高等院校经济管理类规划教材

PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE
商业智能原理与应用

蔡 颖 鲍立威 编著

PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE
PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE
PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

高等院校经济管理类规划教材

PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE

商业智能原理与应用

蔡 颖 鲍立威 编著

PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE

PRINCIPLES AND APPLICATIONS OF BUSINESS INTELLIGENCE

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

商业智能原理与应用 / 蔡颖，鲍立威编著. —杭州：浙江大学出版社，2011.9
ISBN 978-7-308-09120-6

Ⅰ.①商… Ⅱ.①蔡…②鲍… Ⅲ.①关系数据库—数据库管理系统 Ⅳ.①TP311.138

中国版本图书馆 CIP 数据核字（2011）第 191456 号

商业智能原理与应用

蔡　颖　鲍立威　编著

责任编辑　周卫群
封面设计　东方博
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　浙江时代出版服务有限公司
印　　刷　富阳市育才印刷有限公司
开　　本　787mm×1092mm　1/16
印　　张　19.5
字　　数　475 千
版 印 次　2011 年 9 月第 1 版　2011 年 9 月第 1 次印刷
书　　号　ISBN 978-7-308-09120-6
定　　价　36.00 元

前 言

计算机与信息技术经历了半个多世纪的发展，已经渗透到社会生产与生活的方方面面。目前，绝大部分大中型企事业单位都已经建立了比较完善的 OA、ERP、CRM 等信息化系统。这些系统为日常事务、办公管理和生产经营等提供了高效的手段，同时也产生了海量的数据。管理人员和经营人员也越来越不满足于传统信息化系统的事务处理功能，而更希望从所积累的数据中，发现对支持其科学决策更有价值的信息。

商业智能(Business Intelligence，简写为 BI)，又称商务智能，是指将企事业单位积累的数据转化为知识，帮助企事业单位做出科学决策的工具。这里的数据包括来自企事业单位自身业务系统中的数据，如订单、库存、交易账目、客户和供应商资料等，来自企业所处行业和竞争对手的数据，以及来自外部环境中的各种相关数据。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。

现代管理越来越表现出高度科学化和高度组织化的特征，它通过信息获取、计划、决策、组织、领导、控制、实施等一系列活动和过程，以最小的人力、物力和财力支出，最大化地发挥资源的效力，来实现组织的目标。其中，信息的分析处理与利用是达到这个目的的基本技术，是当代社会对经管类应用人才的基本技能要求。对企业而言，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，帮助他们做出对企业发展更加有利的决策(对事业单位也是如此)。随着互联网的普及，基于互联网的各种信息系统在企业中的应用，企业搜集到越来越多的关于客户、产品及市场和销售等情况的各种信息，这些信息能帮助企业更好地预测和把握未来。所以，互联网的广泛应用，特别是电子商务的发展有力地推动了商业智能的进一步应用。

本书全面系统地介绍了商业智能的基本概念、基本方法和基本技术，并以 Microsoft SQL Server 作为数据仓库管理平台，以 SQL Server Business Intelligence Development Visual Studio 作为商业智能开发平台，进行了丰富的案例演示。本书共分 11 章，第 1 章介绍了数据挖掘的发展过程和商业智能的定义；第 2 章讲解了数据仓库的概念、体系结构和设计与实施；第 3 章讲解了数据预处理的主要方法及如何使用 SQL Server Integration Services 进行数据的清理、转换和装载；第 4 章主要讲解了多维数据分析的方法和如何使用 SQL Server Analysis Services 构建多维数据集；第 5 章主要讲解了如何使用 SQL Server Report Services 构建智能报表；第 6 章至第 10 章主要讲解了数据挖掘技术和主要的数据挖掘算法，并使用 SQL Server 中的数据挖掘算法对挖掘过程作了详细的描述；第 11 章主要介绍基于多维数据集的数据挖掘及相关案例。

