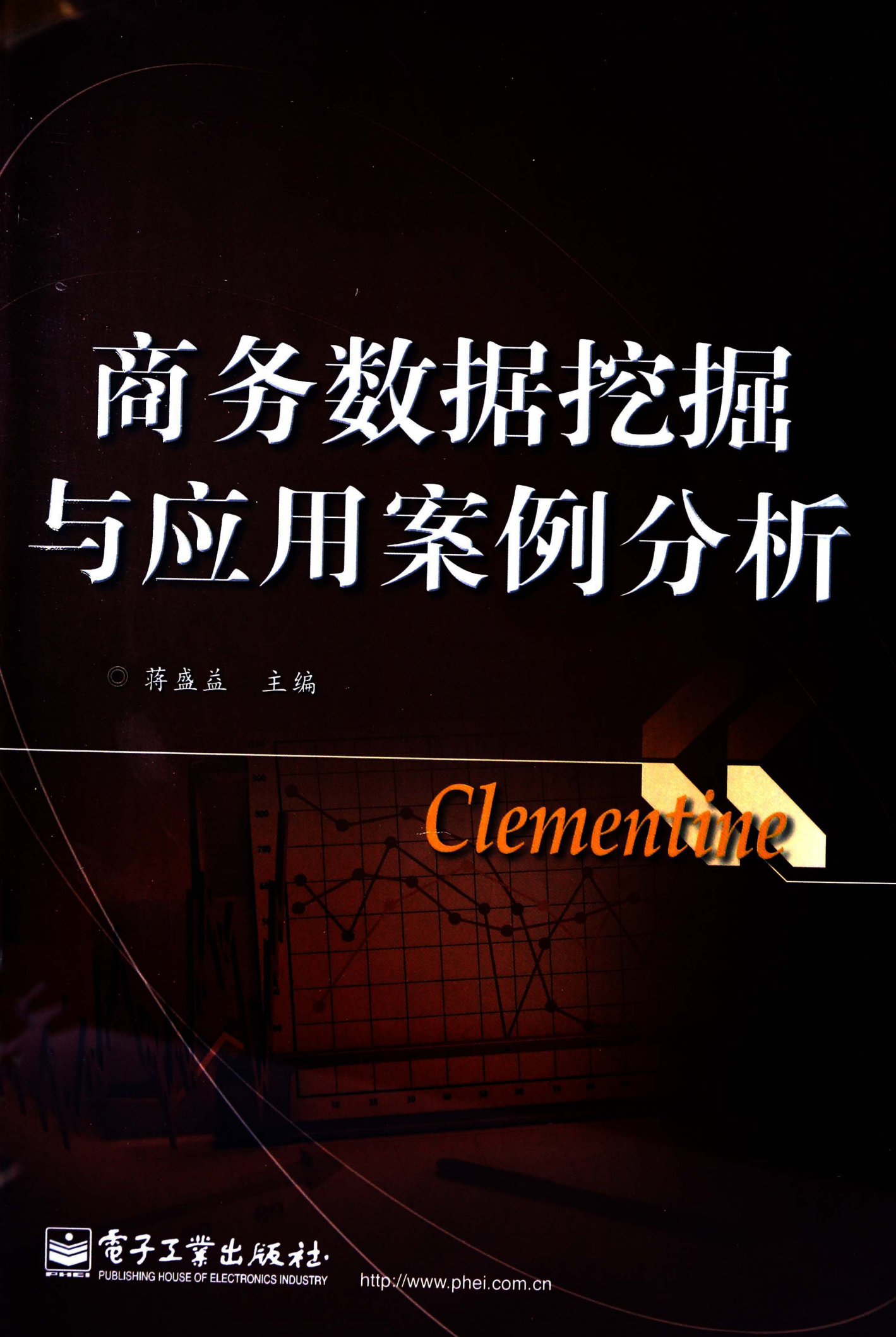
商务数据挖掘与应用案例分析

◎ 蒋盛益　主编



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

商务数据挖掘与应用案例分析

蒋盛益　主编

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

内 容 简 介

本书由认识篇、技术篇和案例篇三部分组成，以商业领域中的问题为背景，重点在于讲解数据挖掘技术的应用。认识篇从整体上介绍了数据挖掘的各种技术和数据挖掘建模过程，可使读者了解数据挖掘技术在商业领域中的应用概貌；技术篇介绍了数据挖掘中的聚类分析、分类、回归、关联规则挖掘、离群点检测等方法；案例篇展示了数据挖掘在 6 个不同行业中的应用案例，期望通过案例的分析使读者能够理解如何应用数据挖掘技术解决商业领域中的问题。

