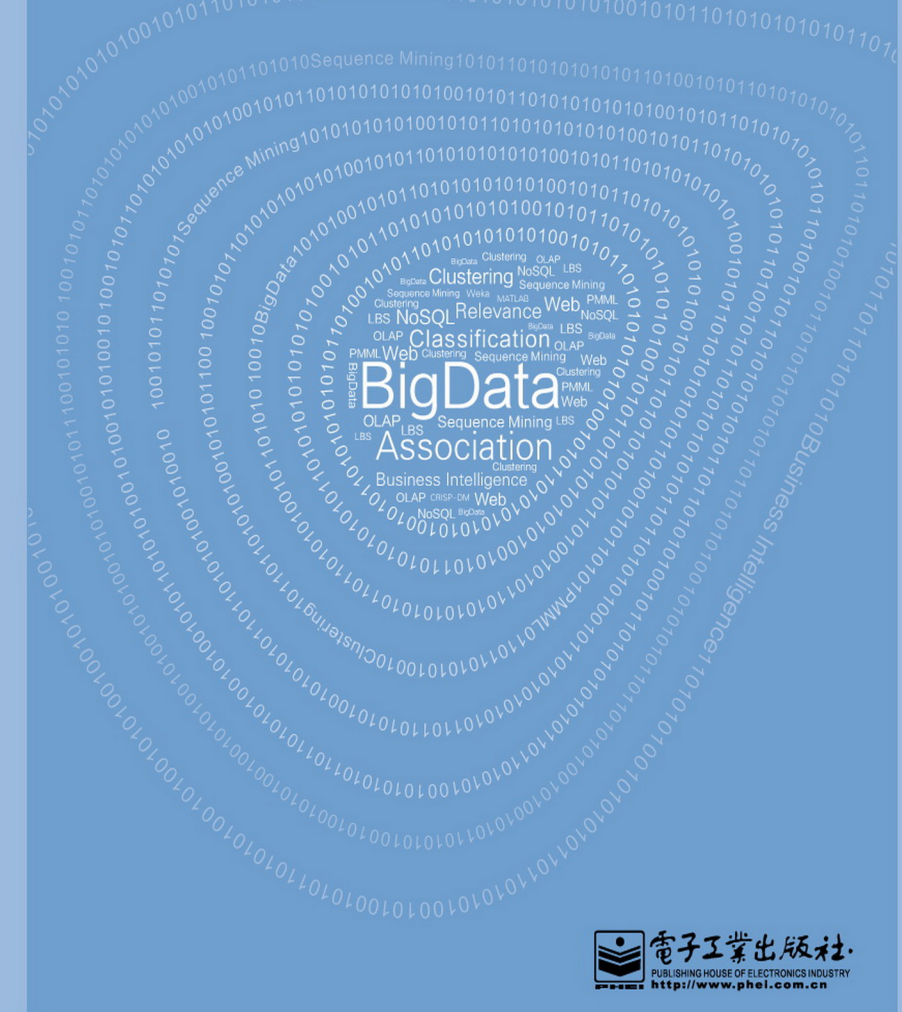
Broadview®
www.broadview.com.cn

这可能是最通俗易懂的一本数据挖掘书籍
——互动通邓广涛 PPTV陶闯 联合力荐

New Internet
大数据挖掘

谭磊 著



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

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。

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。

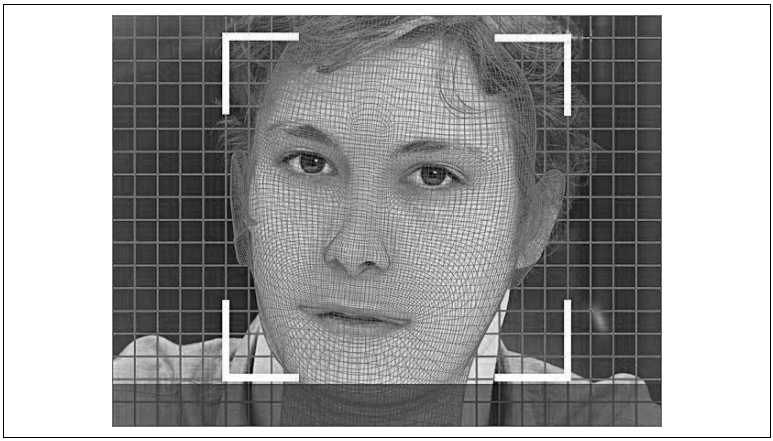


图 2-2　人脸识别示意图

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

		0	1
实际值	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。

