高等院校计算机教育系列教材

数据挖掘实践教程

吴思远 主 编
邹 洋 黄梅根 贾 玲 副主编

- 结构清晰，知识完整
- 入门快速，易教易学
- 实例丰富，实用性强
- 学以致用，注重能力

赠送实验数据集、电子课件和课后习题答案

清华大学出版社

高等院校计算机教育系列教材

数据挖掘实践教程

吴思远　主　编

邹　洋　黄梅根　贾　玲　副主编

清华大学出版社

北 京

内 容 简 介

本书注重数据挖掘理论，将理论与实践相结合、知识理论与具体实现方法相结合，由浅入深地介绍了数据分析与挖掘的相关知识。全书分为3部分。第1部分介绍了数据挖掘理论(第1~3章)，第2部分介绍了Excel 2010数据分析与挖掘、SQL Server 2012数据挖掘、SPSS数据分析与挖掘的实践过程(第4~9章)，第3部分介绍了SQL Server和SPSS数据挖掘的实验内容(第10章)。

本书为教师提供了配套的教学资源，可以作为计算机、智能科学类专业本科生的数据挖掘课程教材，也可以作为专业技术人员的自学参考书及数据挖掘爱好者的自学用书。

图书在版编目(CIP)数据

数据挖掘实践教程/吴思远主编. —北京：清华大学出版社，2017
(高等院校计算机教育系列教材)
ISBN 978-7-302-45204-1

Ⅰ. ①数…　Ⅱ. ①吴…　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(2016)第263917号

责任编辑：吴艳华
封面设计：刘孝琼
责任校对：周剑云
责任印制：沈　露
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn, 010-62791865
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　张：23.25　　字　数：561千字
版　　次：2017年1月第1版　　印　次：2017年1月第1次印刷
印　　数：1～2000
定　　价：48.80元

产品编号：068564-01

前　言

数据挖掘涉及数据库技术、人工智能、统计学、机器学习、知识发现等多个学科的领域。随着信息技术的高速发展、数据量的飞速增长，数据挖掘已经在各行各业有了较为广泛的应用。

Microsoft SQL Server 2012 是集成了数据挖掘技术的第 5 版的 SQL Server。SQL Server 数据挖掘是业界部署最广泛的数据挖掘服务器，由于其可伸缩性大，容易获得，使用也较为简便，政府机构、企事业单位、学术人员和科学家也开始采用或转而使用 SQL Server 进行数据挖掘。IBM SPSS Statistics 是全世界最早的统计分析软件，其主要功能包括统计学分析运算、数据挖掘、预测分析等，由于其具有数据分析深入、使用方便、功能齐全等诸多优点，被广泛应用于自然科学、技术科学、社会科学的各个领域。

Microsoft SQL Server Analysis Services(SSAS)是本书的核心内容，Excel 的数据分析与挖掘，也是基于 SSAS 的服务引擎在进行。使用本书时，可以先学习数据挖掘基本理论；接下来学习 Excel 2010 数据分析与挖掘、SQL Server 2012 数据挖掘、SPSS Statistics 数据分析与挖掘；然后再通过完成教程设计的实验内容，真正地理解数据挖掘理论，掌握数据挖掘的实践技能。

本书结合作者多年从事数据挖掘教学、开发数据挖掘项目的经验，从实际出发，以实用的例子，系统地介绍了数据挖掘。全书分为三个部分，共 10 章。

第 1 部分由第 1～3 章组成，包括商业智能的概念和发展、数据挖掘和数据仓库的基本概念以及它们之间的关系；数据仓库的基本概念和设计步骤，并介绍了联机分析技术的分类和特点，以及回归分析、关联规则、聚类分析、决策树分析等数据挖掘常用分析方法的概念和算法。

第 2 部分由第 4～9 章组成，包括 Excel 2010 数据分析和预测的功能、Excel 2010 的数据挖掘功能；SQL Server 2012 的 Analysis Services 功能、设置数据源、设置数据源视图、设置挖掘结构、处理挖掘模型、查看挖掘结果等；Microsoft SQL Server Analysis Services 中提供的最常用的 6 个数据挖掘算法原理与参数；SPSS Statistics 的界面和基础操作；SPSS Statistics 在数据挖掘中常用的基础统计分析方法和高级统计分析方法。

第 3 部分由第 10 章组成，包括 SQL Server 2012 的数据挖掘实验、SPSS Statistics 的数据挖掘实验。

在内容的选择、深度的把握上，本书充分考虑到初学者的特点，在内容安排上力求循序渐进，不仅可以作为大专院校教学用书，也可以作为数据挖掘的培训教材和数据挖掘爱好者的自学用书。