本书可作为经济、管理类等相关专业的学生学习数据挖掘技术的教材或参考书，也可作为计算机相关专业学生学习数据挖掘技术的参考书，还可作为企事业单位管理者、信息分析人员、市场营销人员和研究与开发人员的参考资料。

图书在版编目（CIP）数据

商务数据挖掘与应用案例分析 / 蒋盛益主编. —北京：电子工业出版社，2014.1
ISBN 978-7-121-22211-5

Ⅰ. ①商… Ⅱ. ①蒋… Ⅲ. ①商务—数据采集—高等学校—教材 Ⅳ. ①F715

中国版本图书馆 CIP 数据核字（2013）第 308904 号

策划编辑：章海涛　　文字编辑：任欢欢
责任编辑：郝黎明
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：787×1 092　1/16　印张：19.5　字数：499.2 千字
印　　次：2014 年 1 月第 1 次印刷
定　　价：42.00 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。

前　言

随着数据挖掘技术应用越来越广泛，越来越多的学校在经济、管理类专业逐步开设了数据挖掘课程，以适应社会对掌握数据挖掘技术人才的需求。

本书旨在介绍数据挖掘的基本原理、方法以及数据挖掘应用流程。通过案例的分析使读者能够应用这些方法解决商业领域中的问题。全书分为三部分，共15章。

上篇，认识篇。数据挖掘技术具有广泛的应用，认识篇从整体上观察、认识数据挖掘，使读者了解数据挖掘的各种技术，了解数据挖掘技术在商业领域中的应用概貌；熟悉数据挖掘建模过程。第 1 章介绍数据挖掘的基本概念以及数据挖掘在商业领域中的一些重要应用，第2章围绕“跨行业数据挖掘过程标准”CRISP-DM介绍数据挖掘过程的6个阶段。

中篇，技术篇。从应用的角度看，数据挖掘是一个工具，为了很好地应用数据挖掘，须知道什么时候应该使用何种数据挖掘技术，需要对数据挖掘主流算法有一定程度的了解。此外，还需要了解模型内部机制，这样才可以知道如何有效地准备建模所用的数据集，以及如何使用不同参数来改进模型的输出结果。数据挖掘所涉及的学科领域和方法很多，数据挖掘的常用技术有聚类分析、分类方法、关联分析、离群点检测和回归分析等，为了有效实施数据挖掘，需要对实际领域的数据进行适当的预处理。技术篇详细讲解了这些经典算法，同时对同一类数据挖掘技术的不同算法进行对比。本篇内容包括第3～8章，第3章介绍聚类分析的方法，第4章介绍分类方法，第5章介绍关联分析方法，第6章介绍离群点检测方法，第7章介绍回归分析方法，第8章介绍数据预处理方法。

下篇，案例篇。在一个特定的行业，进行数据分析时可能需要应用多种数据挖掘技术；同一种数据挖掘技术在应用到不同行业时，会存在一定的差异。案例篇通过 6 个不同行业中的案例展示数据挖掘技术在不同行业的应用。所有案例均采用 CRISP-DM 规范进行描述。本篇内容包括第9～15章，第9章介绍 Clementine 12.0 的基本用法；第10章介绍数据挖掘在通信行业中的客户细分、客户流失分析、客户社会关系挖掘、业务交叉销售等方面的应用，并通过实际案例进行了分析；第11章介绍数据挖掘在银行业中的应用，重点介绍数据挖掘在信用风险分析中的应用，并通过信用卡的欺诈检测进行分析；第12章介绍数据挖掘在目录直销业中的应用，