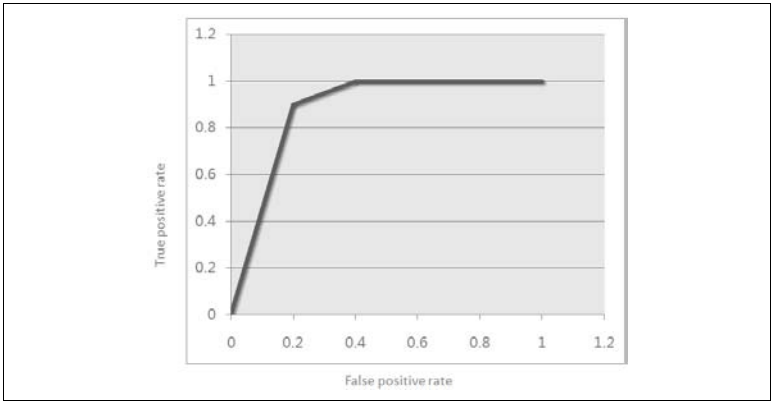


图 2-3　ROC 曲线图

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 $RPP = (B+D)/(A+B+C+D)$。

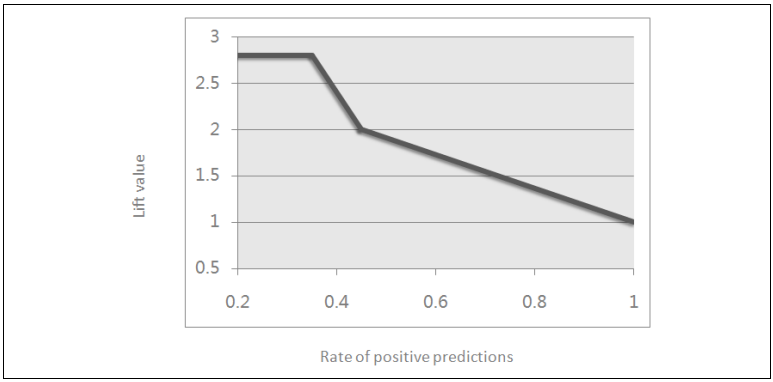


图 2-4　Lift 曲线图

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。

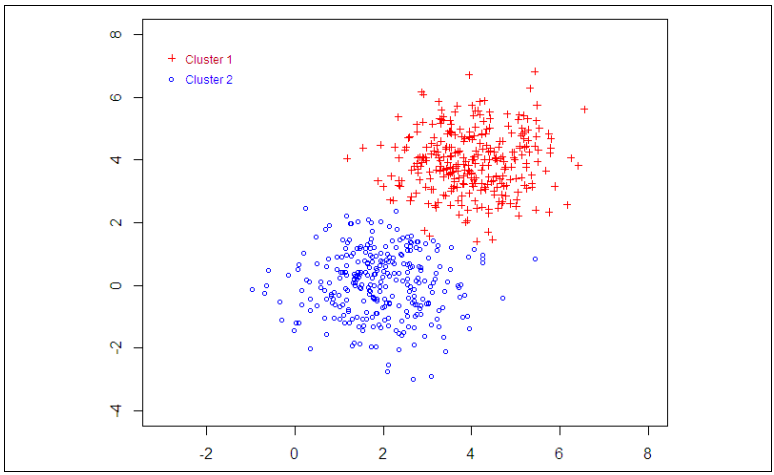


图 2-7　聚类算法示意图

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。

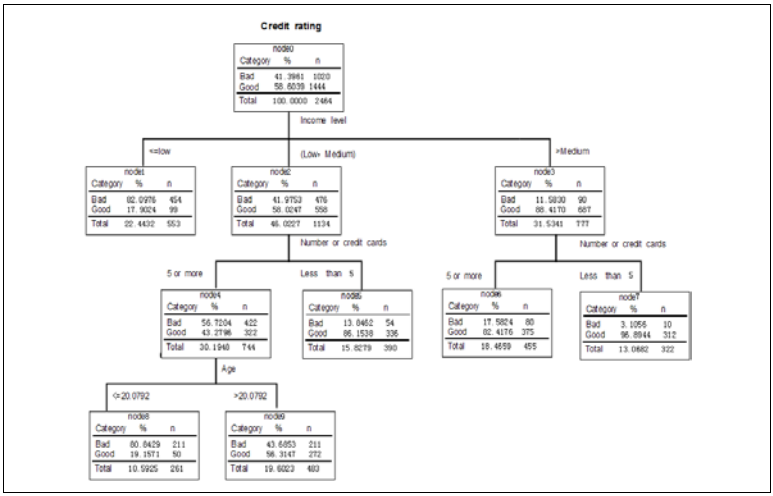


图 2-8　信用决策树示意图

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。

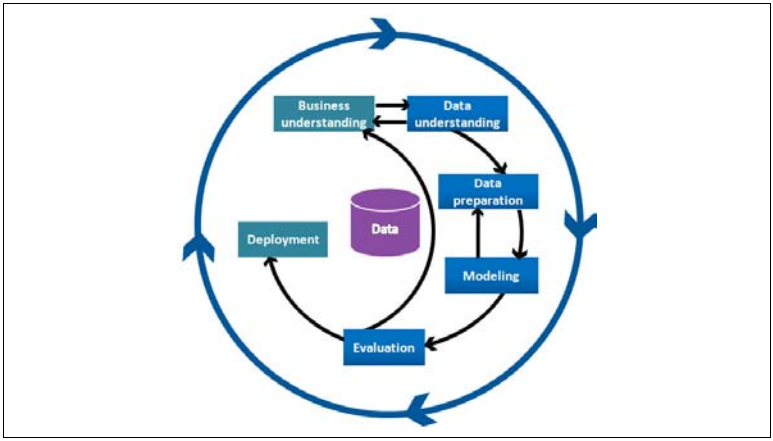


