计算机科学丛书

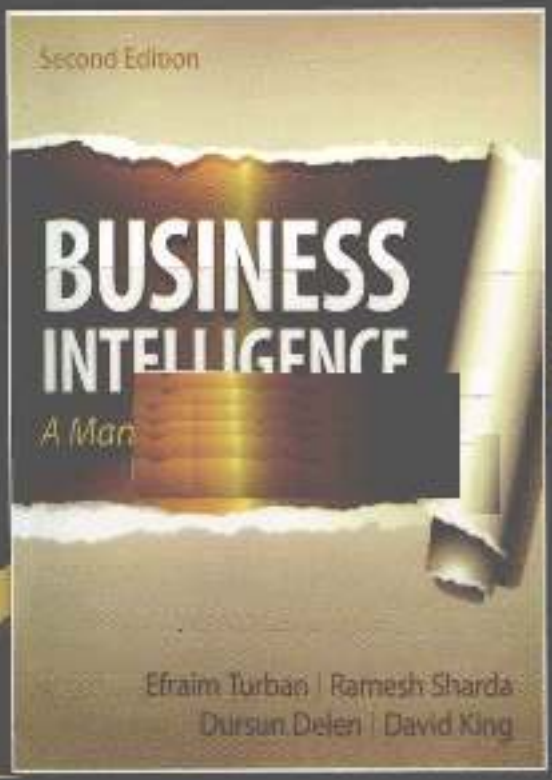
PEARSON

原书第2版

商务智能
管理视角

（美）Efraim Turban　Ramesh Sharda　Dursun Delen　David King 著
秦秋莉　姚家奕　王英 译

Business Intelligence
A Managerial Approach, Second Edition



机械工业出版社
China Machine Press

商务智能 管理视角（原书第2版）

Business Intelligence A Managerial Approach Second Edition

本书是电子商务领域著名学者Turban教授的又一力作，全书围绕于商务智能和为企业决策提供支持的商务分析，书中不仅介绍了传统的商务智能基本理论和应用，还介绍了当前商务智能涉及的新技术和新趋势，如文本挖掘、Web挖掘、社交网络和云计算等。

本书既可作为商务智能课程的教材，也可作为管理信息系统导论或商务战略的教材，同时还可以作为MBA技术管理课程的补充读物。

本书特色

- 管理导向：本书主要从管理视角详细介绍商务智能的应用和实施，同时也非常重视商务智能技术层面的应用，介绍了数据仓库、数据挖掘、数据可视化和人工神经网络等基本理论及其在管理决策方面的应用。
- 真实案例：给出大量全球大公司、小企业、政府机构和非营利组织的真实案例来生动形象地阐述相关概念和理论。
- 集成系统：本书强调那些支持企业和企业众多用户的系统，而不是孤立的基于互联网的商务智能系统。
- 热点研究：无线射频识别、云计算、社交网络、Web 2.0、虚拟世界等，本书分别从理论和应用角度对它们进行了阐述。

作者简介

Efraim Turban 执教于夏威夷大学，畅销教材作者、著名的电子商务学者，曾任美国加利福尼亚长岛州立大学管理信息系统教授，香港城市大学和香港科技大学访问教授，他曾撰写出版了十多部著作和大量论文，并担任多家杂志的编辑以及多家跨国公司和政府的顾问，是商务和企业计算机决策支持领域最多产的学者之一。

Ramesh Sharda 是俄克拉荷马州立大学西尔斯工商管理学院管理科学与信息系统的杰州支撑教授，信息系统研究所所长，ConocoPhillips公司技术管理主席。

Dursun Delen 是俄克拉荷马州立大学西尔斯工商管理学院管理科学与信息系统的副教授。

David King 有25年主持决策支持开发、性能管理和企业系统软件的经验，他还服务于许多工业咨询委员会和大学董事会。

PEARSON

客服热线：(010) 88378991, 88361066
购书热线：(010) 68326294, 88379649, 68995259
投稿热线：(010) 88379604
读者信箱：hzjsj@hzbook.com

华章网站 http://www.hzbook.com

网上购书：www.china-pub.com

PEARSON
www.pearson.com

ISBN 978-7-111-36727-7

9 787111 367277

定价：55.00元

计 算 机 科 学 丛 书

原书第2版

商务智能

管理视角

[美] Efraim Turban Ramesh Sharda Dursun Delen David King 著
常松 赵宏宏 王乘 译

Business Intelligence

A Managerial Approach Second Edition

机械工业出版社
China Machine Press

本书主要介绍商务智能、知识管理、数据挖掘和其他智能系统的基础和应用知识，并通过例子、产品、服务和练习，以及基于网络的问题讨论扩展了读者对网络世界的理解。第2版的改进主要集中在3个领域：数据挖掘、文本挖掘和Web挖掘、实施和新技术。

本书可以作为商务智能课程的教材，也可以作为管理信息系统简介或者商务战略的教材，还可以作为MBA技术管理课程的补充读物，或者注重管理视角的管理科学和管理信息系统项目的教材。

Authorized translation from the English language edition, entitled BUSINESS INTELLIGENCE, 2E, 9780136100669 by Efraim Turban, Ramesh Sharda, Dursun Delen, David King, published by Pearson Education, Inc., publishing as Prentice Hall, Copyright © 2011.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc.

CHINESE SIMPLIFIED language edition published by PEARSON EDUCATION ASIA LTD., and CHINA MACHINE PRESS Copyright © 2012.

本书中文简体字版由Pearson Education（培生教育出版集团）授权机械工业出版社在中华人民共和国境内（不包括中国台湾地区和中国香港、澳门特别行政区）独家出版发行。未经出版者书面许可，不得以任何方式抄袭、复制或节录本书中的任何部分。

本书封面贴有Pearson Education（培生教育出版集团）激光防伪标签，无标签者不得销售。

封底无防伪标均为盗版

版权所有，侵权必究

本书法律顾问　北京市展达律师事务所

本书版权登记号：图字：01-2010-6650

图书在版编目（CIP）数据

商务智能：管理视角（第2版）/（美）特班（Turban, E.）等著；秦秋莉，施家奕，王英译．—北京：机械工业出版社，2011.12

（计算机科学丛书）

书名原文：Business Intelligence：A Managerial Approach

ISBN 978-7-111-36727-7

Ⅰ.商…　Ⅱ.①特…　②秦…　③施…　④王…　Ⅲ.电子商务—教材　Ⅳ.F713.56

中国版本图书馆CIP数据核字（2011）第254566号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：盛思源

北京瑞德印刷有限公司印刷

2012年2月第1版第1次印刷

185mm×260mm·14.75印张

标准书号：ISBN 978-7-111-36727-7

定价：55.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展若六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅凝刻了研究范式，还揭示了学术传承，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到"出版要为教育服务"。自1998年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson等大师名家的一批经典作品，以"计算机科学丛书"为总称出版，供读者学习、研究及珍藏。大师石纹理的封面，也正体现了这套丛书的品位和格调。

"计算机科学丛书"的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，"计算机科学丛书"已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版"经典原版书库"作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com
电子邮件：hzjsj@hzbook.com
联系电话：(010) 88379604
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

华章教育
华章科技图书出版中心

译 者 序

Business Intelligence: A Managerial Approach, 2E

商务智能是一门新兴的边缘学科，近年来引起了学术界和产业界的广泛关注。商务智能是企业利用现代信息技术收集、管理和分析结构化和非结构化的商务数据和信息，创造和累计商务知识和见解，改善商务决策水平，采取有效的商务行动，完善各种商务流程，提升各方面商务绩效，增强综合竞争力的智慧和能力。

商务智能的出现是一个渐进、复杂的演进过程，而且仍处在发展中。20 世纪 90 年代后期，商务智能出现了突飞猛进的发展，越来越多的企业提出了他们对商务智能的需求，把商务智能作为帮助企业达到经营目标的一种有效手段。计算机界很多著名公司已经认识到商务智能巨大的发展潜力，纷纷加入到商务智能研究和软件开发的行列。商务智能技术已从决策支持系统开始，经历了数据仓库、联机分析、数据挖掘的发展，最后到达了可视化信息技术阶段。2010 年，从全球范围来看，商务智能领域并购不断，商务智能市场已经超过 ERP 和 CRM 市场成为最具增长潜力的领域。从中国市场来看，商务智能已经度过了一个从知到行、从概念到实践的阶段。当前金融业、电信业、零售业、服务业都在广泛地应用各种类型的商业智能工具，体验到了数据分析、报告、挖掘的力量，在经营分析、客户选择、绩效管理、运行预警方面得到很大的帮助。

本书作者是具有多年在 IT 相关领域的学术界和产业界工作的博士和专家，不仅承担了很多相关领域研究项目，在国际学术一流期刊发表了大量的学术论文，而且具有在跨国公司从事咨询工作的经验，因此可以从全方位角度向读者介绍商务智能的技术，展现如何从管理的视角去理解商务智能。

正如本书作者所言，本书可作为商务智能课程的教材，也可作为管理信息系统简介或商务战略的教材，同时还可以作为 MBA 技术管理课程的补充读物，或者是注重管理视角的管理科学和管理信息系统项目的教材。另一个目标是向实践管理者提供商务智能、知识管理、数据挖掘和其他智能系统的基础和应用。

本书第 2 版较第 1 版有了很大的改动，内容更加丰富、形式更加新颖、资料更加全面，同时配有生动及时的网站，相信可以满足广大读者的要求。

本书第 2 版包括 6 章。第 1 章和第 6 章、目录、前言和术语，由北京交通大学秦秋莉博士、刘教授负责翻译；第 2 章、第 3 章和第 5 章由北京交通大学姚家奕副教授负责翻译；第 4 章由北京交通大学王英老师负责翻译。全书的译文由秦秋莉负责统一定稿。

感谢机械工业出版社的编辑们，是他们的远见使得本书能够尽快与读者见面。

需要特别指出的是，2010 年我接手了本书的翻译工作。当时正值我将以访问学者身份前往美国亚利桑那大学 ELLER 商学院管理信息系统系人工智能实验室（AI LAB）访学交流之际。通过这一年的访学过程，我更好地理解了数据挖掘和商务智能，对翻译本教材具有很大的帮助。特别是 Chen Hsinchun 教授的 AI LAB 实验室研究课题和他教授的“数据挖掘”课程资料，Sudan Ram 教授的“商务智能”课程，Paulo Goes 教授的“商务智能中的数据挖掘”课程，Zhu Zhang 助理教授的“数据挖掘与网络控制”课程，对我深入了解数据挖掘和商务智能提供了很大的帮助，在此表示诚挚的感谢。

由于译者水平有限，译文中的错误和不当之处，敬请读者朋友指正。意见请发往 qlqin@sohu.com，我们将不胜感激。

希望读者喜欢这本译著，希望这本译著有助于进一步推动我国的商务智能研究与应用的深入开展。

秦秋莉

2011 年 11 月于美国亚利桑那大学人工智能实验室 AI LAB

基于计算机的决策支持的应用越来越广泛。许多公司，比如 IBM、Oracle 和 Microsoft 正在建立基于分析的组织单元，以便商业运作更加灵活有效。越来越多的懂计算机和网络的决策者使用更多的计算机工具来支持他们的工作。同时，消费者和组织在交互的过程中产生着不计其数的数据。这些数据存储能够用来开发和提升适当的产品、服务，促进消费者消费，优化组织内的运作。

本书的目的就是向读者介绍商务智能技术。在一些圈子里，商务智能（Business Intelligence，BI）也称为商务分析（business analytic）。我们不加区别地使用这两个术语。本书介绍了这些系统构建和使用中的基本技术和方式。

第 2 版的改进主要集中在 3 个领域：数据挖掘、文本和 Web 挖掘、实施和新技术。这一版尽管有这么多的变化，但仍然保留了综合性和用户友好性，这使得本书在市场上占据领先地位，最终呈现给读者最准确及时的知识，而这在别的教科书上是没有的。