重点介绍 FRM 方法，并通过 Charles 读书俱乐部的案例进行了分析；第 13 章介绍数据挖掘在零售行业中的应用；第 14 章介绍数据挖掘在上市公司财务风险预警分析中的应用；第 15 章介绍数据挖掘在电子商务领域中的应用。

本书除了介绍数据挖掘的经典方法之外，也融入了作者的部分研究成果。

本书的出版融汇了许多人的辛勤劳动。全书由蒋盛益策划和统稿，李晓婷参与第 2 章的编写，吴美玲参与第 4、10 章的编写，王卉参与第 5 章的编写，张钰莎参与第 7、8 章的编写，殷倩参与第 9、11 章的编写，殷倩、张圣声参与第 12 章的编写，陈东沂、陈嘉华参与第 13 章的编写，彭金原参与第 14 章的编写，麦智凯、丘燕珍、陈东沂参与第 15 章的编写。参与编写工作的还有庞观松、王连喜、程一芳等。王家兵副教授、李霞副教授认真审阅了初稿，指出了文中存在的纰漏之处，并提出了修改建议。本书的出版也得到了电子工业出版社的大力支持，书中参考了许多学者的研究成果，在此一并表示衷心感谢。

限于作者学识水平，书中难免存在不足和疏漏，敬请读者批评指正。

作　者

目　录

上篇　认　识　篇

中篇　技　术　篇

下篇 案 例 篇

上篇 认识篇

第1章 绪 论

柏拉图曾说过“需要是发明之母”。近年来，数据挖掘引起了整个信息产业界和商界的极大关注，其主要原因是存在可以广泛使用的大量数据，并且迫切需要将这些数据转化成有用的信息和知识。获取的信息可以被广泛应用于市场营销、客户关系管理、欺诈检测、产品质量控制等领域。

数据挖掘是一种将传统的数据分析方法与处理大量数据的复杂算法相结合的技术。本章我们将概述数据挖掘相关概念，及数据挖掘在商业领域中的应用，并列举本书所涵盖的关键主题。数据挖掘是需要下工夫熟练掌握的一种技术，我们需要掌握数据挖掘的常识和基础，了解数据挖掘是什么，以及如何应用它。

1.1 引例

超市货架组织

在大型超市，我们会发现服务员经常整理货架。这不禁使我们产生这样的疑问：货架的组织对销售会产生影响吗？哪些商品放在一起比较好卖？哪些商品会受到货架组织的影响？哪些顾客的消费行为会受到货架组织的影响？

经过一段时间的观察和思考、调查和分析，我们会注意到：超市货架的组织方式会影响某些商品的销售，即会影响某些消费者的购买行为。“啤酒与尿布”的故事就是一个经典案例。

广告精准投放

随着商业竞争日益加剧，公司希望能最大限度地从广告中得到销售回报，希望能有更多的用户来响应他们的广告，所以他们就必须在投放广告之前做大量的市场分析工作。例如，根据自己的产品结合目标市场顾客的家庭收入、教育背景和消费趋向，分析出哪些地区的住户或居民最有可能响应公司的销售广告，购买自己的产品或成为客户，从而使广告只针对这些特定的客户群。这样有的放矢地筛选广告的投放市场既节省开销又提高了销售回报率。

随着 Web 2.0 应用的推广，网络社区服务 SNS（Social Network Service）已成为互联网关注的焦点。

SNS 通过网络服务、数据处理，不仅能够帮助人们找到朋友、合作伙伴，而且能够帮助人们实现个人社会关系管理、信息共享和知识分享，拓展其社交网络、达成更有价值的沟通和协作。基于网络社区独特的用户群和黏性服务，其强大的营销价值日益被发掘。通过挖掘网络中潜在的社区人群，企业可以更好地搜索潜在客户和传播对象，将分散的目标顾客和受众精准地聚集在一起，精确地把广告投放给目标客户。这不但可以有效降低单人营销费用，而且可以减少对非目标客户的干扰，提高广告的满意度，最终实现网络广告投放策略的真正价值。这一技术已被当当网等商务网站广泛使用。