图 2-9　CRISP-DM 数据挖掘过程示意图

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。

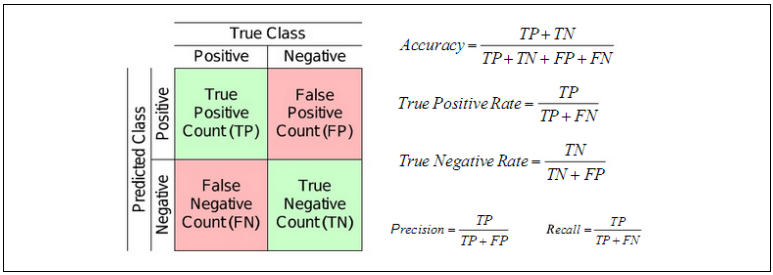


图 2-10　混淆矩阵示意图

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

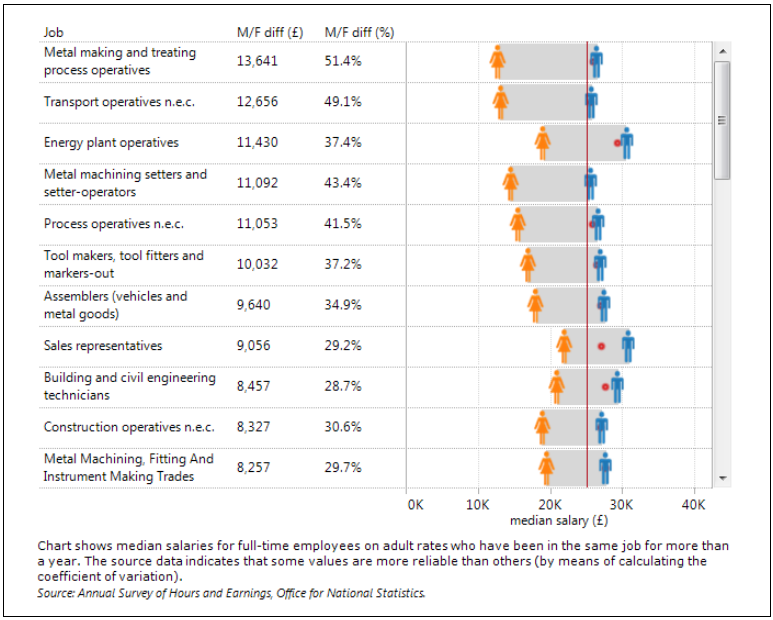


图 2-11　英国男女平均工资差距示意图

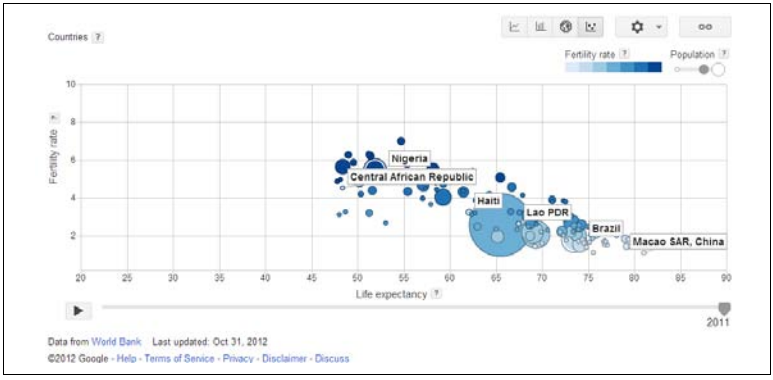


图 2-12　世界受孕率和平均寿命对比图

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。

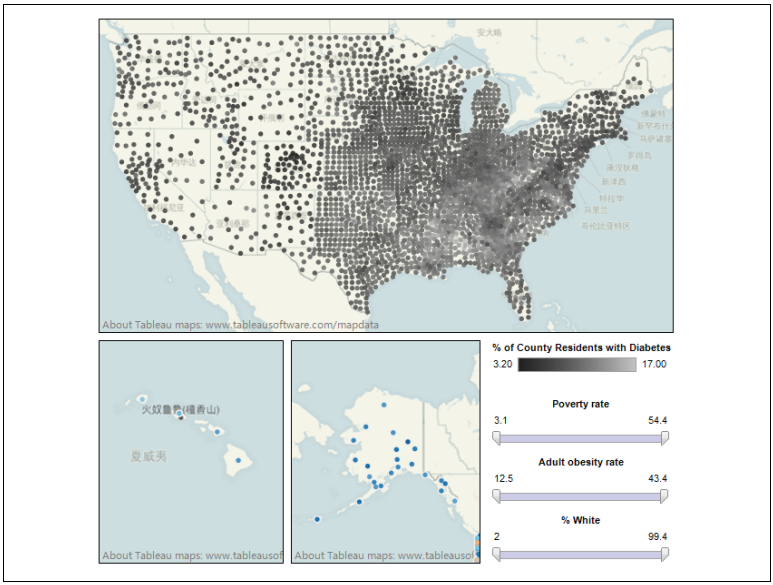


图 2-13　糖尿病占比示意图

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。