本书同时还安排了与各章内容相对应的练习【思考题】和课堂演示实验及课后实验，通

过实际的操作使读者加深对数据挖掘技术的理解和掌握。(如需本书的课件、课堂演示实验及课后实验,请与 caiy@zucc.edu.cn 联系)。本书可以作为高等院校高年级本科生的教材,也可以作为 MBA 的教材以及 IT 相关专业人员、市场营销人员、管理决策支持等实际经济管理领域实务工作者的参考用书。书中不当之处,敬请读者批评指正。

编　者

2011 年 7 月

目 录

第 1 章 数据挖掘和商业智能 …… (1)

1.1 数据挖掘的兴起 …… (1)
1.1.1 数据丰富与知识匮乏 …… (1)
1.1.2 从数据到知识 …… (2)
1.1.3 数据挖掘产生 …… (3)
1.1.4 数据挖掘解决的商业问题 …… (4)
1.2 什么是商业智能 …… (5)
1.2.1 企业决策实现过程的信息需求 …… (5)
1.2.2 企业信息化系统中的商业智能 …… (6)
1.2.3 商业智能的体系结构 …… (7)
1.3 数据挖掘和商业智能工具 …… (9)
1.3.1 商业智能工具的选择 …… (9)
1.3.2 SQL Server 2008 的商业智能构架 …… (10)
1.4 数据挖掘应用案例 …… (13)
【本章小结】 …… (15)
【练习题】 …… (16)

第 2 章 数据仓库 …… (17)

2.1 数据仓库的概念 …… (17)
2.1.1 从传统数据库到数据仓库 …… (17)
2.1.2 数据仓库的定义与基本特性 …… (19)
2.2 数据仓库的体系结构 …… (21)
2.2.1 数据仓库的物理结构 …… (21)
2.2.2 数据仓库的系统结构 …… (22)
2.2.3 数据仓库的数据模型 …… (23)
2.3 元数据 …… (26)
2.3.1 元数据的定义 …… (26)
2.3.2 元数据的分类及作用 …… (26)
2.4 数据集市 …… (28)
2.4.1 两种数据集市结构 …… (28)
2.4.2 数据集市与数据仓库的差别 …… (29)
2.4.3 关于数据集市的误区 …… (29)

2.5 数据仓库设计与实施 …… (30)
2.5.1 自上而下还是自下而上的设计方法 …… (30)
2.5.2 数据仓库的设计步骤 …… (31)
2.5.3 数据仓库的实施 …… (40)
2.5.4 数据仓库的使用和维护 …… (41)
2.6 Microsoft 数据仓库(DW)和商业智能(BI)工具 …… (41)
2.7 数据仓库设计案例 …… (43)
2.7.1 业务数据库 AdventureWorks …… (44)
2.7.2 业务数据分析 …… (46)
2.7.3 项目需求分析 …… (47)
2.7.4 构建数据仓库 …… (48)
【本章小结】 …… (50)
【练习题】 …… (50)

第 3 章 数据预处理 …… (51)

3.1 数据预处理的重要性 …… (51)
3.2 数据清洗 …… (53)
3.2.1 遗漏数据处理 …… (53)
3.2.2 噪声数据处理 …… (54)
3.2.3 不一致数据处理 …… (55)
3.3 数据集成与转换 …… (56)
3.3.1 数据集成处理 …… (56)
3.3.2 数据转换处理 …… (56)
3.4 数据消减 …… (58)
3.4.1 数据立方合计 …… (59)
3.4.2 维数消减 …… (59)
3.4.3 数据块消减 …… (60)
3.5 离散化和概念层次树生成 …… (63)
3.5.1 数值概念层次树生成 …… (64)
3.5.2 类别概念层次树生成 …… (66)
3.6 使用 SSIS 对数据进行 ETL 操作 …… (67)
3.6.1 SSIS 的主要功能 …… (68)
3.6.2 SSIS 的体系结构 …… (70)
3.6.3 SSIS 包主要对象 …… (74)
3.6.4 创建并运行一个简单的包 …… (76)
【本章小结】 …… (88)
【思考题】 …… (88)

第 4 章　多维数据分析 …… (89)