客户流失分析

客户是企业生存的基础，在市场化程度高的行业，企业之间竞争激烈，为了获取更多的客户资源和占有更大的市场份额，企业往往采取名目繁多的促销活动和层出不穷的广告宣传来吸引新客户、留住老客户。研究发现：发展一个新客户比保持一个老客户的费用要高出 5 倍以上。针对这一问题，电信、银行、保险等行业都非常关注客户流失问题。以电信的客户流失为例，有研究者以 2009 年电信的流失客户历史消费行为数据、客户的基础信息、客户拥有的产品信息为基础，通过研究综合考虑流失的特点和与之相关的多种因素，从中发现流失客户的特征，以此建立可以在一定时间范围内预测客户流失倾向的预测模型，以便对流失进行预测，并对流失的后果进行评估，为电信公司有关部门提供有流失倾向的用户名单和这些用户的行为特征，以便制定恰当的营销策略，开展客户挽留工作，防止因客户流失而引发的经营危机，同时提升电信公司的竞争力。

智能搜索

在海量网络数据中，用户试图通过网络来快速发现有用信息变得非常困难，如何提高信息获取的效率成为研究人员广泛关注的课题。Web 信息检索，即搜索引擎，是有效解决这一问题的重要工具。传统的搜索引擎，在用户输入关键词进行查询后，返回的是成千上万的相关结果，这往往导致用户需要花费大量的时间来浏览与选择，因此不能满足用户快速获取信息的愿望。另外，对于同一搜索引擎使用相同关键词进行搜索时，不同人得到的返回结果是相同的，然而不同的人期望的或关注的结果是不同的。如提交查询词“苹果”的两个人可能希望看到不同类型的信息，可能一个对水果的相关产品信息有兴趣，而另一个则倾向于获取电子产品的相关信息。因此大量研究人员开始研究行业化、个性化、智能化的第三代搜索引擎。例如，通过跨语言信息检索可以方便地检索出不同语种的网络资源；通过文本聚类算法可对搜索返回结果进行分组处理，这样用户可以根据聚类结果快速定位到所需的资源上；通过显式或隐式地收集用户偏好信息，深层次地挖掘用户个人兴趣，以便为用户提供个性化的搜索和查询服务；通过交互的查询扩展功能改善用户查询用词，同时也可使系统能更好地理解用户的检索意图。

免费用户到付费用户的转化

在网络游戏试玩初期，游戏运营商为了测试和完善网络游戏以及快速扩大玩家群，通常都会推出一段相对较长的免费试玩期。因此，在网络游戏正式运营前就会存在大量的注册用户，这些注册用户会在网络游戏运行后存在很长一段时间。如何把这些注册用户转化成付费客户，真正为游戏运营商带来收益呢？如何对注册用户采取差别化营销手段，从而提高营销活动效果，使运营商利润得到最大化？

上述例子来自不同的商业应用领域，但背后都以数据挖掘为核心处理技术，即利用数据挖掘技术发现隐藏的规律，为领域的决策提供支持。

1.2 数据挖掘产生的背景及概念

1.2.1 数据挖掘产生的背景

随着通信、计算机和网络技术的快速发展，以及日常生活自动化技术的普遍使用，如超市 POS 机、自动售货机、信用卡和借记卡、在线购物、自动订单处理、电子售票、RFID、客服中心等，数据正以空前的速度产生和被收集。包括通信、银行、交通、零售商等在内的一些企业，已经与客户建立了自动化的交互关系，生成大量交易记录。在各行各业，许多公司已经开始认识到客户对业务非常重要，客户信息是它们的宝贵财富。

对于从事服务业的公司来说，信息意味着竞争优势，信息就是产品。很多公司发现，它们拥有的有关客户的某些信息不仅对自己非常有用，对别人也非常有用。信用卡公司也有航空公司想知道的信息，即谁购买了大量的机票，在这里信用卡公司处在信息经纪人（中间人）的位置。信用卡公司可以针对经常乘坐飞机的人们进行促销，吸引以前坐其他航空公司飞机的人。