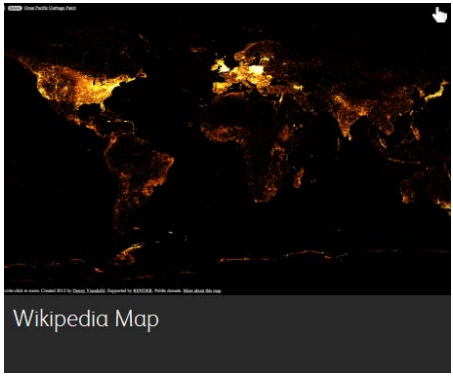


图 2-14　维基百科带地理位置文章发表示意图

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。

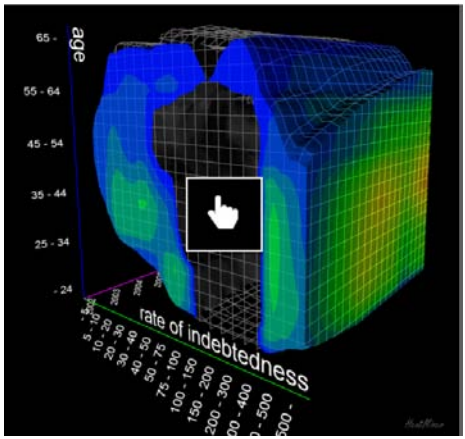


图 2-15　芬兰首都人民的年龄和负债率的对比示意图

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 Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance*, 2011-4-22.

[2] Thuraisingham, B. *Web Data Mining and Applications 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 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%。