4.1　多维数据分析基础 …… (89)
4.2　多维数据分析方法 …… (92)
4.3　多维数据的存储方式 …… (95)
4.3.1　三种存储方式 …… (95)
4.3.2　三种存储方式的比较 …… (97)
4.4　多维表达式(MDX) …… (98)
4.4.1　MDX 中的重要概念 …… (98)
4.4.2　MDX 基本语法 …… (100)
4.4.3　MDX 与 SQL 的区别 …… (101)
4.4.4　MDX 核心函数 …… (102)
4.5　使用 SQL Server Analysis Services(SSAS)构建维度和多维数据集 …… (108)
4.5.1　SSAS 的体系结构 …… (108)
4.5.2　SSAS 的统一维度模型(UDM) …… (109)
4.5.3　SSAS 示例 …… (111)
4.6　使用 Excel 数据透视图浏览多维数据集 …… (141)
【本章小结】…… (146)
【思考题】…… (147)

第 5 章　用 Microsoft SSRS 处理智能报表 …… (148)

5.1　SSRS 商业智能报表 …… (148)
5.1.1　商业智能报表与商业智能 …… (148)
5.1.2　SSRS 的结构 …… (150)
5.1.3　SSRS 报表的 3 种状态 …… (151)
5.2　使用 SSRS 创建报表 …… (151)
5.2.1　创建一个简单报表项目 …… (151)
5.2.2　增强基本报表的功能 …… (153)
5.2.3　发布报表 …… (160)
【本章小结】…… (160)

第 6 章　数据挖掘技术 …… (161)

6.1　数据挖掘的任务 …… (161)
6.1.1　分类 …… (162)
6.1.2　回归 …… (163)
6.1.3　时间序列分析 …… (163)
6.1.4　预测 …… (164)
6.1.5　聚类 …… (164)

6.1.6 关联规则……(165)
6.1.7 序列分析……(166)
6.1.8 偏差检测……(166)
6.2 数据挖掘的对象 ……(167)
6.3 数据挖掘系统的分类 ……(171)
6.4 数据挖掘项目的生命周期 ……(172)
6.4.1 商业理解……(173)
6.4.2 数据准备……(173)
6.4.3 模型构建……(173)
6.4.4 模型评估……(174)
6.4.5 应用集成和实施……(174)
6.5 数据挖掘面临的挑战及发展 ……(175)
6.5.1 数据挖掘面临的挑战……(175)
6.5.2 数据挖掘的发展趋势……(176)
【本章小结】……(178)
【思考题】……(179)

第7章 关联挖掘 ……(180)

7.1 关联规则挖掘 ……(181)
7.1.1 购物分析:关联挖掘 ……(181)
7.1.2 基本概念……(181)
7.1.3 关联规则挖掘分类……(182)
7.2 单维布尔关联规则挖掘 ……(183)
7.2.1 Apriori 算法 ……(183)
7.2.2 关联规则的生成……(186)
7.3 挖掘多层级关联规则 ……(186)
7.3.1 挖掘多层次关联规则……(186)
7.3.2 挖掘多层次关联规则方法……(188)
7.3.3 多层次关联规则的冗余……(190)
7.4 多维关联规则的挖掘 ……(191)
7.4.1 多维关联规则……(191)
7.4.2 利用静态离散挖掘多维关联规则……(192)
7.5 关联挖掘中的相关分析 ……(193)
7.5.1 无意义强关联规则示例……(193)
7.5.2 从关联分析到相关分析……(194)
7.6 利用 Microsoft SSAS 进行关联挖掘……(195)
7.6.1 Microsoft 关联规则模型简介 ……(195)
7.6.2 关联规则数据挖掘示例……(197)
【本章小结】……(206)

【思考题】…… (206)

第 8 章　分类与预测 …… (207)

