databases, Bombay, India, 1996, 134-145.

[88] William N Venables and David M Smith. *An Introduction to R*, Second Edition, Network Theory Ltd, 2009.

[89] Leo Breiman. *Bagging p redictors.* Machine Learning, 24（2):123-140, 1996.

# 附录 C
# 中文参考文献表


[1] （加）Jiawei Han; Micheline Kamber. 数据挖掘概念与技术. 范明, 孟小峰, 译. 机械工业出版, 2007.

[2] （美）Olivia Parr Rud. 数据挖掘实践[M]. 朱扬勇等, 译. 机械工业出版, 2003.

[3] 夏火松. 数据仓库与数据挖掘技术[M]. 科学出版社, 2004.

[4] （美）Richard J.Roiger,（美）Michael W.Geatz**.** 数据挖掘教程[M]. 翁敬农, 译. 清华大学出版社, 2003.

[5] 段云峰等. 数据仓库及其在电信领域中的应用[M]. 电子工业出版社, 2003.

[6] （美）荫蒙（Inmon,W.H）. 数据仓库（第 4 版）[M]. 王志海等, 译, 机械工业出版社, 2006.

[7] 陈京民. 数据仓库原理、设计与应用[M]. 中国水利水电出版社, 2004.

[8] （美）Lou Agosta. 数据仓库技术指南[M]. 潇湘工作室, 译. 人民邮电出版社, 2000.

[9] （美）Pang-Ning Tan, Michael Steinbach, Vipin Kumar. 数据挖掘导论[M]. 范明, 范宏建,译. 人民邮电出版社, 2010.

[10] （美）拉贾拉曼（Anand Rajaraman）,（美）厄尔曼（Jeffrey David Ullman）. 大数据：互联网大规模数据挖掘与分布式处理[M]. 王斌, 译, 人民邮电出版社, 2012.

[11] （美）Haralambos Marmanis, Dmitry Babenko. 智能 Web 算法[M]. 阿稳, 陈钢, 译, 电子工业出版社, 2011.

[12] （土）阿培丁. 机器学习导论[M]. 机械工业出版社, 2009.

[13] （美）Matthew A. Russell. 社交网站的数据挖掘与分析[M]. 师蓉, 译. 机械工业出版社华章公司, 2012.
[14] 曾万聃, 周绪波, 戴勃等. 关联规则挖掘的矩阵算法[J]. 计算机工程, 2006（1）32（2）：45-47.
[15]杨怡玲, 管旭东, 陆丽娜等. 一个简单的 Web 日志挖掘系统. 上海交通大学学报, 2000, 34（7）.
[16] 马征, 李建华. 基于多代理技术的分布式 Web 日志挖掘系统[J]. 微计算机信息（测控仪表自动化）, 2004, 4.
[17] 戴小华. 基于 R 语言的中国柑橘主要病虫害空间分布图[J]. 江西农业学报, 2009, 21（04）.
[18] 董祥军, 陈建斌, 崔林等. 正、负关联规则间的置信度关系研究[J]. 计算机应用及研究, 2005, 22（7）.
[19] 张朝晖, 陆玉昌, 张钹. 发掘多值属性的关联规则[J]. 软件学报, 1998, 9（11）：801-805.
[20] 铁治欣, 陈奇, 俞瑞钊. 关联规则采掘综述[J]. 计算机应用研究, 2000, 17（1）：1-5.
[21] 潘立武, 王保保, 李绪成. 零售业销售数据关联规则挖掘算法关键思想研究[J]. 信阳师范学院学报（自然科学版）, 2003（1）, 16（1）：90-92.
[22]冯玉才, 冯剑林. 关联规则的增量式更新算法[J]. 软件学报, 1998, 9（4）：301-306.
[23] 肖劲橙, 林子禹, 毛超. 关联规则在零售商业的应用[J]. 计算机工程, 2004（2）, 30（3）：301-306.
[24] 李宝东, 宋瀚涛. 关联规则增量更新算法研究[J]. 计算机工程与应用, 2002, 38（23）：6-8.
[25] 张伟, 郑涛, 李辉. 一种并行化的分组关联规则算法[J]. 计算机工程, 2004（11）, 30（22）：84-86.
[26] 李逸波, 于吉红, 白晓明. 合理选择数据挖掘工具[J]. 计算机与信息技术, 2005（08）, 6.
[27] 邹丽, 孙辉, 李浩. 分布式系统下挖掘关联规则的两种方案