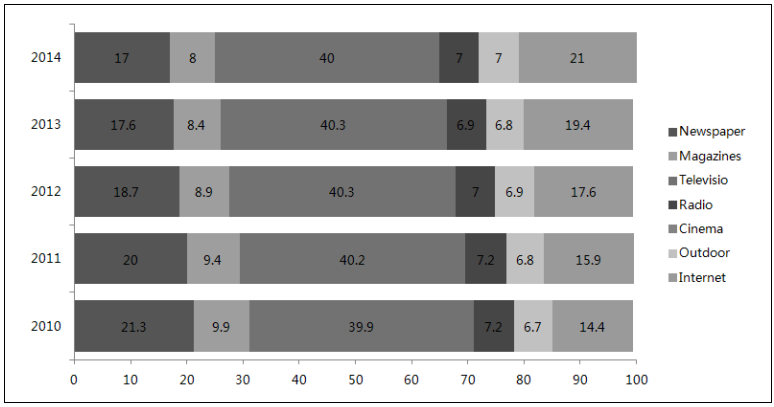


图 9-1　来自 Zenith Optimedia 的全球广告支出变化图

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。

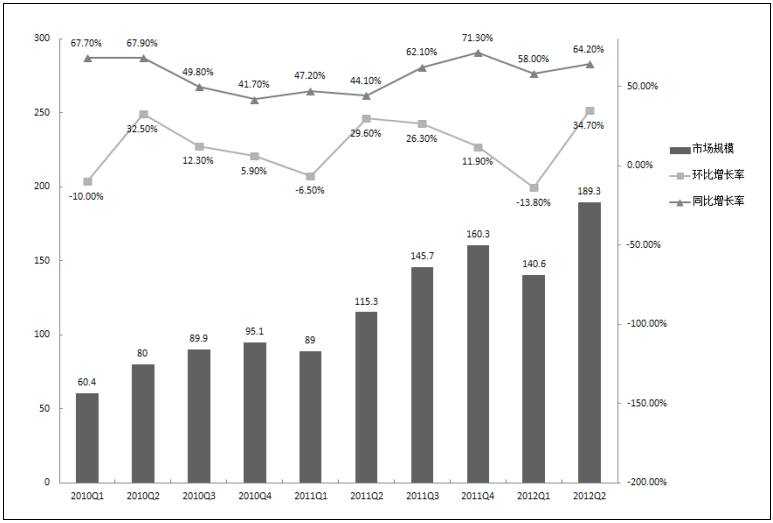


图 9-3　艾瑞咨询公司提供的中国网络广告市场规模

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。

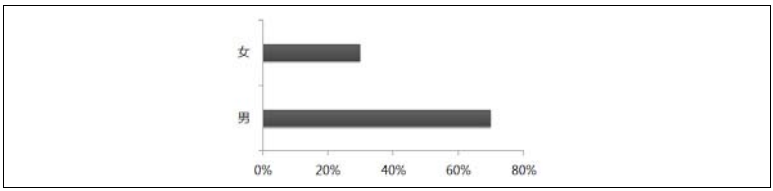


图 9-6　性别比例示意图

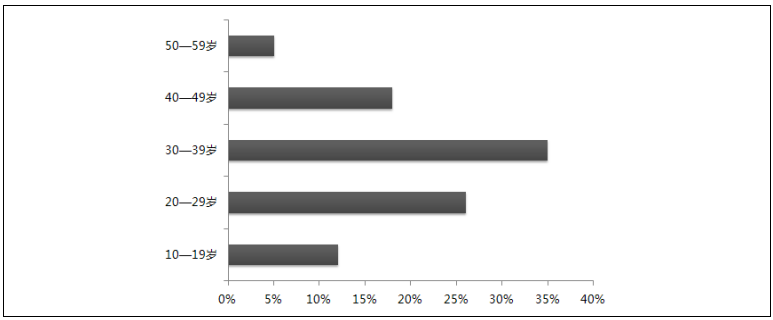


图 9-7　年龄分布示意图

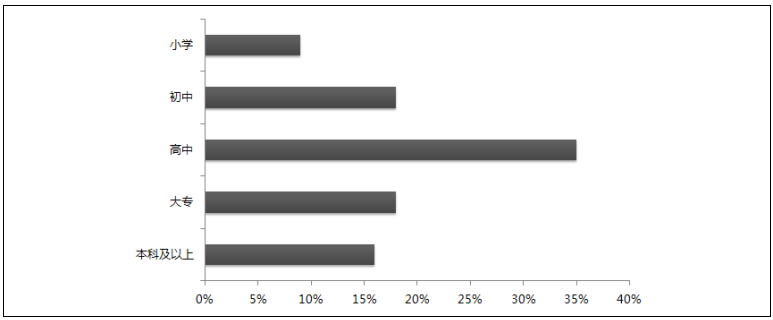


图 9-8　学历分布示意图