8.1　分类与预测基本知识 …… (207)
8.2　有关分类和预测的几个问题 …… (209)
8.3　基于决策树的分类 …… (210)
8.3.1　决策树生成算法 …… (210)
8.3.2　属性选择方法 …… (211)
8.3.3　树枝修剪 …… (213)
8.3.4　决策树分类规则获取 …… (214)
8.3.5　级别决策树方法的改进 …… (215)
8.3.6　数据仓库技术与决策树归纳的结合 …… (216)
8.4　贝叶斯分类方法 …… (217)
8.4.1　贝叶斯定理 …… (217)
8.4.2　基本贝叶斯分类方法 …… (218)
8.5　神经网络分类方法 …… (220)
8.5.1　多层前馈神经网络 …… (220)
8.5.2　神经网络结构 …… (221)
8.5.3　后传方法 …… (221)
8.5.4　后传方法和可理解性 …… (224)
8.6　分类器准确性 …… (225)
8.6.1　分类器准确性估计 …… (225)
8.7　预测方法 …… (226)
8.7.1　线性与多变量回归 …… (226)
8.7.2　非线性回归 …… (227)
8.7.3　其它回归模型 …… (228)
8.8　Microsoft 贝叶斯算法 …… (228)
8.8.1　贝叶斯算法的参数 …… (228)
8.8.2　使用贝叶斯模型 …… (229)
8.8.3　浏览贝叶斯模型 …… (231)
8.9　Microsoft 决策树算法 …… (234)
8.10　Microsoft 神经网络算法 …… (240)
【本章小结】…… (242)
【思考题】…… (242)

第 9 章　聚类分析 …… (243)

9.1　聚类分析概念 …… (243)
9.2　聚类分析中的数据类型 …… (245)
9.2.1　间隔数值属性 …… (246)

9.2.2 二值属性 …… (247)
9.2.3 符号、顺序和比例数值属性 …… (248)
9.2.4 混合类型属性 …… (250)
9.3 主要聚类方法 …… (251)
9.4 划分方法 …… (252)
9.4.1 传统划分方法 …… (252)
9.4.2 大数据库的划分方法 …… (255)
9.5 层次方法 …… (256)
9.5.1 两种基本层次聚类方法 …… (256)
9.6 基于密度方法 …… (258)
9.6.1 基于密度方法:DBSCAN …… (258)
9.7 异常数据分析 …… (259)
9.7.1 基于统计的异常检测方法 …… (260)
9.7.2 基于距离的异常检测方法 …… (261)
9.7.3 基于偏差的异常检查方法 …… (261)
9.8 Microsoft 聚类算法 …… (263)
【本章小结】 …… (270)
【思考题】 …… (270)

第 10 章 时序数据和序列数据挖掘 …… (271)

10.1 时间序列模型 …… (271)
10.2 Microsoft 的时序算法 …… (273)
10.2.1 自动回归 …… (273)
10.2.2 自动回归树 …… (274)
10.2.3 数据中的季节性处理 …… (275)
10.2.4 使用预测函数预测值 …… (275)
10.3 Microsoft 时序算法示例 …… (276)
10.4 Microsoft 的序列模式挖掘 …… (281)
10.4.1 Microsoft 序列聚类算法 …… (281)
10.4.2 序列聚类挖掘示例 …… (284)
【本章小结】 …… (289)
【思考题】 …… (290)

第 11 章 基于多维数据集的数据挖掘 …… (291)

11.1 OLAP 和数据挖掘之间的关系 …… (291)
11.2 构建 OLAP 挖掘模型 …… (293)
【本章小结】 …… (300)

第 1 章

数据挖掘和商业智能

导入语

数据挖掘作为一个新兴的多学科交叉应用领域，正在各行各业的决策支持活动中扮演着越来越重要的角色。数据挖掘技术与普通的数据分析有质的不同，数据挖掘技术以高度精确和高度可靠的手段从海量数据中挖掘和发现新的知识。从商业角度看，数据挖掘是一种强大的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和模型化处理，从中提取可用于辅助商业决策的关键性数据和知识。

本章将着重介绍以下内容：

- 数据挖掘的兴起
- 商业智能的概念
- 数据挖掘和商业智能工具
- 一些成功运用数据挖掘的案例

1.1　数据挖掘的兴起

1.1.1　数据丰富与知识匮乏

计算机与信息技术经历了半个多世纪的发展，使得我们能以更快速更容易更廉价的方式获取和存储数据。早在 20 世纪 80 年代，据粗略估算，全球信息量每隔 20 个月就增加一倍。而进入 90 年代，全世界所拥有的数据库及其所存储的数据规模增长更快。一个中等规模企业每天要产生 100MB 以上来自各生产经营等多方面的商业数据。美国政府部门的一个典型大数据库每天要接收约 5TB 数据量，在 15 秒到 1 分钟时间里，要维持的数据量达到