第 2 版更新了什么

第 2 版增加了许多最新的内容，删掉了过时的内容。以下是几个重要的变化：

- **全部修订或新增的章节。**第 5 章（彻底进行了修订）以全面但易懂的方式研究了 2 个最流行的商业分析工具。这章提供了很多应用案例，使得主题有趣并且吸引读者。第 6 章是新增加的，这章调查了几个新的现象，这些现象正在或有可能改变商务技术和实践，它们分别是：无线射频识别（Radio Frequency Identification，RFID）、云计算、社交网络、Web 2.0、虚拟世界等。商务智能实施的重要主题和面向需求的计算战略也增加到了这一章中，同时第 6 章还更新了计算机化的决策支持对个人、组织和社会的影响。
- **流线型覆盖。**我们通过删除在线的格式文本来缩短教材篇幅，但是我们仍然保留了大量通用的内容。我们通过网站及时提供最新的内容和一些链接。另外，减少了每一章参考文献的数量。而且，我们对第 1 章中商务智能的介绍性的报道和数据挖掘实现了流水线式，这种综述形式能够让学生在学期开始提前准备如何以一个学术项目的方式进行学习和思考（指导老师可能也需要）。我们还删除了本章网上已有的内容并且合并了一些内容。
- **新的作者团队。**与上一版相比，本书增加了一名作者，还有一名作者的角色扩充了。前一版作者（Turban、Sharda、Aronson 和 King）构建了很好的内容，Ramesh Sharda 和 Dursun Delen 主要修订这一版。Ramesh 和 Dursun 主要工作在决策支持系统和数据挖掘领域，他们具有相关的行业和研究经验。
- **新的幻灯片图形。**尽管印刷版本中图形保留了以前版本中的内容，但是新内容中增加了新的图，所有的图都重新画成了彩色，并且可以从在线图库中获得，用于幻灯片演示。
- **更新及时的网站。**本书的使用者能够进入网站，网站包括与本书主题相关的新故事、软件、学习指南，甚至还包括 YouTube 的视频。
- **重新修订和更新的内容。**所有的章节都有新的开篇场景和结束案例，这些都是基于最近的真实应用故事和事件。除此之外，更新了贯穿本书的应用案例，包括最近的一个特殊技术/模型应用的例子。新的网站链接也增加到本书中。本书删掉了许多旧产品的链接和参考文献。最后，许多章节都有新的练习、网络练习、讨论题等。

第 2 版其他特殊改动归纳如下：

- 第 1 章　包括新的开篇故事和结尾案例，以及整章内流线型的材料。
- 第 2 章　包括数据仓库的新材料，包括大量的在线分析处理（Online Analytical Processing，OLAP）和多维数据模型。一款改编的可亲自动手的 MicroStrategy 软件演示也加入了本书。除此之外，我们还将架构和实施问题部分实现了流水线型结构。最后，还加入了数据仓库的未来部分。
- 第 3 章　是以上一版多个章节为基础合并而成的。除了更新的开篇场景、结束案例以及整篇关讨论这种流水线型结构，还包括了新的关键绩效指标（Key Performance Indicator，KPI）和运营指标、精益六西格玛、数据可视化、业务流程管理架构等部分。
- 第 4 章　采用体现标准的数据挖掘项目流程的方法，全面深度地展现数据挖掘的内容。与上一版相应的章节相比，本章重新撰写，使得读者更加容易使用和获得数据挖掘的信息。特别是延伸了文本和 Web 挖掘（有专门的章节），更重要的是扩展了数据挖掘的方法和方法论。本章还详细地描述了人工神经网络和它们在管理决策方面的使用，重点详细地描述了最流行的人工神经网络架构，阐述了它们在不同决策类型问题中使用的差别。这一章还加入了人工神经网络模型灵敏度分析的解释。
- 第 5 章和第 6 章　是新增章节。

我们保留了许多上一版不错的内容并更新了相关的内容，这些内容概括如下：

- **Teradata 大学网络（TUN）的连接**。许多章节都包括 TUN 的连接（teradatauniversitynetwork.com）。Teradata 网站（Teradata 学生网站 TSN，teradatastudent.network.com）主要包括了学生的作业。登录 TSN 网站允许学生阅读案例，观看网站研讨会，回答问题，查询资料等。
- **减少穿插的小板块，组织结构更合理**。小板块减少了 50% 以上，重要素材合并到正文中。现在只有两个小板块：应用案例和技术前沿。
- **软件支持**。TUN 网站提供了免费的软件支持。除此之外，网站还提供了软件的练习，我们的网站还包括了附加的一些链接。

目标和覆盖范围

目前组织能够很容易地使用内部网和因特网发布高价值的性能分析应用给全球的决策者。公司定期地开发分布式系统、内部网和外部网，用来方便地访问存储在许多地点的协作和通信数据。各种信息系统应用彼此集成起来，或者与其他基于网络的系统集成起来，有些集成甚至超越了组织的边界。由于管理者手头有很多精确的信息，所以他们能够更好地决策。

今天的决策支持工具使用网络来进行分析，它们使用图形用户界面，使得决策者可以利用熟悉的网络浏览器更加灵活、有效和容易地观看和处理数据和模型。企业信息、知识和其他高级系统将这种容易使用和阅读能力移植进了个人电脑和个人数字助理。管理者们通过使用一系列的无线掌上设备，包括移动电话和个人数字助理，通过电脑和网络进行沟通。这些设备使得管理者能够访问重要的信息和有用的工具，进行沟通和合作。数据仓库及其分析工具（例如在线分析处理、数据挖掘）极大地提高了穿过组织边界的信息接入和分析。

本书可作为商务智能课程的教材，也可作为管理信息系统简介或商务战略的教材，同时还可以作为 MBA 技术管理课程的补充读物，或者是注重管理视角的管理科学和管理信息系统项目的教材。本书的另一个目标是给实践管理者提供商务智能、知识管理、数据挖掘和其他智能系统的基础和应用。

这次修订版本的主题是商务智能和为企业决策提供支持的商务分析。除了传统的商务智能

应用外，这一版全篇通过例子、产品、服务和练习，以及基于网络的问题讨论扩展了读者对网络世界的理解。本书突出网络智能和网络分析，这些和电子商务及其他应用的商务智能和商务分析是平行的。本书有网站支持（pearsonhighered. com/turban），网站上提供了许多在线的文件。通过特殊的网站部分还可以链接到许多软件指导书。

特点

- **面向管理视角** 商务智能能够从两个视角来看：技术和管理。本书主要从第二个视角来看。本书包括许多关于商务智能的应用和实施的展示。然而，我们承认技术的重要性，因此每1章恰当地表述了技术概要。在学习指南中以博客链接到本书网站的形式提供一些技术细节。
- **面向真实世界** 大量有效的有关全世界大公司、小企业、政府、非营利组织的案例使得概念更加生动。这些案例给学生展示了商务智能的能力，成本和使用它的理由，以及现实中公司在内部使用商务智能的创新方法。
- **Teradata 大学网络（TUN）连接** TUN 是由 Teradata 资助的一个免费学习平台，是 NCR 的一个分支，其目标是帮助职员与其他商务智能领域的同仁之间进行学习、教学、沟通和合作。几百所大学和学院都参与和使用 TUN。Teradata 也支持学生平台（teradatastudentnetwork. com），上面包含了大量学习资源，例如案例、网络研讨会、学习指南、练习和资源的链接。本书与 TUN 主要通过给学生提供使用平台来完成各章不同类型的作业相链接。
- **大多是当前的主题** 有证据显示，本书提供的内容大多数是在 2008 年和 2009 年被引证的有关商务智能的主题。
- **集成系统** 相比其他教材重视孤立的基于互联网的商务智能系统，本书强调那些支持企业和企业众多用户的系统。
- **全球视角** 全球竞争、合作伙伴、贸易的重要性正在快速地增长，因此，全书提供的都是国际案例。
- **在线内容** 可以在线获取本书的附加材料。包括家庭作业的数据文件、许多报告、视频和软件的链接。
- **用户友好性** 本书不仅覆盖所有主要的商务智能主题，而且清晰、简单、结构组织合理。它提供了所有基本定义和逻辑概念支持。进一步说，本书易懂，并且充满有趣的真实世界案例，能激发读者的阅读兴趣。在每节结束还提供相关复习题，以使读者复习和消化新的材料。

补充包：Pearsonhighered. com/Turban

一个全面灵活的基于技术支持的补充包可以用来提高教学和学习经验。下面的教师指导补充材料会放在本书网站 pearsonhighered. com/turban 上，采用本书作为教材的教师需要联系所在地的培生代表处，申请下载权限：

- **指导手册** 指导手册包括整门课程和各章节的学习目标，各章结尾的问题和练习答案，教学建议（包括项目指导）。
- **幻灯片** 描述和构建了本书关键概念的幻灯片（可以下载获得）。
- **测试题和 TestGen 测试软件** 测试题包括判断正误、多项选择、填空、问答等类型。问题都是按照难度水平分类的。答案都引用了书的页码。测试题可以采用微软的 Word 格式和 Pearson TestGen 的计算机格式。TestGen 是一个全面的测试和评价软件工具，它允许指导老师创立和发布课程测试题，可以采用传统打印后分发的方法，也可以通过本地局

域网在线分发。TestGen 特征向导帮助用户在整个项目中移动，软件获得了全技术支持。

- **在线课程材料** Pearson Prentice Hall 支持本书使用者使用在线课程，通过黑板课程管理系统提供文件上传，用来测试、考问和其他辅助。请联系你所在地的培生代表处，获取特殊课程的进一步资料。

致谢

许多人帮助我们完成了本书和第1版出版后的修订。首先，我们感谢那些为本书和另一本决策支持系统教材——《Decision Support and Business Intelligence Systems》（第9版，Prentice Hall，2011）正式审稿的人们，他们给我们提供了很大帮助，他们是：

Ann Aksu	中部皮德蒙特社区学院
Bay Arinze	德雷克塞尔大学
Barjit Bose	新墨西哥大学
Kurt Engemann	约那学院
Badie Farah	东密歇根大学
Gary Farrar	哥伦比亚大学
Jerry Fjermestad	新泽西技术学院
Martin Groseman	布里奇沃特州立大学
Jahangir Karimi	科罗拉多大学丹佛分校
Huei Lee	东密歇根大学
Natalie Nazarenko	纽约州立大学弗雷多尼尔分校
Kala Chand Seal	洛约拉玛莉曼特大学
Roger Wilson	斯尔门特州立大学
Vincent Yu	密苏里大学理工学院
Fan Zhao	佛罗里达高尔夫海岸大学

其次，感谢那些为我们的文本和支撑材料提供资料的人。Barbara Wixom 为第1章撰写开场白，介绍了 Teradata 大学网络与本书的关系。Dan Power（Dssresources.com 和北爱荷华州大学）允许我们使用他的"虚拟世界"栏目上的信息。Deborrah C. Turban（菲律宾圣托马斯大学）为本书第6章做出了贡献。还要感谢亚利桑那州立大学的 Haluk Demirkan。

第三，感谢为第1版做出贡献的 Janine Aronson（乔治亚大学），他是我们的合作者，为数据仓库一章做出了贡献，Mike Goul（亚利桑那州立大学）为第1章做了贡献。T. P. Liang（台湾中山大学），为神经网络贡献了资料。

第四，感谢提供开发和演示软件的零售商，Acxiom（小岩城，阿肯色州）、California Scientific Software（内华达城，加利福尼亚州）、Catalyst Development 的 Gary Harwin（尤卡谷，加利福尼亚州）、Demandtec（圣卡罗，加利福尼亚州）、DS Group 公司（格林威治，康涅狄格州）、KDNuggtes.com 的 Gregory Piatetsky-Shapiro（莱斯斯维尔，佛罗里达州）、NeuroDimension 公司（盖恩斯维尔市，佛罗里达州）的 Gary Lynn、Promised Land Technologies（新罕布什尔，康涅狄格州）、Salford Systems（拉荷亚，加利福尼亚州）、Sense Network（纽约，纽约州）、Statsoft 公司的 Gary Miner（诺尔萨，俄克拉何马州）、Ward System Group 公司（弗雷德里克，马里兰州）、Wordtech Systems（奥林达，加利福尼亚州）。

第五，特别感谢 Teradata 大学网络及其执行总裁 Michael Goul；TUN 的副总裁 Barb Wixom；TUN 的创始人 Hugh Watson；项目经理 Susan Baxley；Teradata 的 Mary Gros，她是 Teradata 与学术团体的联络人。正是由于他们的鼓励促成这本书和 TUN 的联系，为本书提供了许多有用的材料。