[J]. 计算机应用研究, 2006, 1：77-78.
[28] 蔡伟杰, 张晓辉, 朱建秋等. 关联规则挖掘综述[J]. 计算机工程, 2004, 27（5）.
[29] 郎璟, 王保保. 关联规则挖掘结果的可视化技术研究[J]. 电子科技, 2004, 10.
[30] 吉根林, 韦索云, 曲维光. 基于平行坐标的关联规则可视化新技术[J]. 计算机工程, 2005（12）, 31（24）.
[31] 孙圣力, 李金玖, 朱扬勇. 高效处理分布式数据流上 skyline 持续查询算法[J]. 软件学报, 2009, 20（7）.
[32] 袁军鹏, 朱东华. 基于 Apriori 算法的多循环关联规则挖掘综述[J]. 计算机工程, 2004, 31（1）.
[33] 宋威, 杨炳儒, 徐章艳等. 基于索引数组和复合频繁模式树的频繁闭项集挖掘算法[J]. 计算机科学, 2007, 34（8）.
[34] 基于 Web 数据挖掘的高效关联规则研究[J]. 计算机工程与科学, 2005, 27（11）.
[35] 沈洁, 薛贵荣. 一种基于 XML 的 Web 数据挖掘模型[J]. 系统工程理论与实践, 2002, 22（9）.
[36] 徐晖. Web 数据挖掘和数据仓库在电子商务市场营销管理中的应用[J]. 新西部（下半月）, 2007, 11.
[37] 王闯舟. 数据仓库解决方案在电子商务中的应用[J]. 通信市场, 2002, 11.
[38] 张榛楠, 钱旭, 江涛. 面向电子商务的 Web 使用挖掘数据仓库设计与实现[J]. 制造业自动化, 2008, 30（9）.
[39] 熊赟, 邱伯仁, 张坤等. Gen-Cluster:一个基因表达数据的高维聚类算法[J]. 复旦学报（自然科学版）, 2008, 47（2）.
[40] 陈耿, 朱玉全, 杨鹤标等. 关联规则挖掘中若干关键技术的研究[J]. 计算机研究与发展, 2005, 42（10）.

# 附录 D
# 微博

下面列出一些在新浪微博 http://www.weibo.com 上的一些与数据分析和数据挖掘相关的微博。

@数据分析精选
@数据分析
@数据分析玩家
@数据挖掘与数据分析
@数据挖掘研究院
@数据化分析
@商业分析数据挖掘
@社会网络与数据挖掘
@数据驱动营销
@李航博士
@程苓峰
@刘铁岩
@沈浩老师
@张栋_机器学习
@黄萱菁
@ ICTCLAS 张华平博士
@王小川_Matlab
@谢幸 Xing