本书由吴思远任主编，邹洋、黄梅根、贾玲任副主编。具体编写分工如下：邹洋编写第 1～3 章，吴思远编写第 4～6 章，黄梅根编写第 7 章，贾玲编写第 8～9 章，吴思远和贾玲共同编写第 10 章。吴思远负责全书架构的组织设计，负责统稿。本书的编写得到重庆邮

电大学教务处、重庆邮电大学计算机科学与技术学院以及重庆市教育评估院和中冶赛迪重庆信息技术有限公司的大力支持，在此感谢以上单位对本书所做出的贡献。

本书为教师提供了配套的教学资源，可从清华大学出版社网站 http://www.tup.com.cn 下载。

由于作者水平有限，书中难免有疏漏和不足之处，希望广大读者给予谅解和指正。

编　者

目　录

第1章 绪 论

数据挖掘是指从大型数据库中提取人们感兴趣的知识，这些知识是隐含的、事先不知的、潜在有用的信息。数据挖掘涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、高性能计算、专家系统等各个领域，其目的在于从大量数据中发现隐含的、新的、令人感兴趣的关系和规律。它不仅面向特定数据库的简单检索、查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导解决实际问题，发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。这样一来，就把人们对数据的应用从低层次的末端查询操作，提高到为各级经营决策者提供决策支持的层次。

本章着重介绍商业智能的概念和发展、数据挖掘和数据仓库的基本概念以及它们之间的关系，帮助读者理解商业智能、数据挖掘、数据仓库的基本要素，为读者学习以后的章节打下理论基础。

1.1 商 业 智 能

1.1.1 商业智能概述

1. 商业智能的定义

商业智能又称商务智能(Business Intelligence，BI)，是指用现代数据仓库技术、线上分析处理技术、数据挖掘和数据展现技术进行数据分析以实现商业价值。加特纳集团(Gartner Group)将商业智能定义为：商业智能描述了一系列的概念和方法，通过应用基于事实的支持系统来辅助商业决策的制定。商业智能技术提供使企业迅速分析数据的技术和方法，包括收集、管理和分析数据，将这些数据转化为有用的信息，然后分发到企业各处。

商业智能作为一个工具，是用来处理企业中现有数据，并将其转换成知识、分析和结论，以辅助业务或者决策者做出正确且明智的决定，是帮助企业更好地利用数据提高决策质量的技术，包含了从数据仓库到分析型系统等。

商业智能通常被理解为将企业中现有的数据转化为知识，帮助企业做出明智的业务经营决策的工具。这里所谈的数据包括来自企业业务系统的订单、库存、交易账目、客户和供应商的数据，来自企业所处行业和竞争对手的数据，以及来自企业所处的其他外部环境中的各种数据。商业智能能够辅助的业务经营决策，既可以是操作层的决策，也可以是战术层和战略层的决策。为了将数据转化为知识，需要利用数据仓库、联机分析处理(OLAP)工具和数据挖掘等技术。因此，从技术层面上讲，商业智能不是什么新技术，它只是数据仓库、OLAP 和数据挖掘等技术的综合运用。

可以认为，商业智能是对商业信息的搜集、管理和分析过程，目的是使企业的各级决策者获得知识或洞察力，促使他们做出对企业更有利的决策。商业智能一般由数据仓库、联机分析处理、数据挖掘、数据备份和恢复等部分组成。商业智能的实现涉及软件、硬件、咨询服务及应用，其基本体系结构包括数据仓库、联机分析处理和数据挖掘三个部分。

因此，把商业智能看成是一种解决方案应该比较恰当。商业智能的关键是从许多来自不同的企业运行系统的数据中提取出有用的数据并进行清理，以保证数据的正确性，然后经过抽取(Extraction)、转换(Transformation)和装载(Load)，即 ETL 过程，合并到一个企业级的数据仓库里，从而得到企业数据的一个全局视图，在此基础上利用合适的查询和分析工具、数据挖掘工具(大数据魔镜)、OLAP 工具等对其进行分析和处理(这时信息变为辅助决策的知识)，最后将知识呈现给管理者，为管理者的决策过程提供支持。商业智能的一般技术架构如图 1.1 所示。