第六，感谢许多帮助我们处理行政事务、编辑、校对、制图和做准备工作的人员，包括：Subramanian Rama Lyer（俄克拉荷马州立大学）、Mike Henry（俄克拉荷马州立大学）、Angie Jungermann（俄克拉荷马州立大学）、Brittany Solomon（俄克拉荷马州立大学）和 Ivan C. Saballos II（菲律宾的 DE LA SALLE LIPA）。Judy Lang 与我们所有的人进行合作，包括编辑，在第 1 版的制作过程指导我们。

最后，Pearson Prentice Hall 团队非常令人称许：Bob Horan 执行主编精心安排了这个项目，编辑项目经理 Kelly Loftus 给我们控制时间表，Shanthi Lakshmipathy 为我们的手稿做文字编辑工作，正是由于 Pearson Prentice Hall 的制作团队和 PreMedialGlobal 的工作人员，才使我们的手稿变成了一本书。

我们在此感谢以上所有人员的合作，没有他们的帮助，本书就不可能创作完成。

K. T.
R. S.
D. D.
D. K.

作者简介

Business Intelligence: A Managerial Approach, 2E

Efraim Turban（工商管理硕士、博士，加利福尼亚大学伯克利分校）是夏威夷大学太平洋管理信息系统研究院的访问学者，他曾经任职于多所大学，包括：香港城市大学、里海大学、佛罗里达国际大学、加利福尼亚州立大学长滩分校、东伊利诺伊大学、以及南加利福尼亚大学。Turban 博士在一流学术期刊上发了 100 多篇学术论文，例如《Management Science》、《MIS Quarterly》、《Decision Support Systems》。他还出版了 20 本书，包括：《Electronic Commerce：A Management Perspective》，《Information Technology for Management》。他还是世界各地许多公司的咨询师。Turban 博士目前的研究领域是基于 Web 的决策支持系统、智能代理在电子商务系统中的使用，以及全球电子商务的协同问题。

Ramesh Sharda（工商管理硕士、博士，美国威斯康星大学麦迪逊分校）是信息系统研究所（the Institute for Research in Information Systems，IRIS）的所长，康菲石油公司技术管理主席，俄克拉荷马州立大学（Oklahoman State University，OSU）西尔斯工商管理学院管理科学与信息系统的杰出贡献教授。Ramesh Sharada 博士在许多学术期刊上发表了 100 多篇与他的研究相关的论文，这些期刊包括：《Management Science》、《Information Systems Research》、《Decision Support Systems》、《INFORMS Tournal on Computing》、《Production Operations Management》、《Tournal of Management Information Systems》，《Interface》等。他是信息系统协会决策支持系统和知识管理特殊团体（SIGDSS）的共同创办人。Sharada 博士服务于多个编辑委员会，包括《INFORMS Tournal on Computing》、《Decision Support Systems》、《ACM Transactions on Management Information Systems》。他还是《Annals of Operations Research》的编辑，施普林格（Springer）集成信息系统系列和操作研究计算机科学界面丛书的编辑。他目前的研究兴趣是决策支持系统、协同应用、信息过载管理技术。他为许多组织进行咨询，并且在国际执行教育项目进行讲学。Sharada 博士也是 iTradeFair. com 的共同创办人，这是一家运营虚拟交易的公司。

Dursun Delen（博士，俄克拉荷马州立大学）是俄克拉荷马州立大学西尔斯工商管理学院管理科学与信息系统的副教授。他在 2001 任命为俄克拉荷马州立大学助理教授之前，他在得克萨斯州大学城的基于知识的系统公司工作，作为研究员工作了 5 年。期间，他领导了一系列与决策支持和其他信息系统相关的研究项目，资金是由联邦机构资助的，如：美国国防部（DoD）、美国国家航空和宇宙航行局（NASA）、美国国家标准技术研究所（NIST）、美国能源部（DOE）。他的研究成果发表在许多一流学术期刊上，包括：《Decision Support Systems》、《Communications of the ACM》，《Computers and Operations Research》、《Computers in Industry》、《Journal of Production Operations Management》、《Artificial Intelligence in Medicine》，《Expert Systems with Applications》等。最近他和 David Olson 教授出版了一本有关高级数据挖掘技术的书。他是《International Tournal of RF Technologies：Research and Applications》期刊的副主编，他还是《Journal of Information Technologies》、《International Tournal of Intelligent Information Technologies》、《Tournal of Emerging Technologies in Web Intelligence》、《International Tournational Tournal of Service Sciences》的编委会成员。他的研究和教学兴趣是决策支持系统、数据和文本挖掘、知识管理、商务智能，以及企业建模。

Dave King（博士）具有 25 年主持决策支持开发、性能管理、企业系统软件的经验。目前，他是位于亚利桑那州斯科茨代尔市的 JDA 软件公司新产品开发部的副部长。他在 Comshare 公司作为首席信息官和产品开发部高级副部长工作了几年后，于 2004 年加入 JDA。King 博士发表了许多论文和书籍，他是《Electronic Commerce：A Managerial Perspective》（Prentice Hall）一书的合作者。他还服务于许多工业咨询委员会和大学董事会，包括乔治亚大学的管理信息系统咨询委员会和亚利桑那州立大学专家政治咨询委员会。

目 录

出版者的话
译者序
前言
作者简介

第1章 商务智能简介 …… 1
开篇场景：Norfolk Southern 利用商务智能进行决策支持获取成功 …… 1
1.1 变化的商务环境和计算机化的决策支持 …… 3
1.2 商务智能框架 …… 4
1.2.1 BI 的定义 …… 5
1.2.2 BI 的历史 …… 5
1.2.3 BI 的架构 …… 6
1.2.4 BI 的形式 …… 8
1.2.5 BI 的好处 …… 8
1.2.6 事件驱动预警 …… 10
1.3 智能创造和使用与商务智能治理 …… 11
1.3.1 智能创造和使用的循环过程 …… 11
1.3.2 智能与信取 …… 12
1.4 交易处理和分析处理 …… 12
1.5 成功的 BI 实施 …… 13
1.5.1 典型的 BI 用户群体 …… 13
1.5.2 合适的计划及其与商业战略的一致性 …… 13
1.5.3 实时的，基于需求的 BI 是可达到的 …… 14
1.5.4 开发或获得 BI 系统 …… 14
1.5.5 理由和成本-利润分析 …… 14
1.5.6 隐私安全和保护 …… 15
1.5.7 系统集成和应用 …… 15
1.6 商务智能的主要工具和技术 …… 15
1.6.1 技术和工具 …… 15
1.6.2 选择 BI 供应商 …… 16
1.7 本书计划 …… 16
1.8 相关资源，链接和 Teradata 大学网络的连接 …… 17
1.8.1 资源和链接 …… 17
1.8.2 案例 …… 17
1.8.3 供应商、产品和演示 …… 17
1.8.4 期刊 …… 17
1.8.5 Teradata 大学网络的连接 …… 17
1.8.6 本书的网站 …… 18
本章重点 …… 18
关键术语 …… 18
讨论题 …… 18
练习 …… 19
本章结尾应用案例 …… 19
参考文献 …… 20

第2章 数据仓库 …… 21
开篇场景：DirecTV 的连续发展与实时数据仓库 …… 21
2.1 数据仓库的定义和概念 …… 23
2.1.1 什么是数据仓库 …… 23
2.1.2 数据仓库的特点 …… 23
2.1.3 数据集市 …… 24
2.1.4 业务数据存储 …… 24
2.1.5 企业数据仓库 …… 24
2.1.6 元数据 …… 26
2.2 数据仓库流程概述 …… 27
2.3 数据仓库架构 …… 28
2.3.1 可选的数据仓库架构 …… 30
2.3.2 哪种架构是最好的 …… 33
2.4 数据集成以及提取、转换和加载的过程 …… 34
2.4.1 数据集成 …… 34
2.4.2 提取、转换和加载 …… 36
2.5 数据仓库的开发 …… 38
2.5.1 数据仓库供应商 …… 39
2.5.2 数据仓库开发方法 …… 40
2.5.3 数据仓库开发的其他思考 …… 41
2.5.4 数据仓库中的数据表示 …… 42
2.5.5 数据仓库中的数据分析 …… 43
2.5.6 OLAP 与 OLTP …… 43
2.5.7 OLAP 操作 …… 43
2.6 数据仓库的实施问题 …… 46

2.7 实时数据仓库 …… 56
2.8 数据仓库管理系统、安全问题和未来发展趋势 …… 53
2.9 相关资源，链接和 Teradata 大学网络的连接 …… 55
2.9.1 资源和链接 …… 55
2.9.2 案例 …… 55
2.9.3 供应商、产品和演示 …… 56
2.9.4 期刊 …… 56
2.9.5 其他参考文献 …… 56
2.9.6 Teradata 大学网络的连接 …… 56
本章重点 …… 57
关键术语 …… 57
讨论题 …… 58
练习 …… 58
本章结尾应用案例 …… 60
参考文献 …… 61
第 3 章 业务绩效管理 …… 64
开篇场景：Harrah 公司加倍下注 …… 64
3.1 业务绩效管理概述 …… 66
3.1.1 BPM 定义 …… 67
3.1.2 比较 BPM 和 BI …… 67
3.2 制定战略：我们想到那里去 …… 68
3.2.1 战略规划 …… 69
3.2.2 战略差距 …… 70
3.3 计划：我们如何达到那里 …… 70
3.3.1 运营计划 …… 70
3.3.2 财务计划和预算 …… 71
3.4 监控：我们做得怎么样 …… 71
3.4.1 诊断控制系统 …… 72
3.4.2 差异分析的困难 …… 72
3.5 行动和调整：我们需要做什么不同的吗 …… 74
3.6 绩效评价 …… 76
3.6.1 KPI 和业务指标 …… 76
3.6.2 现在绩效评价系统存在的问题 …… 77
3.6.3 有效的绩效指标 …… 78
3.7 BPM 方法 …… 80
3.7.1 平衡记分卡 …… 81
3.7.2 六西格玛 …… 83
3.8 BPM 技术和应用 …… 87
3.8.1 BPM 架构 …… 87
3.8.2 商业 BPM 套件 …… 89
3.8.3 BPM 市场与 BI 平台市场对比 …… 90
3.9 绩效仪表盘和记分卡 …… 90
3.9.1 仪表盘和记分卡 …… 91
3.9.2 仪表盘设计 …… 91
3.9.3 仪表盘展示的内容 …… 92
3.9.4 数据可视化 …… 92
本章重点 …… 94
关键术语 …… 96
讨论题 …… 96
练习 …… 96
本章结尾应用案例 …… 98
参考文献 …… 100
第 4 章 商务智能中的数据挖掘 …… 102
开篇场景：数据挖掘来到好莱坞 …… 102
4.1 数据挖掘概念和定义 …… 104
4.1.1 定义、特征和好处 …… 106
4.1.2 数据挖掘的工作原理 …… 109
4.2 数据挖掘应用 …… 112
4.3 数据挖掘流程 …… 115
4.3.1 步骤 1：理解业务 …… 115
4.3.2 步骤 2：理解数据 …… 115
4.3.3 步骤 3：数据准备 …… 116
4.3.4 步骤 4：建模 …… 118
4.3.5 步骤 5：测试和评估 …… 118
4.3.6 步骤 6：部署 …… 119
4.3.7 其他标准化数据挖掘过程和方法 …… 120
4.4 数据挖掘方法 …… 121
4.4.1 分类 …… 121
4.4.2 分类模型正确性估算 …… 122
4.4.3 数据挖掘聚类分析 …… 127
4.4.4 关联规则挖掘 …… 128
4.5 数据挖掘中的人工神经网络 …… 130
4.5.1 人工神经网络的要素 …… 131
4.5.2 人工神经网络应用 …… 132
4.6 数据挖掘软件工具 …… 134
4.7 关于数据挖掘的一些谣言和谬误 …… 139
本章重点 …… 139
关键术语 …… 140
讨论题 …… 141
练习 …… 141