# 附录 E
# 博客和其他网址

http://www.kdnuggets.com 国外数据挖掘社区，由数据挖掘知名学者 Gregory Piatetsky 主持

http://www.dmg.org/ 由数据挖掘服务提供商组织的制作数据挖掘标准的数据挖掘组织官方网站

http://www.hbr.org 哈佛商业评论

http://www.apache.org/ Apache 开源基金会的首页

http://www.techopedia.com/ 与计算机信息系统相关的词典

http://visual.ly/ 数据可视化社区

http://www.analyticbridge.com/ Vincent Granville 缔造的数据分析社区

http://www.r-project.org/ R 语言的官方网址

http://nginx.org/en/docs/ Nginx 的相关文档

http://wiki.nginx.org/Main Nginx 的维基百科

http://www.itongji.cn/ 中国统计网

http://www.iresearch.cn/ 艾瑞网

http://www.chinabyte.com/ 比特网

http://www.csdn.net/ IT 中文社区

http://bbs.paidai.com/ 派代网电子商务社区

http://www.oschina.net/ 开源中国社区

http://mall.cnki.net/index.aspx 中国知网

http://www.techxue.com/portal.php 互联网分析沙龙

http://www.donews.com/ 互联网新闻

http://www.chinaz.com/ 站长之家

http://olap.com/w/index.php/OLAP_Education_Wiki 专门为OLAP打造的维基网站

https://developers.google.com/prediction/ 谷歌预测API的网址

http://www.r-bloggers.com 专注于R语言的博客

http://r-pbd.org/ Programming in Big Data with R的官方网站

http://pagelever.com/blog/ pageLever分析公司的官方博客

http://anti-spam.org.cn/ 中国防垃圾邮件联盟

http://blog.echen.me/ Twitter数据科学家Edwin Chen的博客

http://decisionstats.com/ 数据挖掘行业专家Ajay Ohri的博客

http://www.thearling.com/index.htm Kurt Thearling关于数据挖掘和分析技术的博客

http://www.dataminingblog.com/ data mining research，数据挖掘研究

http://www.simafore.com/blog/ 商业数据分析公司SimaFore的博客

http://solveforinteresting.com/ 分析师Alistair Croll主持的网站

http://www.the-art-of-web.com/ 互联网技术信息，包括服务器配置，CSS代码等

http://www.mathworks.cn/products/matlab/ MATLAB的主页

http://datamining.typepad.com/data_mining/ Matthew Hurst的博客

http://www.rexeranalytics.com/ Rexer Analytics公司的主页

http://www.information-management.com/sdm/1052295.html Steve Miller的博客

http://www.information-management.com/sdm/2000329.html Boris Evelson的博客

http://www.information-management.com/sdm/32112.html Mark Smith的博客

http://www.information-management.com/sdm/1033156.html Jim Ericson 的博客

http://davebeulke.com/blog/ DB2 专家 Dave Beulke 的博客

http://www.julianbrowne.com/ Julian Browne 的博客

http://www.deep-data-mining.com/ 数据挖掘在技术层面的博客

http://ldfry.wordpress.com/ Larry Fry 的数据挖掘博客

http://www.khabaza.com/ Tom Khabaza 的数据挖掘应用博客

http://www.datawrangling.com/ Pete Skomoroch 关于数据挖掘和机器学习的博客

http://abbottanalytics.blogspot.com/ DEAN ABBOTT 关于数据挖掘和预测分析的博客

http://timmanns.blogspot.com/ 数据挖掘博客

http://blog.markus-breitenbach.com/ markus breitenbach 关于人工智能，数据挖掘和机器学习的博客

http://andrewgelman.com/ 统计模型博客

http://textanddatamining.blogspot.com/ 文字和数据挖掘博客

http://intelligencemining.blogspot.com/ Venkatesh U 的数据挖掘博客

http://sherpablog.marketingsherpa.com/about/ 关于市场数据分析的博客

http://brendantierneydatamining.blogspot.com/ 基于 Oracle 的数据挖掘博客

http://matlabdatamining.blogspot.com/ 基于 MATLAB 的数据挖掘博客

http://lifeanalytics.blogspot.com/ 数据挖掘，文字挖掘和信息抽取的博客

http://datamininglab.blogspot.com/ 数据挖掘实验室的博客

http://cowlingreport.blogspot.com/ 数据挖掘博客

http://flowingdata.com/ 数据可视化博客

http://www.chinakdd.com/index.html 数据挖掘研究院

http://www.dataintoresults.com/ 数据挖掘博客

http://jtonedm.com/ 在决策管理中的数据挖掘和分析

http://spotfire.tibco.com/blog/ 软件公司 TIBCO 关于数据分析的博客

http://SPSS -market.r.blog.163.com/ 数据挖掘与数据分析

http://blog.digitalforest.cn 数据林网站分析博客

http://shenhaolaoshi.blog.sohu.com/178101622.html 沈浩老师的博客

http://chemyhuang.blog.163.com/blog 数据化管理

http://blog.sina.com.cn/s/blog_72e6be57010146qb.html 数据挖掘疯狂者博客

http://www.chinawebanalytics.cn/ 网站分析在中国