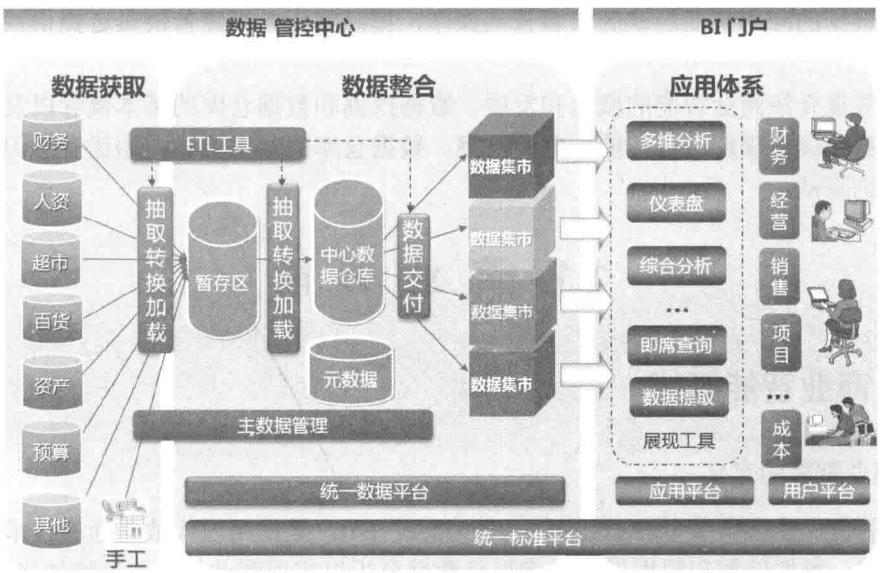

图 1.1　商业智能的一般技术架构

提供商业智能解决方案的著名 IT 厂商包括微软、IBM、Oracle、SAP、Informatica、Microstrategy、SAS 等。

2. 商业智能的应用

商业智能的应用可以大略分为业务分析和决策管理两个方面。

1)　业务分析方面

通过了解各种受众以及相关利益方的独特分析需求，可以发挥商业智能解决方案的全部潜能。企业所需的分析功能应该能够访问几乎所有企业数据源，而不受平台限制；同时可以为所有用户提供便于理解的详细信息视图，而不受用户角色或所在位置的影响。这些解决方案应具有创新的工具，以帮助这些不同的业务用户组轻松地通过台式机或移动设备分析信息。

企业需要广泛的分析功能，但不同的分析工具、信息壁垒、多种平台，以及过度依赖

于电子表格，让企业难以准确地分析信息。企业使用的分析解决方案必须能够满足所有业务用户的需求，包括一线员工，部门主管，以及高级分析员。这些用户希望能够自己分析数据，而无须等待部门提供所请求的信息，从而做出更出色、更智慧的业务决策。

需要说明的是，业务分析并非放之四海而皆准。用户需求可能会有很大的不同。通过了解不同类型的分析需求，并将其与组织中的特定角色相联系，企业可以从中受益。

2) 决策管理方面

决策管理是用来优化并自动化业务决策的一种卓有成效的方法。它通过预测分析，让组织能够在制定决策以前有所行动，以便预测哪些行动在未来最有可能获得成功。从广义角度来看，主要存在三种组织决策类型，即战略型、业务型和战术型。

其中，战略决策通常为组织设定长远方向，其制定者是部门主管人员、副总裁、业务线经理；业务决策通常包括策略或流程的制定，它们专注于在战术级别上执行特定项目或目标，其制定者为业务经理、系统经理和业务分析师；战术决策通常是将策略、流程或规则应用到具体事例的“前线”行动。这些类型的决策适用于自动化，使结果更具一致性和可预测性，其制定者包括消费者服务代表、财务服务代表、分支经理、销售人员，以及网站推荐引擎等自动化系统。

决策管理使改进成为可能。它使用决策流程框架和分析来优化并自动化决策、优化成果，且解决特定的业务问题。决策管理通常专注于大批量决策，并使用基于规则和基于分析模型的应用程序实现决策。因此，虽然决策管理相对较新，但有效性已得到证实。

了解了组织中的决策类型和可用的决策管理选择后，就可以着手建立决策管理基础架构了。业务经理首先应该在影响他们决策的范围内定义其业务挑战，然后通过为特定业务问题开发的以决策为中心的应用程序，利用决策管理优化目标决策。这些应用程序展示了业务人员熟悉的相关信息，并在影响问题的决策范围内加入了预测分析。

3. 商业智能的实施步骤

实施商业智能系统是一项复杂的系统工程，整个项目涉及企业管理、运作管理、信息系统、数据仓库、数据挖掘和统计分析等众多门类的知识。因此，用户除了要选择合适的商业智能软件工具外，还必须使用正确的实施方法，才能保证项目获得成功。 商业智能项目的实施步骤如下。