大量信息在给人们带来方便的同时也带来了一大堆问题：信息冗余；信息真假难以辨识；信息安全难以保证；信息形式不一，难以统一处理等。

随着信息技术的高速发展，数据库应用的规模、范围和深度不断扩大，互联网已成为信息传播的主流平台。“数据过剩”、“信息爆炸”与“知识贫乏”等现象相继产生，人们淹没在数据中而难以快速制定合适的决策! 在强大的商业需求驱动下，商家开始注意到有效地解决大容量数据的利用问题具有巨大的商机；学者们开始思考如何从大容量数据集中获取有用信息和知识。然而，面对高维、复杂、异构的海量数据，提取潜在的有用信息就成为巨大的挑战。面对这一挑战，数据挖掘技术应运而生，并显示出强大的生命力。

数据丰富加上对强有力的数据分析工具的需求这种现象可描述为数据丰富、信息贫乏。快速增长的海量数据、存放在大型和大量数据存储库中，没有强有力的工具从而使理解它们已经远远超出了人的能力。利用数据挖掘工具进行数据分析，可以发现重要的数据模式，这对商务策略、知识库、科学有巨大贡献。数据挖掘的迅速发展，使商业受益匪浅，如市场营销组织应用客户细分来识别那些对不同形式营销传媒敏感的客户群，许多公司应用数据挖掘技术来识别高价值客户，从而为他们提供所需的服务以留住他们。

案例 1-1：梅隆银行的数据挖掘

美国的梅隆银行（Bank of New York Mellon）在 1997 年设定了争取 20 万新户头的目标，为此，计划向 1000 万可能的顾客邮寄邀请函。然而，这家银行却利用了数据挖掘技术产生了 3000 个最可能的顾客的模式。对这些模式的子集再加以精选，产生了更小的一个数目。测试表明，这个更小的数目会产生 12%的回应率。这个回应率使得这家银行只需发出 200 万份邀请函即可获得他们想要的 20 万名顾客，而不是原定的向 1000 万人发出信函。因此，利用数据挖掘技术除了削减成本之外，还提高了每位新开户的顾客的平均利润率，其利润要比通常高 3 倍，因为该技术瞄准了那些需求最适合梅隆银行服务项目的顾客。

这个例子说明了数据挖掘的两个重要方面：第一个方面仅就其规模而言，牵涉到的数据量和所探索的模式数目要比传统的数据分析量大得多；第二个方面就是，即使是受过高级培训的专家也能获益于数据挖掘。正如我们在梅隆银行的例子中所见，一个外部专家小组得出的结果比本公司专职数据分析部门用常规方法得出的结果高 6 倍，而所花时间只是后者的四分之一。我们的一个主要目标就是使

数据挖掘工具使用简易，即可以使最终的商务用户，而不止专家，都会使用它们。

资料来源：http://www.tianyabook.com/zhexue/weilai/014.htm

案例 1-2：Yahoo 的数据挖掘

Usama Fayyad 博士是 Yahoo 的首席数据官，KDnuggets 的 Gregory 对他进行了访谈。下面是访谈中介绍的 Yahoo 在数据挖掘方面的成功案例。

（1）产品整合：一个例子就是你今天在 Yahoo 电子邮箱上看到的数据挖掘的可视结果。通过对用户使用行为的意外模式分析，我们发现在每次会话中，人们阅读邮件和阅读新闻的行为之间存在很强的相关关系。我们把这个发现传达给 Yahoo 电子邮箱产品小组，他们首先想到的就是验证这种关系的影响：在一组测试用户的邮箱首页上显示一个新闻模块，其中的新闻标题被醒目显示。

对于像电子邮箱这种产品，最头痛的问题就是如何获取新的“轻量级用户”，并推动他们的用量，使之变成“重量级用户”。如果你做到了，那么流失率就会显著下降。实际上，在我们的试验中，最弱的一组流失率下降了 40%。于是 Yahoo 立刻开发并完善了新闻模块，并嵌入 Yahoo 电子邮箱的首页，现在，上亿的消费者都可以看到并使用这种产品。我喜欢提及这个故事，因为它很好地说明了我们产品团队的及时反应能力，也证明了在用户使用行为数据中蕴含着很多极具价值的潜在模式。

（2）即时通信：我们对雅虎通（Instant Messenger）的使用情况进行了分析，以了解激励用量的关键因素是什么。结果发现，最重要的因素是让用户扩大他们的“好友列表”，至少增加 5 个新的好友。据此 Yahoo 精心设计了相应的营销活动，鼓励用户增加好友列表中的好友数，从而显著激励了雅虎通的用量。