300TB，存档数据达 15～100PB。在科研方面，以美国宇航局的数据库为例，每天从卫星下载的数据量就达 3～4TB 之多；而为了研究的需要，这些数据要保存七年之久。90 年代互联网的出现与发展，以及随之而来的企业内部网和企业外部网以及虚拟网的产生和应用，使整个世界互联形成一个小小的地球村，人们可以跨越时空地在网上交换信息和协同工作。这样，展现在人们面前的已不是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋。据 IDC（互联网数据中心）与 EMC（提供全球信息存储及管理产品、服务和解决方案的公司）的研究报告指出：数字资源正以 52%的复合年均增长率飞速增长，至 2006 年，所创建、存储、复制的数字信息总量达到 161EB（1EB＝10 亿 GB），而到 2010 年，创建的数字信息总量将达到 988EB。面对如此极度膨胀的数据信息量，人们深刻感受到“信息爆炸”、“混沌信息空间”和“数据过剩”的巨大压力。

然而，人类的各项活动都是基于人类的智慧和知识，通过对外部世界的观察和了解，做出正确的判断和决策以及采取正确的行动，而数据仅仅是人们用各种工具和手段观察外部世界所得到的原始材料，它本身没有任何意义。从数据到知识到智慧，需要经过分析加工处理精炼的过程。如图 1-1 所示，数据是原材料，它只是描述发生了什么事情，并不能构成决策或行动的可靠基础。通过对数据进行分析找出其中关系，赋予数据以某种意义和关联，这就形成所谓信息。信息虽给出了数据中一些有一定意义的东西，但它往往和人们需要完成的任务没有直接的联系，也还不能作为判断、决策和行动的依据。对信息进行再加工，即进行更深入的归纳分析，从信息中理解其模式，方能获得更有用的信息，即知识。在大量知识积累基础上，总结出原理和法则，就形成了所谓智慧。

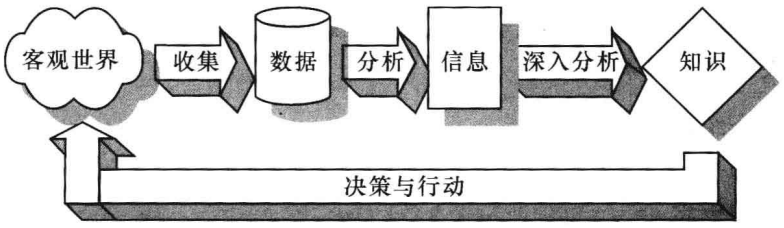

图 1-1　人类活动所涉及数据与知识之间的关系描述

计算机与信息技术的发展，加速了人类知识创造与交流的这种进程。据德国《世界报》的资料分析，如果说 19 世纪时科学定律（包括新的化学分子式、新的物理关系和新的医学认识）的认识数量一百年增长一倍，到 20 世纪 60 年代中期以后，每五年就增加一倍。这中间知识起着关键的作用。当数据量极度增长时，如果没有有效的方法，由计算机及信息技术来帮助人们从中提取有用的信息和知识，人类显然就会感到像大海捞针一样束手无策。据估计，目前一个大型企业数据库中数据，约只有百分之七得到很好应用。由此可见，目前人类陷入了一个尴尬的境地，即“丰富的数据”而“贫乏的知识”。

1.1.2　从数据到知识

早在 20 世纪 80 年代，人们在“物竞天择，适者生存”的大原则下，就认识到“谁最先从外部世界获得有用信息并加以利用，谁就可能成为赢家”。而今置身市场经济且面向全球性激烈竞争的环境下，任何商家的优势不单纯地取决于如产品、服务、地区等方面因素，而在于创