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。

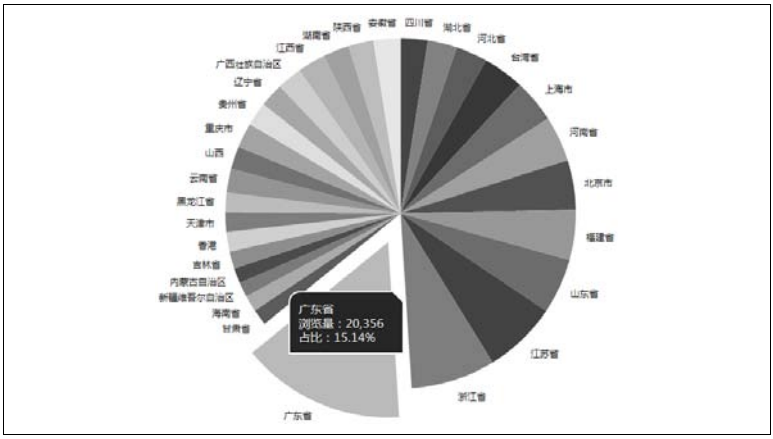


图 9-9　地域分布示意图

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。

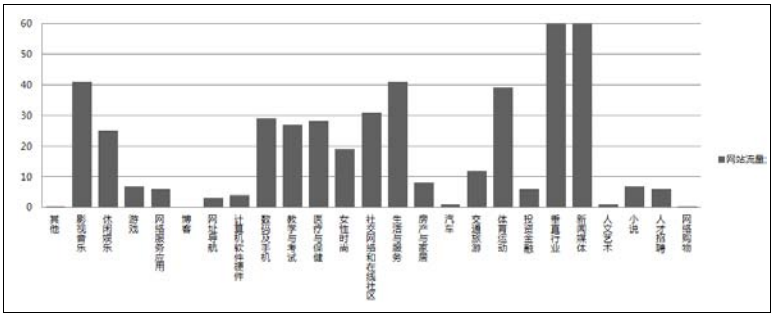


图 9-10　媒体种类分布示意图

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。

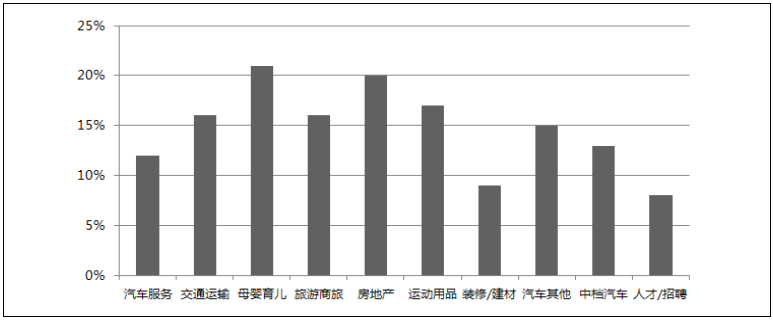


图 9-11　兴趣爱好分布示意图

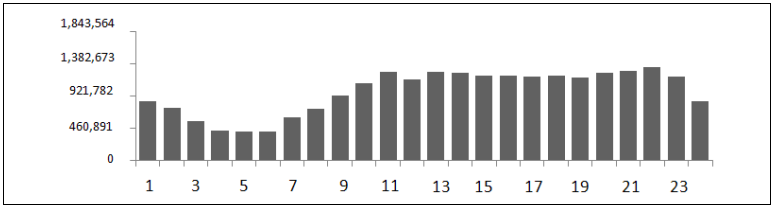


图 9-12　时段分布示意图

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。