（3）Yahoo 首页的搜索框：一个简单的例子就是我们发现，在 Yahoo 的首页上，把搜索框放在居中的位置（而不是以前的左侧）将提高用户的用量。这样一方面可以促进用户的积极使用，对 Yahoo 来说也没有额外成本支出。这个结果的发现过程也很有趣，我们首先发现 Netscape 浏览器的用户比 IE 的用户更多地使用了搜索功能，进一步探查发现两个浏览器在视觉上的唯一区别就是：二者中的搜索框位置不同！搜索框在 Netscape 浏览器中是居中放置的，而在 IE 中则是靠近左侧。很不明显的差别，但却很重要。一般谁会想到呢？

资料来源：http://blogger.org.cn/blog/more.asp?name=idmer&id=9729

1.2.2 数据挖掘概念

数据挖掘可以从技术和商业两个层面上来定义：从技术层面上看，数据挖掘是探查和分析大量数据以发现有意义的模式和规则的过程。从商业层面看，数据挖掘就是一种商业信息处理技术，其主要特点是对大量业务数据进行抽取、转换、分析和建模处理，从中提取辅助商业决策的关键性数据。

数据挖掘与传统数据分析方法（如查询、报表、联机应用分析等）有着本质区别：数据挖掘是在没有明确假设的前提下去挖掘信息和发现知识。数据挖掘所得到的信息具有先前未知、有效和实用三个特征。先前未知的信息是指该信息是事先未曾预料到的，即数据挖掘是要发现那些不能靠直觉或是经验而发现的信息或知识，甚至是违背直觉的信息或知识。挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最经典的例子是“尿布和啤酒”的故事——尿布和啤酒之间销售关联的发现。

数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询提升到从数据中挖掘知识，提供决策支持。在市场对人才需求的引导下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，他们投身到数据挖掘这一新兴的研究领域，使其形成新的技术热点。数据挖掘是需要下工夫熟练掌握的一种技术，我们需要掌握数据挖掘的

常识和基础，了解数据挖掘是什么，以及如何应用它。

1.3 数据挖掘任务及过程

1.3.1 数据挖掘任务

通常，数据挖掘任务可以分为预测型任务和描述型任务。预测型任务就是根据其他属性的值预测特定属性的值，如回归、分类、离群点检测。描述型任务就是寻找概括数据中潜在联系的模式，如聚类分析、关联分析、序列模式挖掘。

（1）聚类（Clustering）分析

“物以类聚，人以群分”。聚类分析技术试图找出数据集中数据的共性和差异，并将具有共性的对象聚合在相应的簇中。聚类分析可以帮助判断哪些组合更有意义，聚类分析已广泛应用于客户细分、定向营销、信息检索等领域。

（2）分类（Classification）分析

分类分析就是通过分析示例数据库中的数据，为每个类别做出准确的描述，或建立分析模型，或挖掘出分类规则，然后用这个分类模型或规则对数据库中的其他记录进行分类。分类分析已广泛应用于用户行为分析（受众分析）、风险分析、生物科学等领域。

（3）关联（Association）分析

关联分析就是发现特征之间的相互依赖关系，通常是在给定的数据集中发现频繁出现的模式知识（又称为关联规则）。关联分析广泛应用于市场营销、事务分析等领域。

（4）离群点（Outlier）检测

离群点检测就是发现与众不同的数据。离群点检测已广泛应用于（商业、金融、保险等领域）欺诈行为的检测，网络入侵检测，反洗钱、犯罪嫌疑人调查，海关、税务稽查等领域。

（5）回归（Regression）分析

回归分析是确定一个变量与一个或多个变量间相互依赖的定量关系的分析方法，常应用于风险分析、销售预测等领域。

（6）序列模式（Sequential Pattern）挖掘

序列模式挖掘是指分析数据间的前后序列关系，包括相似模式发现、周期模式发现等，应用于客户购买行为模式预测、Web 访问模式预测、疾病诊断、网络入侵检测等领域。

1.3.2 数据挖掘过程

从商业应用的角度看，数据挖掘过程就是从数据到决策的过程，主要包括三个步骤：首先是数据准备，然后利用数据挖掘相关方法提取出有用的知识，最后以提取出来的知识来辅助相应决策者进行决策。数据挖掘过程如图 1-1 所示。

