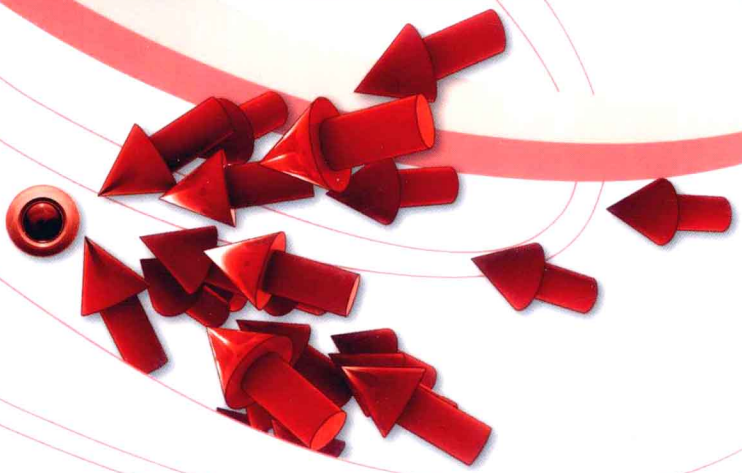
精品课程教材 省



高等学校计算机规划教材

数据挖掘原理与实践

蒋盛益 李霞 郑琪 编著

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

广东省精品课程主教材

高等学校计算机规划教材

数据挖掘原理与实践

蒋盛益　李　霞　郑　琪　编著

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书分为数据挖掘理论和数据挖掘实践两大部分。数据挖掘理论部分的主要内容包括数据挖掘的基本概念、数据挖掘的预处理、聚类分析、分类与回归、关联规则挖掘、离群点检测。数据挖掘实践部分讨论了数据挖掘在通信行业、文本挖掘等方面的实际应用；通过四个案例展示了在通信行业中如何利用数据挖掘进行客户细分、客户流失分析、客户社会关系挖掘、业务交叉销售；通过跨语言智能学术搜索系统和基于内容的垃圾邮件识别两个案例展示了数据挖掘在文本挖掘方面的应用。

本书可作为高等院校计算机、电子商务、信息科学等相关专业的教材或参考书，也可供从事数据挖掘研究、设计等工作的科研、技术人员参考。

图书在版编目（CIP）数据

数据挖掘原理与实践 / 蒋盛益，李霞，郑琪编著. —北京：电子工业出版社，2011.8
高等学校计算机规划教材
ISBN 978-7-121-14050-1
Ⅰ. ①数… Ⅱ. ①蒋…②李…③郑… Ⅲ. ①数据采集－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2011）第 134200 号

策划编辑：章海涛
责任编辑：章海涛　　特约编辑：曹剑锋
印　　刷：涿州市京南印刷厂
装　　订：涿州市桃园装订有限公司
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：787×1092　1/16　印张：17.75　字数：500 千字
印　　次：2011 年 8 月第 1 次印刷
印　　数：3 000 册　　定价：32.00 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

数据挖掘技术应用越来越广泛，社会对掌握数据挖掘技术的人才需求越来越大，越来越多的高校在计算机相关专业及经济、管理类专业开设了数据挖掘课程，以适应社会的需求。

本书旨在向读者介绍数据挖掘的基本原理、方法，数据挖掘应用流程，通过原理、方法应用的背景介绍，使读者理解、掌握如何选择数据挖掘方法解决实际问题，通过案例的分析使读者能够应用这些方法解决现实世界中的问题。

全书分为上、下两篇，共 8 章。上篇包括第 1～6 章，下篇包括第 7～8 章。

第 1 章介绍数据挖掘的基本概念以及数据挖掘的重要应用领域。

第 2 章介绍数据的基本统计量以及数据预处理的常用方法。

第 3 章介绍分类的基本概念、应用背景，重点介绍决策树、贝叶斯、最近邻分类方法。

第 4 章介绍聚类分析的基本概念、应用背景，重点介绍常用的聚类方法。

第 5 章介绍关联分析的基本概念、应用背景，重点介绍频繁模式挖掘算法（Apriori 算法和 Fp-growth 算法）、序列模式挖掘算法。

第 6 章介绍离群点挖掘的基本概念、应用背景，重点介绍基于距离、基于相对密度、基于聚类的离群点挖掘方法。

第 7 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 8 章介绍数据挖掘在文本处理方面的应用，介绍文本挖掘和 Web 挖掘的基本概念，通过跨语言智能学术搜索系统和基于内容的垃圾邮件识别两个案例进行分析。