(1) 需求分析：需求分析是商业智能实施的第一步，在其他活动开展之前必须明确企业对商业智能的期望和需求，包括需要分析的主题，各主题可能查看的角度(维度)；需要发现企业哪些方面的规律，必须明确用户的需求。

(2) 数据仓库建模：通过对企业需求的分析，建立企业数据仓库的逻辑模型和物理模型，并规划好系统的应用架构，将企业各类数据按照分析主题进行组织和归类。

(3) 数据抽取：数据仓库建立后，必须将数据从业务系统中抽取到数据仓库中，在抽取的过程中还必须将数据进行转换、清洗，以适应分析的需要。

(4) 建立商业智能分析报表：商业智能分析报表需要专业人员按照用户指定的格式进行开发，用户也可自行开发(开发方式简单、快捷)。

(5) 用户培训和数据模拟测试：对于开发—使用分离型的商业智能系统，最终用户的使用是相当简单的，只需要点击操作就可针对特定的商业问题进行分析。

(6) 系统改进和完善：任何系统的实施都必须是不断完善的，商业智能系统更是如此。在用户使用一段时间后，可能会提出更多、更具体的要求，这时需要再按照上述步骤对系统进行重构或完善。

4. 商业智能与企业效益

商业智能帮助企业的管理层进行快速、准确的决策，迅速地发现企业中的问题，提示管理人员加以解决。但商业智能软件系统不能代替管理人员进行决策，不能自动处理企业运行过程中遇到的问题。因此，商业智能系统并不能为企业带来直接的经济效益。但必须看到，商业智能为企业带来的是一种经过科学武装的管理思维，给整个企业带来的是决策的快速性和准确性，发现问题的及时性，以及发现那些对手未发现的潜在的知识和规律，而这些信息是企业产生经济效益的基础。不能快速、准确地制定决策方针，等于将市场送给对手；不能及时发现业务中的潜在信息，等于浪费自己的资源。比如，通过对销售数据的分析，可发现各类客户的特征和喜欢购买商品之间的联系，就可更有针对性地进行精确的促销活动或向客户提供更具个性的服务，这都会为企业带来直接的经济效益。

1.1.2 商业智能的发展

提到“商业智能”这个词，网上普遍认为是加特纳集团在 1996 年第一次提出来的，但事实上 IBM 的研究员 Hans Peter Luhn 早在 1958 年就用到了这一概念。他将“智能”定义为“对事物相互关系的一种理解能力，并依靠这种能力去指导决策，以达到预期的目标”。

1. 商业智能的发展历程

对商业智能发展有着“里程碑”意义的事件如下。

1970 年，IBM 的研究员埃德加·弗兰克·科德(E.F. Codd)发明了关系型数据库。

1979 年，一家以创建决策支持系统为己任，致力于构建单独的数据存储结构的公司 Teradata 诞生。1983 年，该公司利用并行处理技术为美国富国银行建立了第一个决策支持系统。

1988 年，为解决企业集成问题，IBM 公司的研究员 Barry Devlin 和 Paul Murphy 创造性地提出一个新的术语：数据仓库(Data Warehouse)。

1992 年，比尔·恩门出版了《如何构建数据仓库》一书，他主张由顶至底的构建方法，强调数据的一致性，拉开了数据仓库真正得以大规模应用的序幕。

1993 年，拉尔夫·金博尔出版了《数据仓库的工具》一书，他主张务实的数据仓库应该由下往上，从部门到企业，并把部门的数据仓库叫作“数据集市”。

2. 商业智能的发展趋势

与其他系统相比，商业智能具有更美好的发展前景，其发展趋势可以归纳为以下几点。

1) 功能上具有可配置性、灵活性、可变化性

BI 系统的范围从为部门的特定用户服务扩展到为整个企业所有用户服务。同时，由于企业用户在职权、需求上的差异，BI 系统提供广泛的、具有针对性的功能，从简单的数据获取，到利用 Web 和局域网、广域网进行丰富的交互、决策信息，以及知识的分析和使用，

解决方案更开放、可扩展，可按用户定制，在保证核心技术的同时，提供客户化的界面。

针对不同企业的独特需求，BI 系统在提供核心技术的同时，使系统又具个性化，即在原有方案基础上加入自己的代码和解决方案，增强客户化的接口和扩展特性；可为企业提供基于商业智能平台的定制工具，使系统具有更大的灵活性和使用范围。