（1）数据准备：包括数据收集（集成）和预处理。数据收集看似容易且不引人注意，但它却是数据挖掘的基础。知识是从海量的数据里提取出来的，要挖掘知识必须得收集一定量的数据。收集到的原始数据通常会存在缺失值、错误值、不一致值等问题，许多数据与数据挖掘的目标无关，不能直接用作知识提取的数据源，需要进行数据预处理。

（2）知识提取：基于预处理后的数据，使用各种数据挖掘方法（如分类、聚类、关联分析等）进行知识提取，这是数据挖掘的核心部分。

（3）知识辅助决策：将提取出来的知识提供给决策者以辅助制定相应决策。

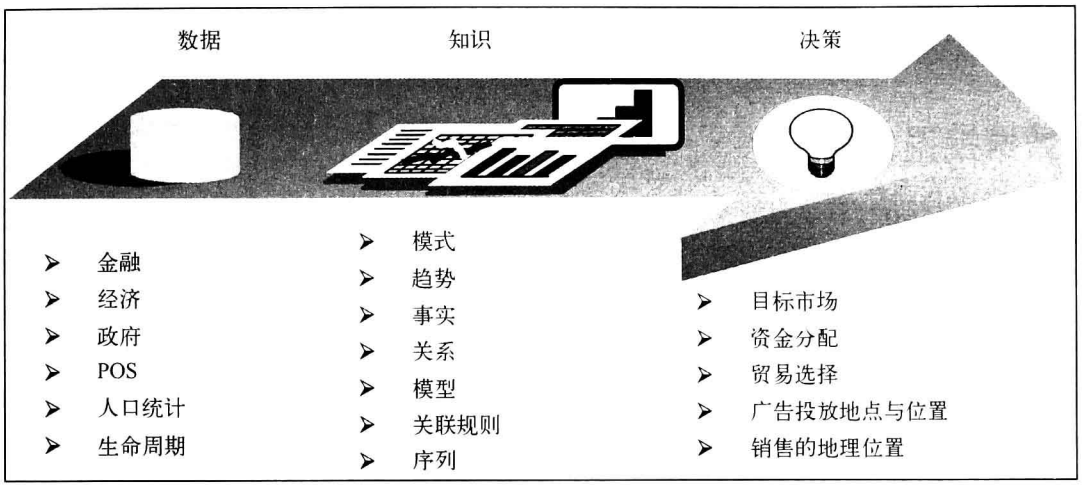


图 1-1　数据挖掘过程

1.4　数据挖掘常用软件简介

数据挖掘产品有很多，比较著名的商用数据挖掘软件有 SPSS Clementine、SAS Enterprise Miner 、IBM Intelligent Miner 、SQL Server 2005 Data Mining 和 Oracle DM 等，它们都能够提供常规的挖掘过程和挖掘模式。Matlab、Excel（Data mining in Excel: XLMiner）、R 语言等提供了数据挖掘模块。开源数据挖掘工具有 Weka、RapidMiner（YALE）、ARMiner 以及 AlphaMiner 等。

（1）商用软件

● SPSS Clementine

Clementine 是 ISL（Integral Solutions Limited）公司开发的数据挖掘工具平台。1999 年 SPSS 公司收购了 ISL 公司，对 Clementine 产品进行重新整合和开发，2009 年 10 月，IBM 收购了 SPSS Inc。作为一个数据挖掘平台，Clementine 结合商业技术可以快速建立模型，进而应用到商业活动中，帮助人们改进决策制定过程。强大的数据挖掘功能和显著的投资回报率使得 Clementine 在业界久负盛誉。Clementine 拥有功能强大的数据挖掘算法，将数据挖掘贯穿业务流程的始终，在缩短投资回报周期的同时极大地提高了投资回报率。来自 KDnuggets （http://www.kdnuggets.com/polls / ）的调查报告显示：Clementine（2000—2009）曾有 9 年摘获数据挖掘产品用户数排行榜桂冠。本书中实验所用环境主要是 Clementine。