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。

本书为广东省精品课程建设成果。

本书的出版融会了许多人的辛勤劳动。第 1、2、4、6、7、8 章由蒋盛益负责，第 3 章由李霞负责，第 5 章由郑琪负责。参与编写工作的还有庞观松、王连喜、吴美玲、谢照青、阳垚、苗邦、谢植林、邝丽敏等。印鉴教授、王家兵副教授认真审阅了初稿，指出了一些纰漏，并提出了修改建议。本书的出版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，书中肯定存在不足和疏漏，敬请读者批评指正。

本书为任课教师提供配套的教学资源（包含电子教案、实验用数据集、习题及参考答案、部分综述文献和常用资源列表），需要者可登录华信教育资源网（http://www.hxedu.com.cn），注册之后进行下载。

读者反馈：unicode@phei.com.cn。

作　者

反侵权盗版声明

目　录

上篇　原理篇

下篇 实践篇

上篇　原理篇

第1章 绪 论

数据收集与数据存储技术的快速发展，使得各种组织机构积累了海量数据。如何从这些海量数据中提取有价值的信息以辅助决策，成为巨大的挑战。面对这种挑战，一种数据处理的新技术——数据挖掘（Data Mining）应运而生。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。本章将概述数据挖掘，并列举本书所涵盖的关键主题。

引例

啤酒与尿布的故事

在一家超市，人们发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然摆在一起。但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不是一个笑话，而是一直被商家所津津乐道的发生在美国沃尔玛连锁超市的真实案例。原来，美国的妇女通常在家照顾孩子，所以她们经常会嘱咐丈夫在下班回家的路上为孩子买尿布，而丈夫在买尿布的同时又会顺手购买自己爱喝的啤酒。这个发现为商家带来了可观的利润。

这个故事是营销界的神话。“啤酒”和“尿布”两个看上去没有直接关系的商品摆放在一起进行销售，并获得了很好的销售收益，这种现象就是卖场中商品之间的关联性。研究“啤酒与尿布”关联的方法就是购物篮分析，购物篮分析可以帮助零售商在销售过程中找到具有销售关联的商品，并以此指导货架的组织，促进销售收益的增长！

广告精准投放

随着 Web 2.0 应用的推广，SNS（Social Network Service，网络社区服务）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，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。所谓客户流失，是指客户终止与企业的服务合同或转向其他同类企业提供的服务，在市场基本饱和的情况下，对老客户的保留将直接关系到企业的利益，客户流失将对企业的经营产生深远影响。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。客户流失分析是以客户的历史消费行为数据、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，从中发现与流失密切相关的特征和流失客户的特征，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，为相关业务部门提供有流失倾向的用户名单和这些用户的行为特征，以便相关部门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，提升公司的竞争力。

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和选择，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，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，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，同时可使系统能更好地理解用户的检索意图。

入侵检测

随着互联网的发展，各种网络入侵和攻击工具、手段也随着出现，使得入侵检测成为网络管理的重要组成部分。入侵可以定义为任何威胁网络资源（如用户账号、文件系统、系统内核等）的完整性、机密性和可用性的行为。目前，大多数商业入侵检测系统主要使用误用检测策略，这种策略对已知类型的攻击通过规则可以较好地检测，但对新的未知攻击或已知攻击的变种则难以检测。新的网络攻击或已知攻击的变种可以通过异常检测方法来发现，异常检测通过构建正常网络行为模型（称为特征描述），来检测与特征描述严重偏离的新的模式。这种偏离可能代表真正的入侵，或者仅是需要加入特征描述的新行为。异常检测主要的优势是可以检测到以前未观测到的新入侵。与传统的入侵检测系统相比，基于数据挖掘的入侵检测系统通常更精确，需要更少的专家的手工处理。

上述例子来自不同应用领域，但背后都以数据挖掘为核心处理技术，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。

1.1 数据挖掘产生的背景

四种技术激发了人们对数据挖掘技术的开发、应用和研究的兴趣：① 超大规模数据库的出现，如商业数据仓库和计算机自动收集数据记录手段的普及；② 先进的计算机技术，如更快和更大的计算能力和并行体系结构；③ 对海量数据的快速访问，如分布式数据存储系统的应用；④ 统计方法在数据处理领域应用的不断深入。

近年来，计算机软件和硬件技术快速发展，互联网用户急剧增加，社会已进入网络化时代。在网络化时代背景下，通信、计算机和网络技术正改变着整个人类和社会。如果用芯片集成度来衡量微电子技术，用 CPU 处理速率来衡量计算机技术，用信道传输速率来衡量通信技术，摩尔定律告诉我们，它们都是以每 18 个月翻一番的速率在增长，这一势头已经维持了十多年。在美国，广播用户达到 5000 万户用了 38 年，电视用户用了 13 年，Internet 拨号上网