新。用知识作为创新的原动力，就能使商家长期持续地保持竞争优势。因此要能及时迅速地从日积月累庞大的数据库中，以及互联网上获取与经营决策相关的知识，自然而然就成为满足易变的客户需求以及因市场快速变化而引起激烈竞争局面的唯一武器。因此，如何对数据与信息快速有效地进行分析加工提炼以获取所需知识，就成为计算机及信息技术领域的重要研究课题。

事实上，计算机及信息技术发展的历史也是数据和信息加工手段不断更新和改善的历史。早年受技术条件限制，一般用人工方法进行统计分析和用批处理程序进行汇总和提出报告。在当时市场情况下，月度和季度报告已能满足决策所需信息要求。随着数据量的增长，多数据源所带来的各种数据格式不相容性，为了便于获得决策所需信息，就有必要将整个机构内的数据以统一形式集成存储在一起，这就形成了数据仓库(data warehouse，DW)。数据仓库不同于管理日常工作数据的数据库，它是为了便于分析针对特定主题的集成化的、提供存贮 5～10 年或更长时间的数据，这些数据一旦存入就不再发生变化。

数据仓库的出现，为更深入对数据进行分析提供了条件。针对市场变化的加速，人们提出了能进行实时分析和产生相应报表的在线分析工具 OLAP(On Line Analytical Processing)。OLAP 能允许用户以交互方式浏览数据仓库内容，并对其中数据进行多维分析。例如 OLAP 能对不同时期、不同地域的商业数据中变化趋势进行对比分析。

OLAP 是数据分析手段的一大进步，以往的分析工具所得到的报告结果只能回答“什么”，而 OLAP 的分析结果能回答“为什么”。但 OLAP 分析过程是建立在用户对深藏在数据中的某种知识有预感和假设的前提下，是在用户指导下的信息分析与知识发现过程。由于数据仓库(通常数据贮藏量以 TB 计)内容来源于多个数据源，因此其中埋藏着丰富的不为用户所知的有用信息和知识，而要使企业能及时准确地做出科学的经营决策，以适应变化迅速的市场环境，就需要有基于计算机与信息技术的智能化自动工具，来帮助挖掘隐藏在数据中的各类知识。这类工具不应再基于用户假设，而应能自身生成多种假设，再用数据仓库(或大型数据库)中的数据进行检验或验证，然后返回用户最有价值的检验结果。此外，这类工具还应能适应现实世界中数据的多种特性(即量大、含噪声、不完整、动态、稀疏性、异质、非线性等)。要达到上述要求，只借助于一般数学分析方法是无法达到的。多年来，数理统计技术方法以及人工智能和知识工程等领域的研究成果，诸如推理、机器学习、知识获取、模糊理论、神经网络、进化计算、模式识别、粗糙集理论等等诸多研究分支，给开发满足这类要求的数据深度分析工具提供了坚实而丰富的理论和技术基础，这是从数据到知识演化过程中的一个重要里程碑。利用数据库技术和信息技术辅助从数据提取知识的演化过程如图 1-2 所示。

图 1-2　数据到知识的演化过程示意描述

1.1.3　数据挖掘产生

1989 年 8 月，在第 11 届国际人工智能联合会议的专题研讨会上首次提出了基于数据

库的知识发现(KDD,Knowledge Discovery in Database)技术。该技术涉及机器学习、模式识别、统计学、智能数据库、知识获取、专家系统、数据可视化、高性能计算等领域,技术难度较大,一时难以满足实际需要。到了1995年,在美国计算机年会上,提出了数据挖掘(DM,Data Mining)的概念,即通过数据库抽取隐含的、未知的、具有潜在使用价值信息的过程。