本章结尾应用案例 …… 144
参考文献 …… 145
第5章 文本挖掘与Web挖掘 …… 147
开篇场景：文本挖掘与安全和反恐 …… 147
5.1 文本挖掘的概念和定义 …… 149
5.2 自然语言处理 …… 151
5.3 文本挖掘应用 …… 155
5.3.1 市场营销应用 …… 155
5.3.2 安全应用 …… 155
5.3.3 生物医学应用 …… 157
5.3.4 学术应用 …… 158
5.4 文本挖掘过程 …… 159
5.4.1 任务1：确定素材 …… 160
5.4.2 任务2：创建文献术语矩阵 …… 161
5.4.3 任务3：提取知识 …… 162
5.5 文本挖掘工具 …… 166
5.5.1 商业软件工具 …… 166
5.5.2 免费软件工具 …… 166
5.6 Web挖掘概述 …… 167
5.7 Web内容挖掘和Web结构挖掘 …… 168
5.8 Web使用挖掘 …… 170
5.9 Web挖掘的成功实例 …… 171
本章重点 …… 174
关键术语 …… 175
讨论题 …… 175
练习 …… 175
本章结尾应用案例 …… 176
参考文献 …… 177
第6章 商务智能实施：整合和新兴趋势 …… 179
开篇场景：BI Eastern Mountain Sports 消费合作和生产力 …… 179
6.1 BI实施：概述 …… 181
6.1.1 BI实施因素 …… 181
6.1.2 BI实施中的管理问题 …… 182
6.2 BI和整合实施 …… 184
6.2.1 整合的类型 …… 184
6.2.2 为什么进行整合 …… 184
6.2.3 BI整合的水平 …… 185
6.2.4 嵌入式智能系统 …… 185
6.3 BI系统与数据库和其他企业系统的连接 …… 185
6.3.1 与数据库连接 …… 186
6.3.2 BI应用和后端系统的整合 …… 186
6.3.3 中间件 …… 187
6.4 面向需求的BI …… 188
6.4.1 传统BI的限制 …… 188
6.4.2 面向需求的选择 …… 188
6.4.3 关键特性和好处 …… 188
6.5 法律、隐私和道德问题 …… 190
6.5.1 法律问题 …… 190
6.5.2 隐私 …… 190
6.5.3 决策和支持中的道德问题 …… 191
6.6 BI中的新兴话题：概述 …… 192
6.7 Web 2.0创新 …… 193
6.7.1 Web 2.0的典型特征 …… 193
6.7.2 Web 2.0公司和新的商业模型 …… 194
6.8 在线社交网络：基础和示例 …… 194
6.8.1 定义和基本信息 …… 194
6.8.2 移动社交网络 …… 195
6.8.3 主要的社交网络服务：Facebook和Orkut …… 195
6.8.4 商业和企业社交网络的意义 …… 196
6.9 虚拟世界 …… 198
6.10 社交网络和BI：协同决策 …… 201
6.10.1 协同决策的概念 …… 202
6.10.2 虚拟团队决策中的协同 …… 202
6.11 RFID和新的BI应用机会 …… 203
6.12 现实挖掘 …… 206
关键术语 …… 208
讨论题 …… 209
练习 …… 209
本章结尾应用案例 …… 211
参考文献 …… 211
术语 …… 213

商务智能简介

学习目标

- 理解今天动荡的商务环境，描述组织机构如何在这样的环境（解决问题和探索商机）中生存并且取得成就。
- 理解管理决策中对计算机化支持的需要。
- 描述商务智能的方法论和相关概念，并将它们与决策支持系统相联系。
- 理解商务智能实施中存在的主要问题。

当今的商务环境在不断地改变并且变得越来越复杂。组织、个人、公众都面临着巨大的压力，这些压力迫使他们要对变化的环境做出快速的反应，同时还要求他们在运作方法上有创新精神。这就需要组织机构灵活并且频繁快速地在战略层、战术层、操作层做出决策。有些决策是非常复杂的，做出这样的决策需要大量相关的数据、信息和知识。在需要决策的框架中，处理这些数据就需要企业能够做出非常迅速、实时的行动，这通常需要某些计算机化的支持。

本书讲述了如何将商务智能作为一种计算机化的支持应用到管理决策中。在重点讲述针对决策支持的商务智能的理论和概念基础的同时，也涉及有效的商务工具和技术。本章一方面详细介绍了这些内容，另一方面也对本书的内容进行了概述。

开篇场景：Norfolk Southern 利用商务智能进行决策支持获取成功

在美国有4个大型的铁路货运公司，Norfolk Southern（以下简称为NS）是其中之一。每天，公司在东部的22个州、哥伦比亚区、安大略、加拿大有大约500辆货运火车在运行，运行的总里程有21 000英里。公司有超过260亿的固定资产和超过30 000名员工。

在一个多世纪的时间里，铁路行业一直是一个受到严格管制的行业。NS及其前身主要是依靠管理自己的成本来盈利的。管理者将主要的精力放在了对现有轨道车辆的优化利用上，依靠公司的固定资产来获得更多的成果。在1980年，行业开始部分放松管制，这就为公司之间合并提供了机会。与此同时，公司可以基于自己的服务来收费并和顾客订立合同。准时送货成了影响这个行业的重要因素。

在一段时间里，NS公司适应业界变化的对策是变成了一个“预定铁路”。这就意味着公司必须要制定一套固定的火车运行时刻表，为行驶在火车与码头之间的汽车制定一套固定节点。在这种情况下，管理者能够预测什么时候他们可以将货物送达客户。

NS一直用多种复杂的系统来经营自己的业务。然而，变成一个“预定铁路”就需要一个新的系统首先可以应用统计模型来决定最好的路线和连接点以使火车运输的表现最优。然后这个系统还要应用模型来制定可以指导铁路运行的计划。这些新系统叫做TOP（Thoroughbred Operating Plan），TOP是在2002年开始部署的。

NS意识到仅用TOP系统来管理铁路的运行是不够的，公司还要监测和衡量TOP计划的表现。NS的众多系统产生了成千上万的关于货物的记录、轨道车的信息、火车全球定位系统（Global Positioning System，GPS）的信息、火车燃料的信息、收入的信息、机组人员管理和历史

跟踪记录信息。不幸的是，公司在开发利用这些信息的同时还要冒着对系统的运行产生重大影响的风险。

早在1995年，公司投资引进了一个1TB的Teradata数据仓库，这个数据仓库是关于历史数据的仓库[1]。数据仓库是按照以下方式来组织的：一方面数据是很容易得到的（使用网络浏览器），另一方面数据可以用来做决策支持。数据仓库的数据来自公司的运行系统（也就是原始系统），并且一旦这些数据从原始系统移到数据库中，用户就可以得到数据并且使用这些数据而不必冒着影响系统运行的风险。

在2002年，数据仓库变成了TOP系统的关键组成部分。NS建立了一个TOP仪表盘应用，可将数据从数据仓库中抽出，用图描绘出与运输计划不符的性能，包括火车性能和连接点性能。这个应用使用可视化技术使区域经理能够更轻松地解释如此大量的数据（如何用这里有160 000个连接点遍布整个网络）。自从此应用实施以来，消失的连接点的数量已经减少了近60%。并且在过去的5年中，轨道车的运转周期已经减少了一整天，这意味着节省了数百万的资金。

NS拥有一套企业数据仓库（Enterprise Data Warehouse，EDW），这就意味着一旦数据被放到数据仓库中，那么整个公司都可以得到数据，而不仅仅是对某个应用。虽然火车和连接点的性能数据是供TOP使用的，但公司可以将这些数据用于其他类型的应用。例如，市场部门开发了一个叫做AccessNS的应用程序，这个应用程序是为NS的那些想要进入NS广泛的运输网络的客户建立的。这些客户想要知道运送他们货物的船只现在在哪儿，有时客户还想要了解一些历史信息，如：我的货物是从哪里来的？需要多长时间到达？在运送的过程中遇到过什么问题？

AccessNS允许来自8 000多个客户组织的14 500用户随时访问系统，获得预先确定的关于他们账户的客户报告。用户可以得到时时更新的信息，也可以查看过去3年的数据。AccessNS拥有预警功能和真正简单聚合（Really Simple Syndication，RSS）的跟踪能力。事实上，每天有4 500个报告发布给用户。AccessNS提供自主服务的特性，使得NS能够给客户提供他们想要的信息，并且减少了从事客户服务的员工数量。事实上，如果没有AccessNS系统，要维持现在的客户报告水平公司至少需要47人。

公司的各个部门，从工程与战略规划部到成本与人力资源部都在使用EDW系统。公司内部的一个很有意思的应用程序是由人力资源部开发的。最近，为了很好地满足NS公司超过30 000名员工的需要，该部门需要确定区域办公室所在地。通过将员工的人口统计信息（如邮政编码）与原本用在工程部的地理数据整合后，人力资源部就能够清楚地勾画出员工的人口密度可视化地图，这样就使区域服务办公地点选取的优化工作变得很容易了。

现在，NS公司的数据仓库系统已经发展成一个6TB的系统，该系统可以管理公司巨大的铁路和海运服务网的海量信息。NS利用这套数据仓库系统分析趋势、制定预测时间表、存档记录并为顾客的自主服务提供便利。这套数据库系统为超过3 000名的员工和140 000个外部客户和利益相关者提供信息。

NS是第一家提供自助服务商务智能的铁路企业，它的创新使得其他的铁路企业纷纷效仿。公司还是第一家可以为外部客户提供大量历史数据的公司。

---

[1] Dashboard是一个苹果公司Mac OS X V10.4 Tiger操作系统中的应用程序，用做称为"widget"的小型应用程序的执行基础。——译者注

开篇场景的问题

1. 在NS公司，信息系统是如何用来支持决策的？
2. 可视化应用可以获得哪种类型的信息？
3. AccessNS可以提供哪种类型的信息支持？
4. NS公司是如何将数据仓库应用到人力资源管理中的？
5. 同样的数据库是否可以应用到商务智能和优化的应用中？

从开篇场景中能够学到什么

这个开篇场景表明：即使在一个很成熟的行业，数据仓库技术仍然可以通过在公司的经营中获取更高的效率的方式使企业获得竞争的优势。确实，在许多情况下，这就是需要挖掘的前沿。从资产中获得更多的利润需要公司对其业务有及时详细的理解，同时具有使用信息做出更好决策的能力。在本书中可以看到许多这样的应用案例。

可以在Teradata大学网络（简称TUN）上获取更多关于案例的辅助资源，将在后续的章节中进行详细的叙述。这些包括其他的论文和一篇题为《Norfolk Southern Uses Teradata Warehouse to Support a Scheduled Railroad》的播客。

来源：Contributed by Professors Barbara Wixom(University of Virginia), Hugh Watson(University of Georgia;2005), and Jeff Hoffer(University of Dayton).

1.1 变化的商务环境和计算机化的决策支持

开篇场景说明了一个全球化的公司是如何在一个成熟而竞争激烈的市场中取得成就的。公司正在飞速地发展它们业务的计算机化支持。为了理解公司为什么如此地青睐计算机化支持，包括商务智能，我们建立一个商业压力-反应-支持模型来说明这个问题，如图1-1所示。

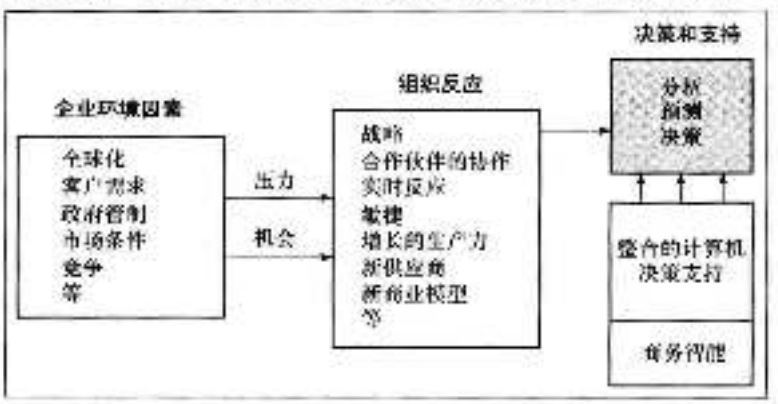


图1-1 商业压力-反应-支持模型

商业压力-反应-支持模型

商业压力-反应-支持模型就如其名字一样包括3个部分：源于今天商业环境的商业压力，为克服压力公司采取的反应（或者是对环境中有利机会的利用），对环境监测提供方便并且能够提高组织反应能力的计算机化支持。

商业环境 组织今天面对的环境变得愈加复杂。这种复杂性一方面创造了机会，另一方面也产生了问题。以全球化为例，今天，你可以轻松地在世界上的许多国家找到供应商和顾客，这就意味着你可以买到更加便宜的原料，并将产品和服务卖给更多的顾客，存在很多机会。然而全球化意味着更多和更激烈的竞争。商业环境因素可以分成主要的4种类型：市场因素、顾客需求