2) 从单独的商业智能向嵌入式商业智能发展

这是商业智能应用的一大趋势，即在企业现有的应用系统中(如财务、人力、销售等系统)嵌入商业智能组件，使普遍意义上的事务处理系统具有商业智能的特性。即便只考虑 BI 系统的某个组件而不是整个 BI 系统，也并非一件简单的事，比如将 OLAP 技术应用到某一个应用系统，一个相对完整的商业智能开发过程(如企业问题分析、方案设计、原型系统开发、系统应用等过程)是不可缺少的。

3) 从传统功能向增强型功能转变

增强型的商业智能功能是相对于早期的用 SQL 工具实现查询的商业智能功能。当前应用中的 BI 系统除实现传统的 BI 系统功能之外，大多数已实现了图 1.1 中应用体系层的功能。而数据挖掘、企业建模是 BI 系统应该加强的应用，以便更好地提高系统性能。

4) 市场增长强势不减

BI 软件市场在最近几年得到了迅速增长。在这个市场中，终端用户查询、报告和 OLAP 工具占绝对主流，达到 65%。用户希望从企业资源规划(ERP)、客户关系管理(CRM)、供应链管理(SCM)和遗留系统中发掘他们的数据资产，因此，对 BI 软件的需求正在不断增加。从这些需求来看，说明企业正逐渐摆脱单纯依赖于软件来处理日常事务的情况，而是明确要利用软件来帮助自己，依据企业数据做出更好、更快的决策。

此外，对分析应用需求的增加将持续刺激对商业智能软件的需求。这些软件主要用来进行复杂的预测，得出相对直接的执行报告，另外也包括以多维分析工具为基础的客户分类应用。

5) 商业智能解决方案走向完整

来自国外的统计结果表明，全球企业的信息量平均每 1.5 年翻一番，而仅仅利用了全部信息数据中的 7%。随着知识经济时代的来临，记录客户与市场数据的信息和信息利用能力已经成为决定企业生死存亡的关键因素，越来越多的国内外企业已经根据信息流和数据分析技术进行企业重整，传统的数据记录方式被更先进的商业智能技术所代替。在商业智能解决方案的帮助下，企业级用户可以通过充分挖掘现有的数据资源，捕获信息、分析信息、沟通信息，发现许多过去缺乏认识或未被认识的数据关系，帮助企业管理者做出更好的商业决策，如开拓什么市场、吸引哪些客户、促销何种产品等。商业智能还能够通过财务分析、风险管理、欺诈分析、销售分析等过程帮助企业降低运营成本，进而获得更高的经营效益。

根据世界权威性的 IDC 公司的调查结果，企业用于商业智能的投资 3 年平均回报率高达 400%。

1.2 数据挖掘

1.2.1 数据挖掘的定义

数据挖掘(Data Mining)，又译为资料探勘、数据采矿。它是数据库知识发现(Knowledge-Discovery in Databases，KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统和模式识别等诸多方法来实现上述目标。

数据挖掘采用了如下一些领域的思想和理论。

(1) 来自统计学的抽样、估计和假设检验。

(2) 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。

(3) 数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。

(4) 一些领域也对数据挖掘起到重要的支撑作用。数据挖掘需要高效的数据库系统提供有效的存储、索引和查询处理支持；源于高性能并行计算的技术在处理海量数据集方面很有用；分布式技术能帮助处理海量数据，并且在数据不能集中到一起处理时更是至关重要。

从数据本身来考虑，数据挖掘的过程通常需要有数据清理、数据变换、数据挖掘过程、模式评估和知识表示等八个步骤。

(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。

(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。

(3) 数据规约：多数的数据挖掘算法即使在少量数据上执行也需要很长的时间，而做商业运营数据挖掘时往往数据量非常大。数据规约技术可以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。

(4) 数据清理：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。

(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。对于某些实数型数据，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。

(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。

(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。

(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。

数据规约、数据清理、数据变换又合称数据预处理。在数据挖掘中，至少 60%的费用可能要花在信息收集阶段，而至少 60%以上的精力和时间是花在数据预处理阶段。

1.2.2 数据挖掘的重要性

据预测，到 2020 年，全球以电子形式存储的数据量将达到 35ZB，是 2009 年全球存储量的 40 倍。而在 2010 年年底，根据 IDC 的统计，全球数据量已经达到了 120 万 PB，或 1.2ZB。如果将这些数据都刻录在 DVD 上，那么光把这些 DVD 盘片堆叠起来就可以从地球垒到月球一个来回(单程约 24 万英里)。