● SAS/Enterprise Miner

SAS/Enterprise Miner 是数据挖掘产品市场上一个强劲的竞争者。它支持 SAS 统计模块，还通过大量数据挖掘算法增强了那些模块。SAS 使用它自身的 SEMMA 方法来提供一个能支持包括关联、聚类、决策树、神经网络和统计回归在内的数据挖掘工具。SAS Enterprise Miner 既方便初学者使用（可视化操作），也能为有编程经验的用户使用（高效的编程）。它的 GUI 界面是由数据流驱动的，所以易于理解和使用，同时，允许分析者通过构造一个使用链接连接数据结点和处理结点的可视数据流图建造一个模型，并且允许把处理结点直接插入到数据流中。由于支持多种模型，Enterprise Miner 允许用户比较（评估）不同模型并利用评估结点选择最适合的。另外，Enterprise Miner 提供了一个能产生被任何 SAS 应用程序所

访问的评分模型的评分结点。

● Microsoft SQL Server 2005 Data Mining

Microsoft SQL Server 2005 Data Mining 属于商业智能技术，它可帮助用户构建复杂的分析模型，并使其与业务操作相集成。Microsoft SQL Server 2005 分析服务中构建了新的数据挖掘平台——一个易于使用、可扩展、方便访问、非常灵活的平台。对于以前从未考虑过采用数据挖掘的组织机构，这无疑是个非常容易接受的解决方案。

（2）开源软件

● Weka

Weka 的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），它是基于 JAVA 环境下的开源机器学习与数据挖掘软件，可在其官方网站上进行下载。

Weka 作为数据挖掘平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。开发者可使用 Java 语言，在 Weka 的架构下开发出更多的数据挖掘算法。

Weka 适合学习研究用，但不适合商用。AlphaMiner 是一款基于 Weka 内核开发的处理能力更强的商用软件（http://bi.hitsz.edu.cn/AlphaMiner/index.htm）。

● RapidMiner

RapidMiner（前身是 YALE）是基于 Weka 构建的一款开源数据挖掘软件，它不仅提供了一个 GUI 的数据处理和分析环境，还提供了 Java API 以便将它的能力嵌入到其他应用程序。其数据挖掘任务涉及范围广泛，能简化数据挖掘过程的设计和评价。

1.5 数据挖掘在商业领域中的应用

在银行、保险、电信、零售等行业，由于高度竞争引发了对数据挖掘的广泛应用。数据挖掘商业应用的目标是公司通过对客户的更多了解来改善其市场、销售和客户支持运作。在市场经济环境下，任何有远见的公司应努力了解每个客户，通过对客户的了解，采取措施促使客户选择与它们进行商业活动，而不是选择它们的竞争对手。通过对客户的了解，学习认识每个客户的价值，进而知道哪些人值得投入资金和人力来保持联系，哪些人可以放弃。为此，公司需要做到：

（1）注意客户正在做什么；

（2）记住公司及其客户曾经做过什么；

（3）学习、挖掘客户与公司交易过程中留下的信息；

（4）按照获得的知识进行商业活动使顾客更加受益。

本书的主要目标是上述第三个方面，即从公司的历史交易记录中挖掘有用的规则与规律。公司的历史数据中包含着客户的身份定位、兴趣、购买规律等对公司未来商业决策有用的信息，其能为高层管理员提供强有力的决策辅助。而这将是一个双赢的结果，既能为客户提供更高质量的服务（如为客户提供个性化的服务），同时，也能降低公司的成本或者提高收入（如公司广告可以针对感兴趣的对象投放）而产生更多利润。

客户对公司的忠诚度怎样？谁可能流失？哪种电话销售方式最适合某个客户？什么产品应该以何种定位面世？是什么决定某个客户是否对某种产品做出回应？某个客户需要的下一种产品或者服务是什么？下一个分支机构应该设置在哪里？类似这些问题的答案就隐藏在公司运营数据中，数据挖掘能为这些问题答案的寻找提供有力支持。