整个知识发现过程是由若干重要步骤组成,如图1-3所示,而数据挖掘仅仅是其中的一个重要步骤。

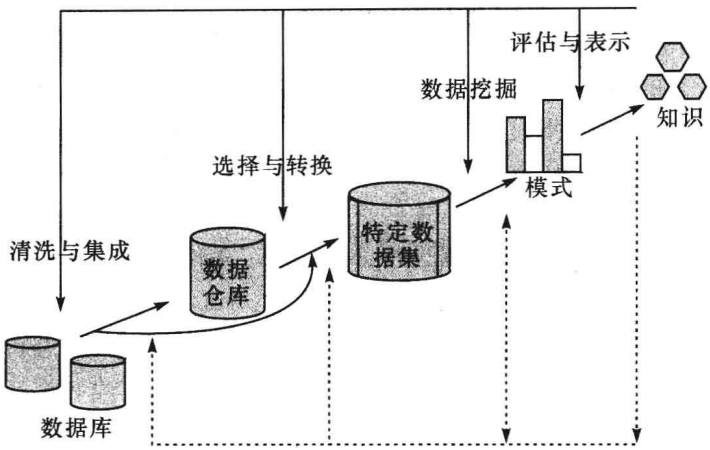

图1-3 知识挖掘全过程示意描述

整个知识挖掘的主要步骤有:

- 数据清洗,其作用就是清除数据噪声和与挖掘主题明显无关的数据;
- 数据集成,其作用就是将来自多数据源中的相关数据组合到一起;
- 数据转换,其作用就是将数据转换为易于进行数据挖掘的数据存储形式;
- 数据挖掘,它是知识挖掘的一个重要步骤,其作用就是利用智能方法挖掘数据模式或规律知识;
- 模式评估,其作用就是根据一定评估标准从挖掘结果筛选出有意义的模式知识;
- 知识表示,其作用就是利用可视化和知识表达技术,向用户展示所挖掘出的相关知识。

尽管数据挖掘仅仅是整个知识挖掘过程中的一个重要步骤,但由于目前工业界、媒体、数据库研究领域中,"数据挖掘"一词已被广泛使用并被普遍接受,因此在实际应用中对数据挖掘和知识发现这两个术语的应用往往不加区别。即数据挖掘就是一个从数据库、数据仓库或其他信息资源库的大量数据中发掘出有用的知识。

1.1.4 数据挖掘解决的商业问题

下面主要介绍数据挖掘技术可以解决的一些典型商业问题。

客户行为分析:如选择正确时间销售就是基于顾客生活周期模型来实施的。数据挖掘还可以对商品进行购物篮分析,分析哪些商品顾客最有希望一起购买,如被业界和商界传诵的经典案例——沃尔玛超市的"啤酒和尿布",就是数据挖掘透过数据找出人与物之间规律的典型。

客户流失分析：电信、银行和保险业如今正面临着激烈的竞争。平均每一个新的移动电话客户要花费电话公司超过 200 元的市场投资。每个公司应当留住尽可能多的客户。流失性分析能够帮助市场部经理了解客户流失的原因，还能够帮助改善与客户的关系，最终增加客户的忠诚度。

交叉销售：客户可能购买什么产品？对零售商来说，交叉销售是很重要的。许多零售商，特别是通过 Internet 销售的零售商，他们通过交叉销售来增加他们的销售量。例如在网上书店购买一本书，Web 站点会推荐一些相关数据，这些推荐的书就来自数据挖掘分析。

欺诈检测：这份保险存在欺诈吗？保险公司一天要处理成千上万个投诉，因此保险公司不可能调查每一个投诉。数据挖掘能够帮助他们鉴别哪些投诉很可能具有欺诈性。

风险管理：给某客户的一项贷款应该批准吗？这在银行业是很常见的问题。数据挖掘技术能用来评价客户的风险，帮助管理者对每一项贷款作出合适的决定。

客户细分：谁是我的客户？客户细分能帮助市场部经理了解客户个人信息的区别，基于客户细分采取适当的市场策略。

广告定位：针对特定用户如何使用适当的广告标语？Web 零售商和门户站点希望为他们的客户提供个性化广告的内容。通过使用客户的导航模式或者在线购买模式，这些站点利用数据挖掘解决方案，在客户的 Web 浏览器中显示个性化广告。

市场和趋势分析：利用检索数据仓库中近年来的销售数据，可预测出季节性、月销售量，对商品品种和库存的趋势进行分析，以确定下一个月的库存应该是多少。数据挖掘预测技术能够回答这种与时间相关的问题。