在信息化的建设过程中，众所周知，数据可以分为结构化数据、半结构化数据和非结构化数据三种。其中，85%的数据属于企业业务过程中产生的文档等非结构化数据。

面对着海量的数据，人们不禁感叹，大数据时代已经到来，悲观者为管理和维护而忧虑，乐观者则看到了大数据的大价值。何谓“大数据”，目前没有统一的定义。通常认为，它是海量的非结构化数据，其特点是数据量很大，数据的形式多样化。如何存储这些快速增长的、海量的数据？如何对大数据进行数据挖掘，挖掘出价值？相关的一系列问题成为所有企业面临的共同挑战。

数据挖掘在各领域的应用非常广泛，只要该产业拥有具有分析价值与需求的数据仓储或数据库，皆可利用挖掘工具进行有目的的挖掘分析。一般较常见的应用案例多发生在科学研究领域、零售业、制造业、财务金融保险业、通信业以及医疗服务等。从目前网络招聘的信息来看，规模大小不同公司采用数据挖掘的应用有 50 多个方面，如表 1.1 所示。

表 1.1 数据挖掘的应用领域

1.数据统计分析	20.风险数据分析	39.数据实验模拟
2.预测预警模型	21.缺陷信息挖掘	40.数学建模与分析
3.数据信息阐释	22.决策数据支持	41.呼叫中心数据分析
4.数据采集评估	23.运营优化与成本控制	42.贸易/进出口数据分析
5.数据加工仓库	24.质量控制与预测预警	43.海量数据分析系统设计、关键技术研究
6.品类数据分析	25.系统工程数学技术	44.数据清洗、分析、建模、调试、优化
7.销售数据分析	26.用户行为分析/客户需求模型	45.数据挖掘算法的分析研究、建模、实验模拟
8.网络数据分析	27.产品销售预测(热销特征)	46.组织机构运营监测、评估、预测预警
9.流量数据分析	28.商场整体利润最大化系统设计	47.经济数据分析、预测、预警
10.交易数据分析	29.市场数据分析	48.金融数据分析、预测、预警

续表

11.媒体数据分析	30.综合数据关联系统设计	49.科研数学建模与数据分析：社会科学，自然科学，医药，农学，计算机，工程，信息，军事，图书情报等
12.情报数据分析	31.行业/企业指标设计	50.数据指标开发、分析与管理
13.金融产品设计	32.企业发展关键点分析	51.产品数据挖掘与分析
14.日常数据分析	33.资金链管理设计与风险控制	52.商业数学与数据技术
15.总裁万事通	34.用户需求挖掘	53.故障预测预警技术
16.数据变化趋势	35.产品数据分析	54.数据自动分析技术
17.预测预警模型	36.销售数据分析	55.泛工具分析
18.运营数据分析	37.异常数据分析	56.互译
19.商业机遇挖掘	38.数学规划与数学方案	57.指数化

在以上的领域中，采用数据挖掘技术都大大提高了效率，数据挖掘已经在国计民生的各个方面扮演着越来越重要的角色。

1.2.3 数据挖掘的功能

1. 数据挖掘信息的种类

数据挖掘是为了从现有数据中获得信息。数据挖掘能够发现的信息主要有以下五种。

(1) 概念信息，类别特征的概括性描述知识。根据数据的微观特征发现同类事物带有普遍性的、较高层次概念的共同性质，是一种对数据的概况、提炼和抽象。

(2) 关联信息，主要反映一个事件和其他事件之间的依赖或者关联性。如果两项或者多项属性之间存在关联，那么其中一项的属性值就可以根据其他属性值进行预测。这类知识发现方法中最有名的就是 Apriori 算法。

(3) 分类信息，主要反映同类事物的共同特征和不同事物之间的差异。

(4) 预测性信息，根据历史数据和当前数据对未来数据进行预测，主要是时间序列预测。

(5) 偏差性信息，这是对差异和阶段特例的揭示，如数据聚类的离群值等。

2. 数据挖掘的功能

为了获取以上信息，对应到数据挖掘的流程方法上，数据挖掘的功能通常表现为：分类 (Classification)、估计(Estimation)、预测(Prediction)、相关性分组或关联规则(Affinity grouping or association rules)、聚类(Clustering)、描述和可视化(Text、Web、图形图像、视频、音频等)，具体如下。

1) 分类

首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分类。例如：

(1) 信用卡申请者，分类为低、中、高风险。