因素、技术因素、社会因素。表1-1对这4类因素进行了总结。

值得注意的是这些因素的影响会随着时间变得更强，这会导致更大的压力和更激烈的竞争。除此之外，为了增加利润，组织和组织内的部门会面临来自更少的预算和上层管理者要求提高绩效和利润的更大压力。在这种环境下，经理们必须以更快的速度来应对，即创新和敏捷。我们来看看他们是如何做到的。

表1-1 给组织制造压力的商业环境因素

因素	描述
市场因素	激烈竞争 扩大的全球市场 网络上日益增加的电子市场 创新的需求手段 支持信息技术外包的机会 实时、基于需求交易的需要
顾客需求	定制产品 追求高质量、产品多样化、快速物流 顾客要求增大和缺少忠诚
技术	更多创新、新产品和服务 不断增加的投资率 不断增加的信息超载 社交网络、Web 2.0等
社会环境	不断增加的政府管理和政治管制 劳动力更多样化、老龄化并包括了更多的女性 关于安全和恐怖袭击的担忧 萨班斯法案的必要性和其他报告相关立法 公司日益增加的社会责任 对于持续性更大的关注

组织反应：反应性、预期性、适应性、前瞻性 不论私人组织还是公共组织都清楚地意识到了今天的商业环境和压力。它们用不同的行动来应对压力。例如Vodafone New Zealand（Krivda，2008年）利用商务智能来改善沟通，在保持现有顾客和在现有顾客中取得更多收益方面支持管理人员（参看本章末的应用案例）。经理们可能采取其他的措施，包括

- 采用战略规划
- 应用新的和创新性的商业模型
- 业务流程再造
- 参加商业联盟
- 改善企业信息系统
- 改善与合作伙伴的关系
- 鼓励创新和创造性
- 完善客户服务和关系
- 转向电子商务（e-商务）
- 转向订单驱动式生产和面向需求的生产和服务
- 利用新的IT技术改善沟通、数据获取（信息发现）和合作
- 面对竞争对手的行动快速做出反应（例如在定价、促销、新产品和服务方面）
- 白领雇员的许多工作的自动化
- 某些决策过程自动化，特别是通过雇佣专业分析人员来提高决策质量

即使不是全部，许多这些反应也都需要计算机化的支持。这些和其他的行为是经常需要计算机化的决策系统支持的。

缩小战略差距 计算机化的决策支持系统的一个主要目标就是为缩小现在组织的绩效和它所期望的绩效之间的差距提供便利，通常表述为组织的使命、方向、目标和达到目标的战略。为了理解为什么需要计算机化支持和它是如何被提供的，让我们来回顾一下商务智能的框架和它在决策支持中的应用。

1.1 节复习题

1. 列举商业压力-反应-支持模型的组成并解释这个模型。
2. 什么是当今商业环境中最主要的因素？
3. 组织在面对当今商业环境时所做出的行动是什么？

1.2 商务智能框架

决策支持概念正在被越来越多提供决策支持工具和方法的供应商以不同的名称逐渐实现。

随着企业规模系统的成长，经理们可以更方便地得到用户友好的报告，这些报告能够使他们更快地做出决策。这些系统，通常叫做高级管理人员信息系统（Executive Information System，EIS），可以提供可视化、预警和绩效考核能力。到了2006年，概括地定义，主要的商业化的产品和服务被称为商务智能（Business Intelligence，BI）。

1.2.1 BI的定义

商务智能是个概括性术语，它包括了构架、工具、数据库、分析工具、应用和方法论。这是一个无内容表述，所以对于不同的人意味着不同的含义。关于商务智能的一部分迷惑在于与它有关的一些缩略语和流行语的混淆（例如业务绩效管理）。商务智能的主要目标就是实现数据的交互（有时候是实时的），实现对数据的操作，使管理者和分析员能够实施合理的分析。通过对历史和现有数据、位置、性能的分析，决策者可以得到有价值的深刻理解，这些研究使决策者可以做出更好的决策。BI的过程是以将数据转换成信息为基础，然后做出决策，最终开始行动。

1.2.2 BI的历史

BI这个词是由Gartner Group在20世纪90年代中期提出来的。然而，这个概念出现的更早，它可以溯源到20世纪70年代的管理信息系统（Management Information System，MIS）的报告系统。在那个时期，报告系统是静态的、二维的，没有分析能力。在20世纪80年代早期，高级管理人员信息系统（Executive Information System，EIS）的概念出现了。这一概念将计算机化支持系统扩展到了高层经理和管理人员。这些能力有动态的多维度的报告、预测和预报、趋势分析、深入到细节、状态访问和关键成功因素（CSF）分析。到了20世纪90年代中期，这些特征才出现在一系列的商业化产品中。然后，相同的能力和具有一些新功能的产品被称为BI。今天，一个好的基于BI的企业信息系统包含了管理人员所需要的所有信息。所以，最初的EIS概念发展成了BI。到了2005年，BI系统开始包括了人工智能和强大的分析能力。图1-2展示了在BI系统中可能包含的各种各样的工具和技术。它也展示了BI的发展历程。图1-2中提供的各种工具为

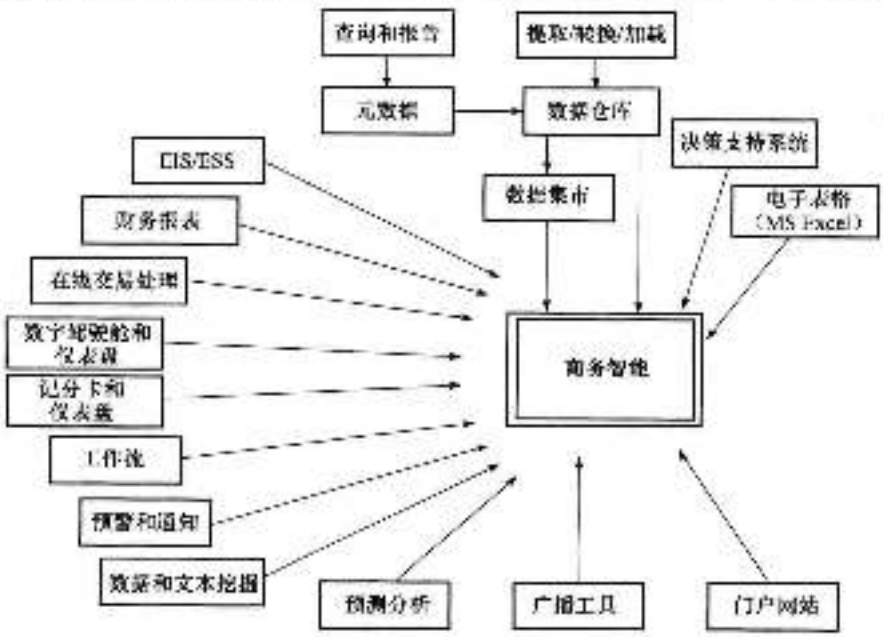


图1-2 BI的发展

BI提供了各种能力。最复杂的BI系统包含了大多数的这些工具，其他一些专业的BI系统只包含了一部分这些工具。第2章到第6章将更详细地讲述这些功能。

1.2.3 BI的架构

BI系统通常包括4个主要的部分：带有数据源的数据仓库；商业分析（一个用于挖掘和操作分析数据仓库中的数据的工具集）；用来监测和分析绩效的业务绩效管理（Business Performance Management，BPM）；用户界面（如仪表盘）。这4个部分之间的关系如图1-3中所示。第2章至第6章将详细讲述这些内容。

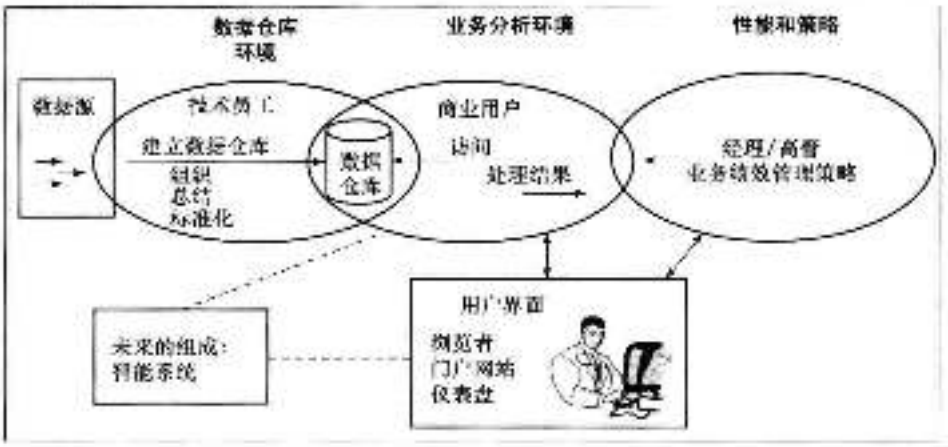


图1-3 BI的高层架构

来源：Based on W. Eckerson, *Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions*. The Data Warehousing Institute, Seattle, WA, 2003, p. 32, Illustration 5.

值得注意的是数据仓库环境主要是技术人员负责，然而分析环境（也称做商业分析）是属于商业用户范围的。任何用户都可以通过用户界面（如浏览器）连接系统。高层管理者也会用到BPM组件和仪表盘。一些商业分析和用户界面工具将在1.7节和第4章和第5章进行简要的叙述。

数据仓库 数据仓库和一些它的衍生物是任何一个大中型BI系统的基石。最初，数据仓库仅仅包括那些总结和组织好的历史数据，所以最终的用户能够容易地查看和处理数据和信息。今天，有些数据仓库也包括现有的数据，所以它们能有提供实时的决策支持（见第2章）。

商业分析 最终用户能够通过一系列的工具和技术利用数据仓库中的信息。这些工具和技术可以分成如下的两大类：

1. **报告和查询** 商业分析包括静态和动态报告、所有类型的查询、信息的发现、多维度视图、深入到细节等。这些将在第3章中进行讲述。这些报告与BPM相关（将在下面介绍）。

2. **数据、文本、Web挖掘和其他复杂的数学和统计工具** 数据挖掘（在第2章至第6章中进行讲述）是一个在大型数据库或是数据仓库中，对于未知关系和信息的寻找过程，这个过程要应用智能工具，例如神经网络、预测分析技术，或是高级统计工具（见第4章）。就像在第5章中讨论的那样，挖掘可以在网络中进行，也可在文本数据中进行。下面就是两个有用的数据挖掘的实际应用。

示例1

对于新产品或服务的成功预测，对于任何一个企业来说都是一个挑战，对于像电影、音乐之类的娱乐产品预测成功是个特别困难的问题。Epagogix专门从事通过对电影剧本的详细分析来预测电影的成功与否。就像Davenport和Harris在2009年报道的，2007年他预测了《Lucky you》这

影电影将会有一个很不错的700万美金的票房，虽然这部电影包含了著名的明星和著名的导演，并且花费了5 000万美元的巨资。这部电影最终只是收到了600万美元的票房。基于喜好和评论的模型也会被类似诸如Netflix的供应商用来预测哪个电影或是音乐会受欢迎。（基于第4章的研究，还会涉及另外一个数据挖掘用于电影票房成功的应用。）

示例2

澳大利亚国家银行应用数据挖掘来帮助进行营销预测。这些工具被用来提取和分析存储在银行的Oracle数据库中的数据。这家应用集中于评估客户对于主动权是如何影响银行的底线。数据挖掘技术被用来从历史数据中产生市场分析模型。银行认为在一个竞争日益激烈的金融服务市场中，主动权是保持边缘的关键。

应用案例1.1描述了用于数据挖掘工具的另一个BI技术——聚类分析。

应用案例1.1 选址

Hoyt Hihland Partners是一个营销智能公司，它主要是帮助卫生保健提供者应对不断增加的患者人数。这个公司也帮助它们决定哪里是它们提供卫生保健业务的最佳选址。Hoyt Hihland正在与一家紧急护理诊所合作。这家诊所面临着来自其他紧急治疗诊所和方便治疗诊所的非常激烈的竞争。为了增加收入，这个诊所需要决定它是否应该改变位置或是改变营销业务。为了帮助诊所做出决定，Hoyt Hihland确定应用Acxiom的PersonicX系统来确定诊所的目标顾客集中的地区。

Acxiom的PersonicX系统将每个美国家庭分成70个部分和21个年龄阶段。地点是基于独特的消费行为和人口特征。信息包括能够概括出重要市场的消费者行为、态度、消费地址特征的调查数据。Hoyt Hihland利用PersonicX来决定哪个集群出现在紧急治疗诊所的数据库中，哪个集群表现出很高的投资回报（Return-on-Investment，ROI）潜质。

应用软件的地理分析能力，Hoyt Hihland发现诊所80%的患者住在诊所现有位置5英里的半径范围内。它还发现年轻的家庭在数据库中很有代表性，而单身和老年人代表性不强。另外，它发现离诊所的邻近程度是在一个紧急诊所选址中的首要条件。这一分析结果帮助诊所认识到最好的行动决策是改变它的营销侧重点而不是改变诊所的位置。今天，诊所将它的营销重点放在了住在距诊所半径5英里范围内的年轻家庭的患者。

来源："Location, Location, Location," Acxiom, acxiom.com (accessed March 26, 2009).