1.2　什么是商业智能

企业通过管理信息系统（MIS）快速收集和处理商业信息，通过企业资源计划系统（ERP）准确监控信息流，从而对企业经营的各个方面进行管理。这些系统除了本身的应用外，还积累了大量的数据，如来自业务系统的订单、库存、交易账目、客户和供应商资料，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，这是一笔宝贵的财富。信息系统应该具备把这些庞大的数据转化为知识，进而辅助企业经营决策，这就是商业智能（BI，Business Intelligence）。信息系统正在经历着“MIS→ERP→BI”的演变过程。

1.2.1　企业决策实现过程的信息需求

管理就是决策，决策需要信息。决策过程实际上就是一个信息输入、信息输出及信息反馈的循环过程。原来的决策支持系统，现在流行的商业智能，其目的都是为了辅助决策，让管理者从拍脑袋做决策到依据数据和事实做决策。这些依赖的数据和事实来源于两个方面：一个来源于竞争环境，这包括内部信息源（主要是存在于决策主体的经验信息）和外部信息源（主要是决策主体和咨询机构从社会上通过各种渠道获取的信息）；另一方面来源于企业多年信息化建设中积累的数据库信息。对于第一个方面，信息的非结构化特征决定了其

随意性和不确定性，这是决策理论中研究的问题；而对于第二个方面的信息，即使用存在数据库中的信息来辅助决策的问题，就是可以通过商业智能从技术上来得到很大程度的解决。

使用计算机辅助商业系统进行决策，需要经过五个步骤：

1. 提出决策信息请求（商务查询需求）。例如，现在某公司的决策层为了确定次年度在不同地区投资的力度，需要知道本年度和前五年华中、华北、华东和华南等区域的销售量和销售额，并且要有直观的图表来表达这些来源于数据库中的数据，这就为此决策发出了信息请求。

2. 调用商业智能应用程序。决策者可以直接使用原来的系统，如 ERP 来访问相关的销售数据，但是，这些数据往往分散在不同的数据库中，原来的系统也可能并没有提供十分富有个性化的查询需求。比如，在上述的决策中，原系统可能只提供了所有年度的销售数据，而不会具体到某一年甚至某一个月，那么这时候要满足决策信息需求就必须使用基于数据仓库技术的商业智能应用程序。

3. 基于已发布的模型、规则或是策略确定适当的决策。这一步是用计算机辅助决策的重要步骤，也是智能化体现的地方。决策（特别是结构化决策）是有一定规律的，这些规律可以从以往的决策过程或者从以往的数据中抽象获得，把抽象得到的这些规律放在经过特别组织的库中，可以构成模型库、规则库和策略库，智能决策可以在这些库的基础上进行。

4. 发布决策。决策最终取决于人的行为，计算机辅助了决策过程中信息的提取和规律性决策的结果，但最终的决策行为还是掌握在决策者自己的手中。

5. 采取行动。这是检验决策正确性的唯一途径。

商业智能系统建设的目标就是要为企业提供一个统一的分析平台，充分利用原有系统中积累的宝贵数据，对其进行深层次的发掘，并从不同的角度分析企业的各种业务指标和构建业务知识模型，进而满足决策的信息需求和实现通过技术辅助决策的功能。

1.2.2 企业信息化系统中的商业智能

商业智能的概念最早是 Gartner Group 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、联机分析、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的技术及其应用。

商业智能过程实际上包含两个层次。

第一个层次是在整合系统数据的基础上提供灵活的前端展现。例如，通过直方图等形式表现来自销售管理系统的地区销售情况报表，对复杂的计算则通过计算机的手段辅助完成。

商业智能的第二个层次是数据库中的知识发现。许多商业、政府和科学数据库的爆炸性增长已远远超出了传统方法能够解释和消化这些数据的能力，需要新一代的工具和技术对数据库进行自动和智能的分析，进而从数据中获取知识。这些工具和技术包括 OLAP，多种挖掘算法等。

因此，我们将商业智能定义为：将存储于各种商业信息系统中的数据转换成有用信息的技术。它允许用户通过查询和分析数据库，得出影响商业活动的关键因素，最终帮助用户做出更好、更合理的决策，其中的报表、在线分析和数据挖掘等工具从不同的层面帮助企业实现这个目标。