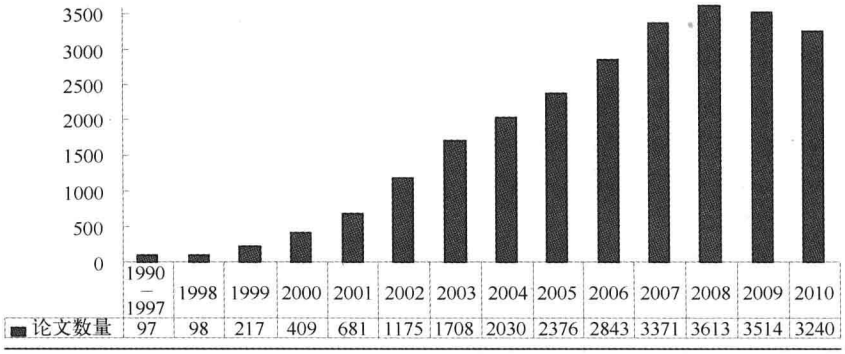
达到 5000 万户仅用了 4 年。全球 IP 网发展速度达到每 6 个月翻一番，国内情况亦然。《纽约时报》由 20 世纪 60 年代的 10～20 版扩张至现在的 100～200 版，最高曾达 1572 版，《北京青年报》也已是 16～40 版，《市场营销报》已达 100 版。然而在现实社会中，人均日阅读时间通常为 30～45 分钟，只能浏览一份 24 版的报纸。大量信息在给人们带来方便的同时也带来了一大堆问题：信息冗余、信息真假难以辨识、信息安全难以保证、信息形式不一、难以统一处理等。

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”和“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，有效地解决海量数据的利用问题具有巨大商机，学者们开始思考如何从海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成为巨大挑战。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。

数据挖掘思想来自于机器学习、模式识别、统计和数据库系统。数据挖掘概念首次出现在 1989 年举行的第十一届国际联合人工智能学术会议上。目前有许多数据挖掘方面的国际会议，如 ACM SIGKDD（ACM's Special Interest Group on Knowledge Discovery and Data Mining）、ACM SIGMOD（ACM's Special Interest Group on Management Of Data）、CIKM（ACM Conference on Information and Knowledge Management）、ICDM（IEEE International Conference on Data Mining）、ECML PKDD（European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases）、PAKDD（Pacific-Asia Conference on Knowledge Discovery and Data Mining）、ICDE（IEEE International Conference on Data Engineering）、VLDB（Very Large Data Base）、ADMA（International Conference on Advanced Data Mining and Applications）、SDM（SIAM Conference on Data Mining）、ICMLC（International Conference on Machine Learning and Computing）。在数据挖掘的发展历程中，其研究重点从最初的侧重发现方法转向侧重系统应用，注重多种发现策略和技术的集成，注重学科间的相互渗透。此外，在 Internet 上还有不少 KDD（Knowledge Discovery in Database，知识发现）电子出版物和自由论坛，如国际权威半月刊 Knowledge Discovery Nuggets（http://www.kdnuggets.com/subscribe.html）、国内的数据挖掘研究院（中科院）http://www.dmresearch.net 和中国商业智能网 http://www.chinabi.net。

国内对数据挖掘的研究起步较晚，1993 年国家自然科学基金首次支持该领域的研究。此后，国家、各省自然科学基金委，国家社科基金，“863”、“963”项目，国家、各省的科技计划，每年都有相关项目支持。众多研究机构和大学都成立有专门的项目组。从事数据挖掘研究与应用的人员越来越多，在中国期刊全文数据库 CNKI 中检索主题词“数据挖掘”得到的各年度论文数如图 1-1 所示。这表明最近十多年数据挖掘经历了快速发展期，2008 年达到了顶峰，数据挖掘的基本理论问题逐步得到了解决，现在更多的是数据挖掘的应用。

在国内召开的许多信息技术学术会议中，数据挖掘也是非常重要的主题，如中国机器学习会议 CCML（China Conference on Machine Learning）、全国数据库学术会议、中国数据挖掘会议 CCDM（China Conference on Data Mining）、全国搜索引擎和网上信息挖掘学术研讨会 SEWM（Symposium of Search Engine and Web Mining）。



	1990—1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
论文数量	97	98	217	409	681	1175	1708	2030	2376	2843	3371	3613	3514	3240