业务绩效管理 也称做企业绩效管理（Corporate Performance Management，CPM），BPM是一个正在兴起的一些应用和方法论的组合，包括演进的BI结构和它的核心工具。BPM通过引进管理和反馈的概念包含了监管、测量、销售对比、利润成本、利润率和其他的一些绩效指标。它包含了将诸如计划和预测作为中心内容的商业战略流程。与传统的能够将底层数据抽取出来变换成信息的DSS、EIS和BI相比较，BPM能够实现一个公司范围内从上到下执行的战略。BPM是第3章的主要内容，通常与平衡记分法和仪表盘结合使用。

用户界面：仪表盘和其他信息广播工具 仪表盘（与自动仪表盘类似）提供一个综合的公司绩效指挥（也称为关键绩效指标）、趋势和例外的可视化视图。它们整合了不同商业地区的信息。仪表盘能够显示与预期的度量标准对比后的真实的绩效图表。从仪表盘能够一眼就看出组织机构运行得是否健康。除了仪表盘外，其他能够发布信息的工具有企业门户，数字驾驶舱以及其他的可视化的工具（见第3章）。许多从多维度立体化的呈现方式到虚拟现实的可视化工具，

都是 BI 的组成部分。BI 是从 EIS 发展过来的，所以许多提供给高级管理人员的可视化辅助手段被移植到 BI 软件中。而且，诸如地理信息系统（Geographical Information System，GIS）的技术在决策支持系统中占有越来越重要的地位。

1.2.4 BI 的形式

BI 的结构取决于它的应用。MicroStrategy Corp 区分了 5 种 BI 形式，并且给每种形式提供了特殊的工具。这 5 种形式是报告交付和预警、企业报告（用仪表盘和积分卡）、立体分析（被称做是切片和骰子分析）、特定查询、统计和数据挖掘。第 2 章将会学习更多的关于 MicroStrategy 的软件。

1.2.5 BI 的好处

就像在开篇故事中提出的，BI 对于一个公司主要的好处就是能够在需要的时候提供精确的信息，包括对于一个公司及其部门绩效的实时视图。这样的信息对于战略规划，甚至是生存等所有类型的决策都是必需的。Thompson（2004）也表明 BI 最常用的应用领域就是常规报告、销售和营销分析、计划和预测、财务合并、法定报告、预算和营利性分析。

组织机构为了提高它们的商业运作，被迫要捕获、理解、利用它们的数据来支持决策的制定。现在立法和法规（例如 2002 年萨班斯法案）要求企业领导人将他们的商业过程文档化，答发他们依赖的合法信息，并向相关利益者报告。另外，商业循环周期变得越来越短、快速和信息化，更好的决策制定是竞争非常必要的。管理者们需要在合适的地点、合适的时候获得正确的信息。组织机构必须更加灵活地工作。所以，一点也不奇怪，越来越多的组织开始拥护支持 BI。开始的引例讨论了 Norfolk Southern 公司 BI 的成功案例。在接下来的第 2 章至第 6 章中将有更多的关于 BI 成功的案例和这些成功的基础。表 1-2 列出了关于 BI 典型应用的例子。在应用案例 1.2 中描述了一个采用预测分析工具进行数据挖掘的案例（将在第 4 章至第 6 章讲述）。

表 1-2 BI 分析应用的商业价值

分析应用	商业问题	商业价值
顾客分割	我的顾客属于市场的什么部分？顾客们有什么特征	能够获得更高满意度和保留的个性化的客户关系
购买倾向	哪些顾客最有可能对公司的促销做出反应	基于顾客的需求锁定顾客 并增加顾客对于公司产品的忠诚度 同时，通过关注最有可能购买的顾客来提高促销利润
顾客盈利	公司顾客的终生利益是什么	基于顾客的终生利益制定个体商业交互
欺诈检测	公司如何识别哪些交易有可能是欺骗公司的	快速识别欺诈并立即采取措施以减少损失
顾客牵拐	哪些顾客有离去的风险	阻止高价值客户流失和放弃低价值顾客
渠道优化	满足每部分顾客的最好的渠道是什么	基于顾客喜好和公司管理成本的需要与客户接触

来源：A. Ziama and J. Kasher(2004), *Data Mining Primer for the Data Warehousing Professional*. Dayton, OH: Teradata.

应用案例 1.2 Alltel Wireless：在准确的时间将准确的信息送给正确的客户

2006 年 4 月，Alltel Wireless（现在已经与 Verizon 合并）发起了一项“我的圈子”的活动并且在手机产业发起了一场革命。第一次，顾客可以免费在任何网络给任何一个 10 位号码不受限制地打电话。为了在越来越多的无线接入率的时代巩固“我的圈子”活动的效果，Alltel 发现需要一个集中的、关注数据的方案来增加新客户的数量并且增进和现有顾客的关系。

通过 Acxiom 的 PersonicX 部门系统（acxiom.com），Alltel 能够通过顾客特殊的消费行为和人口特征将关于美国家庭的信息进行分类。这个功能增加了 Alltel 的顾客，对更好地理解购买行为和顾客订阅周期事件提供了更丰富的展望数据。利用这些分析技术，Alltel 可以通知特定的顾客群体关于一些可以增加客户无线体验的机会，例如文本信息和铃声下载。另外，Alltel 现在能够定位那些很有可能通过低成本网络和客户中心渠道激活订阅的新客户。

应用 Acxiom 的 BI 软件，通过自动化 Alltel 的客户生命周期管理，Alltel 一年可以管理超过 300 个直接的营销主动权；增加了 265% 的客户人数；投资回报率增加 133%；同时创造了超过 3 000 万美元的正在进行的业务。

来源："Customer Lifecycle Management," Acxiom, acxiom.com (accessed March 26, 2009).

自动化决策制定 一个比较新的支持决策制定的方法是**自动决策系统**（Automated Decision System，ADS），有时被称做决策自动化系统（Decision Automation System，DAS）（参见 Davenport and Harris，2005）；ADS 是一个基于规则的系统，通常在一个功能领域（例如金融业，制造业），针对某个行业特定的重复性的管理问题提供解决方案（例如对于一项贷款请求的批准与拒绝，决定商店中的某个商品的价格）。应用案例 1.3 就展示了一个应用 ADS 来解决组织机构都面对如何给产品或是服务定价的问题。

ADS 最初出现在航空产业，在这个产业，ADS 被称做收入（或者产出）管理（或者收益优化）系统。航空公司应用这套系统在实际需求的基础上动态地为机票定价。今天，许多服务行业都应用相似的价格模型。与能够通过模型为通用的结构化问题（例如资源分配，确定存货水平）提供解决方案的管理科学的方法相比，ADS 能够提供基于规则的解决方案。以下是一些商业规则："如果从洛杉矶到纽约航班的 70% 的座位已经在出发前 3 天售出，那么就给非商务旅客一个 x 的折扣"，"如果一个申请者拥有一套房子和每年超过 100 000 美元的改造费，那么就提供一个 10 000 美元的信贷额度"，"如果一件产品的单价是 2 000 美元，并且公司每年只是购买一次，那么采购代理不需要特别批准。"这些通过经验或是通过数据挖掘得到的规则，能够和数学模型结合使用形成解决方案，自动和快速地提供给问题（例如：基于提供的信息和需要证实的科目，你能够被我们大学录取），或者能够提供给做最终决策的人（见图 1-4）。ADS 努力在业务规则的基础上对重复的决策实现高度自动化（为了使计算机化的成本更加合理）。ADS 很适合一线的业务人员，他们能够在线看见客户信息，必须经常地做出快速的决策。关于 ADS 的更多的信息请参看 Davenport and Harris（2005）。

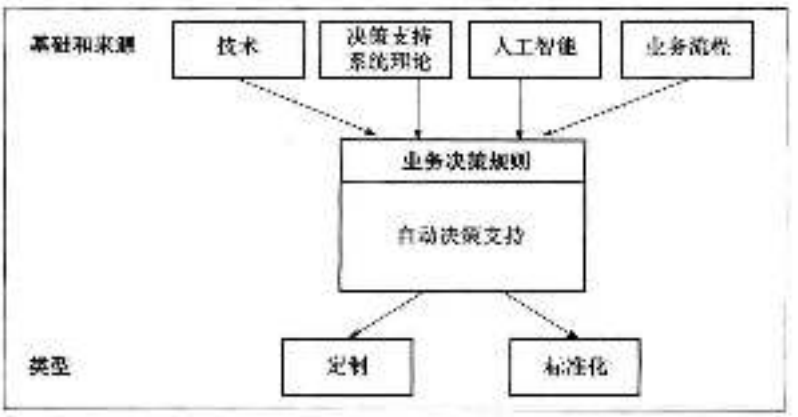


图 1-4 自动决策框架

应用案例1.3 Giant Food Stores为整个商店定价

Giant Food Stores, LLC，是一个总部设在宾夕法尼亚州卡莱尔的区域性美国连锁超市。它实行一个有限品种的每天低价政策，这一政策应用在商店内的大多数产品上。公司拥有一个30年之久的定价和促销系统，这个系统非常地耗费人力并且很难跟上快节奏的零售市场的定价决策要求。这一系统也限制了公司实施更多复杂定价策略的能力。

Giant Foods 想通过一套有限的定价规则（零售业定价规则可能包括国有品牌和私有品牌之间的关系，尺寸大小之间的关系，纯度数字例如9之间的关系）更加持续一致地实施它的定价策略。过去，许多这些规则是写在纸上的，另外一些是在相关人员的脑子中，有些规则文件写得不好，以至于其他人很难理解和确保一致性。在价格到达货架之前，公司也没有可靠的方法去预测规则改变的影响。

Giant Food 与 DemandTec 合作部署了一套用于定价决策的系统。这一系统能够处理大量的销售点和模型的竞争数据并且能够预测消费者需求，能够自动化和流线化复杂的基于规则的定价计划。这个系统能够在不增加员工的情况下处理大量的价格变动。此系统允许 Giant Food 用自然语言来编写价格规则，而不是通过技术员来进行。系统还具有预测功能。这些能力使得 Giant Food 能够在价格到达货架之前预测价格变动和新的促销带来的影响。Giant Food 决定在整个商店连锁链实施这套系统。

这套系统使得 Giant Food 在定价方面变得更加的灵活。它现在能够在一周之内对有竞争的价格变动或是供应商成本的变动做出反应，而不是在得到资源的时候。因为不需要因为价格的变动而增加员工，所以 Giant Food 的生产能力成倍地增加。Giant Food 集中精力在满足客户需求的同时持续盈利和维持它的价格形象。

来源："Giant Food Stores Prices the Entire Store with DemandTec," DemandTec, demandtec.com.（accessed March 26, 2009）.