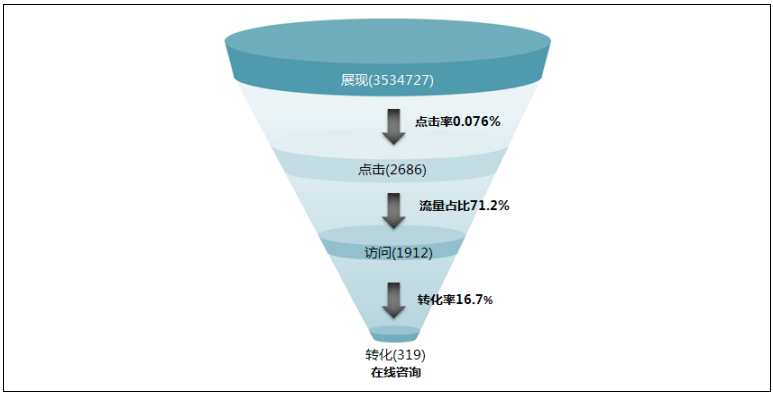


图 9-13　网盟广告投放转化漏斗示意图

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。

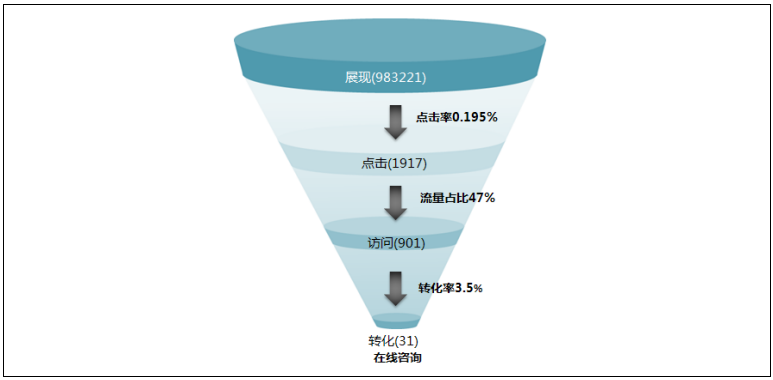


图 9-14　网盟广告投放小说阅读网站转化漏斗示意图

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。

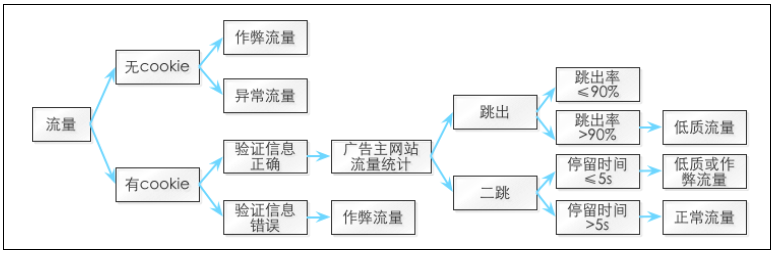


图 9-16　网盟作弊分析决策树示意图

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 Detection*: *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