图 1-1　国内学术期刊网中检索主题词“数据挖掘”得到的年度论文数

1.2　数据挖掘任务及过程

1.2.1　数据挖掘定义

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，数据挖掘就是从大量数据中提取有用信息的过程。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。挖掘出的信息越出乎意料，就可能越有价值。在商业应用中最典型的例子是“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，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，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。

1.2.2　数据挖掘对象

从应用领域的角度看，数据挖掘对象主要包括以下几大类型。

（1）关系数据库

关系数据库是建立在关系数据库模型基础上的数据库，借助于集合代数等概念和方法来处理数据库中的数据。关系数据库可以通过数据库查询、获取信息，当数据挖掘应用于关系数据库时，可以进一步搜索趋势或数据模式。关系数据库广泛应用于各行各业，是数据挖掘最常见、最丰富的数据源。

（2）数据仓库（Data Warehouse）

数据仓库是一个从多个数据源收集的信息存储库，存放在一个一致的模式下。数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化的（Time Variant）数据集合，用于支持管理决策（Decision Making Support），适合于

联机分析处理（On-Line Analysis Processing，OLAP）。银行、电信等行业，数据集中后通常需要保存在数据仓库中。

（3）事务数据库

在事务数据库中，每个记录代表一个事务。通常，一个事务包含唯一的事务标识号和组成该事务的项的列表（如在超市中购买的商品）。超市的销售数据是典型的事务型数据。事务数据库可能有一些与之关联的附加表，如包含关于销售的其他信息：事务的日期、顾客的编号、销售者的编号、连锁分店的编号等。

（4）空间数据库（Spatial Database）

空间数据库是指在关系数据库内部对地理信息进行物理存储。空间数据库中存储的海量数据包括对象的空间拓扑特征、非空间属性特征、对象在时间上的状态变化。常见的空间数据库的数据类型包括地理信息系统、遥感图像数据医学图像数据。空间数据库的特点有：数据量庞大，空间数据模型复杂，属性数据和空间数据联合管理，应用范围广泛。

（5）时态数据库和时间序列数据库（Temporal Database and Time-Series Database）

时态数据库和时间序列数据库都存放与时间有关的数据。时态数据库通常存放与时间相关的属性值，如与时间相关的职务、工资等个人信息及个人简历信息等。时间序列数据库存放随时间变化的值序列，如零售行业的产品销售数据、股票数据、气象观测数据等。时态数据库和时间序列数据库的数据挖掘研究事物发生、发展的过程，有助于揭示事物发展的本质规律，可以发现数据对象的演变特征或对象变化趋势。

（6）流数据（Stream Data）

与传统数据库中的静态数据不同，流数据是连续的、有序的、变化的、快速的、大量的输入数据，主要应用场合包括网络监控、网页点击流、股票市场、流媒体等。与传统数据库相比，流数据在存储、查询、访问、实时性的要求等方面都有很大区别。流数据具有以下特点：数据实时到达；数据到达次序独立，不受应用系统控制；数据规模宏大且不能预知其最大值；数据一经处理，除非特意保存，否则不能被再次取出处理，或者再次提取数据的代价昂贵。

（7）多媒体数据库（Multimedia Database）

多媒体数据库是数据库技术与多媒体技术相结合的产物。多媒体数据库不是对现有的数据进行界面上的包装，而是从多媒体数据和信息本身的特性出发。多媒体数据库用计算机管理庞大复杂的多媒体数据，主要包括图形（graphics）、图像（image）、音频（audio）、视频（video）等，现代数据库技术一般将这些多媒体数据以二进制大对象的形式进行存储。多媒体数据库的数据挖掘需要将存储和检索技术相结合，处理方式不同于数值、文本数据的处理。目前，对多媒体数据的挖掘包括构造多媒体数据立方体、多媒体数据的特征提取和基于相似性的模式匹配等。

（8）文本数据库（Text Database）

文本数据库是一种常用的数据库之一，也是最简单的数据库。任何文件都可以存入文本数据库。文本数据库存储的是对对象的文字性描述。文本数据类型包括：无结构类型（大部分的文本资料和网页）、半结构类型（XML 数据）、结构类型（图书馆数据）——对应于通常的关系型数据库。文本数据的处理广泛应用于办公资料的处理，如法院、检察院的案件资料的处理。文本数据库存在以下缺点：一是并发访问麻烦，无法实现多个程序同时修改数据库里面的不同记录；二是查询、修改、删除非常麻烦，只能顺序查找，修改、删除需要更新整个文件。文本数据库的优点显而易见：程序简单，数据库管理方便。

（9）万维网数据

万维网（Word Wide Web，WWW）被看成是最大的文本数据库。随着 Internet 的广泛使用，