1.2.6 事件驱动预警

ADS的一个例子就是事件驱动预警，它是一个警告或是当预设的或者不寻常的事件发生时被激活的行动。例如，信用卡公司已经建立了广泛的预测分析模型来确定可能的诈骗事件，当不正常的活动被注意到时（例如当一个用户没有这样的交易历史记录，而大宗购买发生在异常或是境外的时候），能够自动地提醒信用卡用户核对交易。如果一个客户存了一大笔钱在银行，银行可能自动地提供一个更高利率的存款证（Certificate of Deposit，CD）或者投资。这样的预警同样应用在基于其生购买完成时产生促销。当然，预警同样通过 BPM 的仪表盘呈现给适当的经理。当一些偏离其结果的显著事件存在时，这些经理有责任监管这些行为指标。

1.2 节复习题

1. BI 的定义。
2. 列出并描述 BI 的主要组成部分。
3. 识别 BI 的典型应用。
4. 列举 ADS 的例子。
5. 列举关于事件驱动预警的例子。

1.3 智能创造和使用与商务智能治理

1.3.1 智能创造和使用的循环过程

数据仓库和BI初始化典型地遵循着一个与军事智能初始化非常相似的过程。实际上，BI的实施者们经常遵循如图1-5所描述的国家安全模型。这个过程是通过一系列相互关联的步骤来实现循环的。分析是将未经加工的信息转变成决策支持信息的主要步骤。然而，精确或可靠的分析是不可能的，除非在该循环中的其他步骤被合理地实施。过程和实施步骤的细节参见Krizan（1999）和第4章。

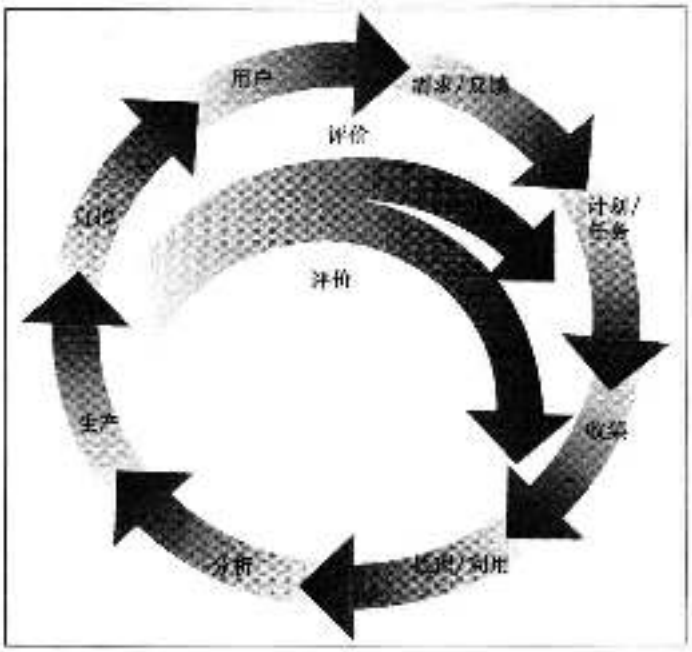


图1-5 智能生成和使用流程

来源：L. Krizan, Intelligence Essentials for Everyone. Washington DC: Joint Military Intelligence College (occasional paper number six) Department of Defense, p. 6.

一旦安装数据仓库，适用的智能创造过程就从识别和确认特定BI项目开始了。对于每个在投资组合中潜在的BI项目，应用投资回报（Return on Investment，ROI）和制有成本措施来估计成本效益率是非常重要的。这意味着每个项目都要经过通常阶段所需费用和维持商业用户应用费用的检查。另外，效益的评估需要涉及对最终用户决策制定影响的检查，包括反应现金流加速效益措施的检查。一些组织将项目优化过程称做商务智能治理（BI governance）的形式（Matney and Larson，2004）。一个主要管理问题就是在BI项目优化中谁应该是决策的制定者。商务智能治理的两个关键组成部分是：（1）功能区领导和产品或是服务区领导（中间层）之间的合作关系；（2）潜在客户和提供者（业务代表和IT方代表）之间的关系。中间层可以纵观整个组织来确保项目的优先次序反映整个业务的需求；他们确保项目在一个地区的实施相对于另一个地区不是局部最优化。顾客可以为项目中产生的智能的潜在用处提供深入的理解，同时供应商在反映交付现实的立场上是非常重要的。商务智能治理团队的典型问题就是：（1）制定项目分类（投资、商业机会、战略、强制等）；（2）定义项目选择的标准；（3）决定和设置管理项目风险的框架；（4）管理和平衡项目内部依赖关系；（5）持续管理和调整投资组合的构成。

1.3.2 智能与窃取

虽然许多人认为智能这个词听起来就像是一个有关间谍秘密运作的缩写，致力于偷取公司秘密或是政府的CIA，但是这与事实相去甚远。虽然这种间谍活动当然会发生，但我们感兴趣的是现代公司如何公正合法地去收集他们的客户、商业环境、相关利益者、业务流程、竞争对手和一些其他的现有具有潜在价值的信息资源。但是，收集信息才刚刚开始。大量的这样的数据需要被分类、标记、分析、分类、过滤，并且实施一系列其他操作来产生有用的、能够影响决策制定和提高底线的信息。随着企业跟踪和积累越来越多的数据，这些主题的重要性与日俱增。例如，未经加工的数据数量呈指数增长是由于传感器数据的出现导致的，包括无线射频识别（Radio Frequency IDentification，RFID）。基于传感器和地理位置数据的应用将会成为下一代BI专家最令人激动和快速增长的应用类别。从"文本挖掘"中获得文本资料，从"网络挖掘"中获得网络资源（见第4章），伴随将这两种资源合成的新方法的出现，表明组织机构正处在BI决策支持的一个爆炸新纪元的边缘。

BI已经形成了一套专门术语、系统和概念，这些可将它清楚地和其他有窃取信息倾向的国内外的智能系统区分开。也就是说，有许多这两类之间的比较，主要的努力就是尽力去获取卓越的智能资源、纯度和可靠度的智能处理过程，将信息恰当地传送给正确客户的机制。

1.3 节复习题

1. 列举出智能创造和使用的步骤。
2. 什么是商务智能治理？
3. 什么是智能收集？

1.4 交易处理和分析处理

为了说明BI的主要特性，我们首先说明从名称上看BI不是交易处理。我们都很熟悉支持我们日常交易的信息系统，例如ATM取款机、银行存款机、杂货店的收银机等。这些交易处理系统一直更新我们可能称做操作数据库的东西。例如，在一次ATM取款交易中，需要相应地减少银行存款余额，一次银行存款会增加相应的金额到银行账户中。一个杂货店的购买最终反映到商店一天的总销售计算中，并且在我们购买商品的同时它应该反映一个相应的商店库存减少量等。这些在线交易处理（Online Transaction Processing，OLTP，也称为联机事务处理）系统处理公司的日常实时业务。相反，数据仓库是一个截然不同的系统，它能够存储分析中将会用到的数据。分析的内容就是为了获得商业信息而清洗数据的管理能力，它能够被用来提供战术和操作决策支持。例如，一线人员可以做出更快和更多识广的决策。第2章会给出一个更技术化的数据仓库定义，但是它足以说明数据仓库是想利用在线分析处理系统的信息来工作。

产生于企业资源计划（Enterprise Resources Planning，ERP）系统，或者是在它的互补相似的供应链管理（Supply Chain Management，SCM）系统，或者是客户关系管理（Customer Relationship Management，CRM）中的大多数交易数据是存储在OLTP系统中，这是一种典型的计算机处理过程，当用户有需求时就快速地做出反应。每个需求被看做是一笔计算机记录的离散事件交易，例如存货收据或者客户订单。换句话说，一笔交易需要一套两个或多个数据库以一种全有或是全无的方式更新。

能够使OLTP系统有效地进行交易处理这种特殊的设计，对于终端用户的特别报告、查询、分析是没有效率的。在20世纪80年代，许多商业用户将他们的主要框架称做"黑洞"，因为所有的信息都进入它，但是信息从来不会出来。所有报告都是通过IT员工编程得到，然而仅仅

"提前装好"的报告能够在一个计划好的基础上产生，并且实时查询根本不可能实现。虽然20世纪90年代的基于客户/服务器模式的ERP系统有点能够面向报告，但是它还是不能满足一个常规的、非技术的，最终用户对于操作报告和交互分析期望的需求。为了解决这些问题，DW和BI的概念产生了。

数据仓库（Data Warehouse，DW）包含很多类型的数据，这些数据能够在某个时间点呈现出关于商业条件情况连贯的描述。想法就是创建一个数据库架构，这个数据库能够实时在线并且包括所有来自OLTP系统的信息，包括历史数据，但是以一种可以快速有效地进行查询、分析和决策支持的方式来重新组织和结构化。

将OLTP从分析和决策支持中分离出来，使前面描述的BI优点可以实现。下面将描述具有竞争性的智能和优势。

1.4 节复习题

1. 定义OLTP。
2. 定义OLAP。

1.5 成功的BI实施

实施和部署一个BI系统可能会耗时、耗费资金并且是失败的。让我们来探讨一些相关问题。

1.5.1 典型的BI用户群体

BI有更大更多样化的用户群体。BI的成功一部分取决于组织中的什么人来使用BI系统。BI成功的最重要方面就是它必须为整个企业带来利益。这就是说很有可能有许多用户会在DW投资描述的开始阶段就会参与进来。一点也不奇怪，可能会有专注战略层的用户，也会有专注交易层的客户。

存在于组织中多阶层的BI用户能够帮助指导DW的构建、BI工具的类型和其他需要的支持软件。当BI实施时，每一组人员都是评估特定BI成本和利润的非常好的信息资源。从上面的讨论中可以看到，任何一个擅长BI方法的企业的特征就是对来自不同层次潜在用户的鉴别。

1.5.2 合适的计划及其与商业战略的一致性

首先，投资BI最根本的原因必须与公司的商业战略相一致。BI不仅仅是信息系统部门的技术实践。它必须是一种通过改进商业过程和将决策制定过程转变为数据驱动型来改善公司行为方式的一种手段。许多曾经参与过成功BI实施的BI咨询和实践专家建议：计划框架是必需的前提条件。Gartner公司（2004年）曾经开发过一个框架，分为计划、业务执行、组织、功能、基础设施几个部分。在业务和组织层，在考虑完成任务需要具有组织能力的同时，需要定义战略和业务目标。高层经理们还应该考虑到以下的内容：BI实施行动所处的组织文化、为实施初始建立激情、为组织内部建立良好的能够分享BI实施的程序。同时高层经理们还要制定组织准备应对变化的计划。在实施过程中，首先要考虑的就是评估IS组织和潜在用户的基本技能、组织文化是否服从改变。从这些评估中，假设公司有理由需要实施，那么公司就可以准备一个更详细的行动计划了。另外一个实施BI成功的关键是多个BI项目之间的整合（多数企业应用多个BI项目）和BI与其他存在于组织和商业伙伴中的IT系统之间的整合。

如果一个企业的战略与实施DW和BI的原因一致，如果公司的IS组织能够在这样的项目中发挥它的作用，如果拥有需要的用户团体并且拥有合适的动机，那么开始实施BI和在公司中建立一个商务智能资格中心（BI Competency Center，BICC）是非常好的。这个中心应该实现下面的功能：

- 中心能够证明BI是如何与企业战略和战略实施联系在一起的。

- 中心能够鼓励潜在业务客户群体与IS组织之间的交流互动。
- 中心应该能够作为不同业务之间的知识库和最好的BI实践传播者。
- BI实施中，好的标准能够在企业中得到拥护和鼓励。
- IS组织能够在与用户团体交流互动的过程中学到很多，例如所需要的各种分析工具的知识。
- 业务用户群体和IS组织能够更好地理解为什么为了适应不断变化的业务要求，数据仓库平台必须非常灵活。
- 它能够帮助重要的利益相关者像高级管理人员一样知道BI是如何发挥重要作用的。

另一个BI成功实施的重要因素就是它本身具有实时的、基于需求的灵活环境，下面将进行介绍。

1.5.3 实时的、基于需求的BI是可达到的

随着销售业务数据与战略目标之间差距的需要越来越迫切，对于实时的、基于需求的分析信息获得的需求也日益增加。结果被称做实时BI应用系统的一类产品出现了（见第3章）。像RFID这类新数据生成技术的使用更加加速了这种增长和后来对于实时BI的需求。传统的BI系统使用大量的经过提取、清洗、载入数据仓库的静态数据来生成报告和分析。然而，需求并不仅仅是生成报告，因为用户需要商业监督、绩效分析和对于事情为什么发生的了解。这些能够提供给用户，这些用户需要知道（通常是实时的）数据变化或者相关报告的有效性、预警，网络、电子邮件或者即时通信（Instant Messaging，IM）应用中事件和新趋势的通知。除此之外，商务应用能够按照这些实时BI系统发现的信息来规划。例如，一个供应链管理系统（Supply Chain Management，SCM）能够在实际库存降到某个程度时自动地发出订单获取库存，当某个用户在线发出的订单金额超过了10 000美元的时候，客户关系管理会自动地触发客户服务代表和信用控制职员来查看用户的用户情况。

一种实现实时BI的方法就是利用传统BI系统的DW模型。在这种情况下来自创新的BI平台制造商（Ascential或者Informatica）的产品能够提供面向服务的、接近实时的解决方案，这些解决方案使得DW比典型的夜间提取/转换/加载（Extract/Transfer/Load，ETL）批次更新的方式更快（见第2章）。第二种方法通常被称做业务活动管理（Business Activity Management，BAM），这种方法被纯粹提供BAM和混合BAM中间件的提供商采用（例如Savvion、Heration、Software、Vitria、WebMethods、Quantive、Tibco、VineyardSoftwar）。它绕过了DW，应用网络服务（Web service）或者其他的监测手段来发现关键商务事件。这些软件监管（或者智能代理）能够安装在网络独立服务器或是交易应用数据库上，同时它们能够使用基于事件和过程的方法主动地、智能地监测业务过程。

1.5.4 开发或获得BI系统

现在，许多供应商都提供多样化的工具，它们中一些工具是完全设计好的（叫做壳（Shell）），你需要做的就是输入你的数字。这些工具能够购买或是租赁。关于产品、演示、白皮书，现行产品信息，可以访问information-management.com。需要进行免费用户注册。几乎所有的BI应用是由供应商自己或者和第三方合作，用它提供的壳为客户构建解决方案。公司面临的问题就是要选择什么样的方案：购买、租赁还是构建。每种方案又会有多种选择。做出决定的主要标准就是理由和成本-利润分析。

1.5.5 理由和成本-利润分析

随着BI应用数目的增加，证明它们合理和对它们进行优化的需要也在增加。由于大量无形利益的存在，这不是一项简单任务。直接或是无形利益需要被识别。当然，这也是为什么其他组

织中相似应用的知识和案例学习非常有用的原因。例如，数据仓库研究院（tdwi.org/）提供了很多关于产品和创新应用与实施的信息，这样的信息能够在评估直接和无形利益时发挥作用。

1.5.6 隐私安全和保护

这是一个在任何一个计算机系统发展中都特别重要的问题，特别是在可能拥有控制战略价值数据的BI中。同样，员工和顾客的隐私也需要保护。

1.5.7 系统集成和应用

除了小部分较小的应用外，所有的BI应用都必须与诸如数据库、遗留系统、企业系统（特别是ERP和CRM）、电子商务系统（买卖双方）等更多系统进行集成。另外，BI应用通常连接到内特网，并多次进入到商业伙伴的信息系统中。

而且，BI工具之间有时候是需要集成的，这样使得它们能够产生协同作用。

集成的需要使得软件供应商们不断地添加新功能到他们的产品中。购买包含所有功能软件包的顾客只需和一个供应商打交道即可，而不用处理系统连接问题。但是，他们可能丧失利用最好的组件构建系统的优势。

1.5 节复习题

1. 描述BI用户的主要类型。
2. 列举一些Gartner报告中强调的实施要点。
3. 列举一些BI成功因素。
4. 为什么很难说出BI应用的理由？

1.6 商务智能的主要工具和技术

DSS和BI如何实施是由应用的工具决定的。

1.6.1 技术和工具

在过去几年中开发了许多支持管理决策制定的技术和工具。它们以不同的名字和定义出现。表1-3描述了主要的计算机化的工具分类。表1-3中的内容将在其他章节进行详细的描述。

表1-3 决策支持的计算机工具

工具分类	工具和缩略语	在本书中的章节
数据管理	数据库和数据管理系统（DBMS）	2
	提取、转换和加载系统（ETL）	2
	数据仓库（DW）、实时数据仓库、数据集市	2
状态跟踪报告	在线分析处理（OLAP）	3
	高级管理人员信息系统（EIS）	3
可视化	地理信息系统（GIS）	3
	仪表盘	3
	多维演示	3
战略和绩效管理	业务绩效管理（BPM）或者企业绩效管理（CPM）	3
商业分析	仪表盘和平衡记分卡	3
	数据挖掘	4，5
	网络挖掘和文本挖掘	5
	网络分析	5
社交网络	Web 2.0	6
海量数据挖掘新工具	流实挖掘	6

1.6.2 选择BI供应商

近来，在BI软件和应用提供商中出现了大的波动。有些公司的名字在完成本书后会变得非常的熟悉，如：Teradata、MicroStrategy、Microsoft、IBM + Cognos + SPSS、SAP + Business Objects、Oracle + Hyperion、SAS，还有许多其他公司。有许多大的软件公司合并其他的公司来使自己产品包含全部的提供方案。例如，SAP合并了Business Objects，2008年IBM合并Cognos，2009年合并了SPSS，Oracle合并了Hyperion。在文本、网络和数据分析方面出现了新的公司。公司之间建立了合作伙伴的关系。例如，SAS和Teradata已经形成了合作伙伴关系来共同提供数据仓库和预测分析能力。

1.6 节复习题

1. 列举决策支持工具6种主要分类。
2. 识别主要BI零售的供应商公司。

1.7 本书计划

本书的6章是按照以下顺序进行安排的。BI包括了几个截然不同的组成部分。第2章主要讲述了在分析和绩效考核中必须应用的数据仓库。第3章讨论了BPM，仪表盘、记分卡和相关内容，主要讲述应用和数据挖掘过程。第4章主要描述了数据挖掘，包括神经网络等算法的技术细节。第5章的主要内容是文本和网络挖掘新出现的应用。第6章是对于全书的总结，并讨论了一些新的趋势，比如说无处不在的移动电话、地理信息系统设备、个人无线数字助理（Personal Digital Assistant，PDA）是如何使得大量数据库产生的。新的数据挖掘产品和BI公司正在研究这些新的数据库，并对顾客行为和活动有了更好更深的理解。应用案例1.4讨论了一个这样的实例——现实挖掘。我们在第6章中将学到这种应用和其他更多的应用。

应用案例1.4 下一代网络

传感网络是当前许多公司正在开发的应用之一，这一应用能够更好地理解顾客的活动。其中有一种应用就是分析超过400万手机用户活动的数据。这些数据来自GPS、移动电话塔和当地的Wi-Fi无线网点。这些数据是匿名的，但是彼此之间是相互联系的。这种联系使得数据挖掘者能够找到在特定时间、特定地点的顾客群。聚类技术能够识别这些顾客是属于哪个"部落"——商务旅行者还是年轻旅行者等。通过详细分析，在合适粒度建立起来的客户档案能够使企业确定目标市场并进行促销。

除了将信息用于更精确地确定目标客户这一传统应用之外，这样的系统某一天能够应用在研究犯罪和疾病的传播上。其他正在研究实施同样分析技术的公司，还有Kinetics、Nokia等。

来源：Compiled from S. Baker, "The Next Net," *Businessweek*, March 2009, pp. 42 – 46; Greene, K., "Mapping a City's Rhythm," *Technology Review*, March 2009, at technologyreview.com/communications/22286/page1/ (accessed January 2010); and Sheridan, B., "A Trillion Points Of Data," *Newsweek*, March 9, 2009, pp. 34 – 37.

1.8 相关资源、链接和 Teradata 大学网络的连接

使用下面部分中描述的内容会提高对于本书学习的效果。

1.8.1 资源和链接[1]

我们推荐下面的主要资源和链接：

- 数据挖掘研究院（tdwi.org）
- 信息管理（information-management.com）
- 在线分析处理（OLAP）报告（olapreport.com）
- 决策支持系统 DSS 资源（dssresources.com）
- 信息技术工具箱（businessintelligence.ittoolbox.com）
- 商业智能网（b-eye-network.com）
- AIS 世界（isworld.org）
- 微软企业财团（enterprise.waltoncollege.uark.edu/mec）

1.8.2 案例

所有的 BI 供应商（例如 MicroStrategy、Microsoft、Oracle、IBM、Hyperion、Cognos、Exsys、SAS、FICO、BusinessObjects、SAP 和 Information Builders）都提供有趣的顾客成功故事。学术案例在哈佛商学院案例收录（hbsp.harvard.edu/b01/en/academic/edu_home.jhtml）、企业绩效提高资源（bpir.com）、集团理念出版社（idea-group.com）、常春藤联盟出版（ivylp.com）、知识风暴（knowledgestorm.com）和其他网站上可以获得。Miller 的《MIS Cases》（2005 年）包含了简单的案例、应用电子表格和数据库练习，支持本书多章的内容。

1.8.3 供应商、产品和演示

许多供应商提供他们产品和应用的软件演示。在 dssresources.com 中有关于产品、架构和软件的信息。

1.8.4 期刊

我们推荐下面的期刊：

- 《Decision Support Systems》（决策支持系统）
- 《CIO Insight》（信息主管视角）（cioinsight.com）
- 《Technology Evaluation》（技术评估）（technologyevaluation.com）
- 《Baseline Magazine》（底线杂志）（baselinemag.com）
- 《Business Intelligence Journal》（智能商务期刊）（tdwi.org）

1.8.5 Teradata 大学网络的连接

本书与 Teradata 大学提供的免费资源紧密联系（TNN；参见 teradatauniversitynetwork.com）。

[1] 在本书送去印刷时，我们验证了本书参考的所有网站都是有效并可用。然而，URL 是动态的，我们在文章中参考的网站，有时由于公司改变名称、被收购或者拍卖、合并、或者失败会改变或者不可用。有时，由于网站维护、修复、重新设计，网站会关闭。许多组织的网站已经不采用“www”的设计，而一些仍在使用。如果您在连接网站时遇到了我们提供的上述问题，请耐心等候并使用网络搜索找到可能的新网站。大多数时候，您能够通过一个流行的搜索引擎快速地找到新的网站。我们担心为这种情况给您带来的不便表示歉意。

TUN门户网站分为两个部分：一部分供学生使用，一部分供员工使用。本书通过每章最后提供的特殊部分与TUN网站相连。这些部分包括特定章节相关资源的链接。另外，在TUN上，我们提供利用软件和其他资料（例如案例）的动手练习。

1.8.6 本书的网站

本书网站是：pearsonhighered.com/turban，包含与本书内容有关的补充材料。

本章重点

- 商业环境正在变得越来越复杂、变化越来越快，这使得决策的制定更加困难。
- 业务必须通过更快更好的决策快速地对变化的环境做出反应和适应。
- 制定决策的时间框架正在缩小，而决策制定的全球化特性正在扩大，这使得计算机化的决策支持系统的发展和使用更加必要。
- 对于一个组织的生存来说，组织的计算机化支持是必需的。
- 自动决策支持为许多行业提供了基于规则的重复性决策解决方案（例如定价）。
- BI方法利用数据仓库能够进行有效的数据挖掘、在线分析处理、业务绩效管理、数据可视化。
- BI包括数据仓库、最终用户使用的商业分析工具、用户界面（例如仪表盘）。
- 许多组织使用BPM系统来监测运行情况，并将它们与标准和目标进行比较之后用图表形式表示出来。
- 数据挖掘是在大量的数据中找到信息和关系的工具。
- 在管理决策系统的使用和发展中发挥关键作用的技术有：网络技术、因特网、内部网、外部网。

关键术语

analytics 分析
Automated Decision Systems（ADS，自动决策系统）
automated decision support 自动决策支持
BI governance 商务智能治理
business analytics 商业分析
Business Intelligence（BI，商务智能）
Business Performance Management（BPM，业务绩效管理）or Corporate Performance Management（CPM，企业绩效管理）
complexity 复杂度
corporate portal 公司门户网站
data 数据
database 数据库
data mining 数据挖掘
decision making 决策制定
Global Positioning System（GPS，全球定位系统）
Geographical Information System（GIS，地理信息系统）
information 信息
information overload 信息过载
intelligence 智能
intelligent agent 智能代理
knowledge 知识
management science 管理科学
Online Analytical Processing（OLAP，在线分析处理）
Online Transaction Processing（OLTP，在线交易处理）
predictive analysis 预测分析
predictive analytics 预测分析学
user interface 用户界面
Web service 网络服务

讨论题

1. 为图1-2中每个部分举个例子。
2. 区分智能收集与获取信息。
3. 什么是商务智能治理？
4. 讨论在BI实施中的主要注意事项。

练习

Teradata 大学网络（TUN）和其他的动手练习题

1. 访问 teradatauniversitynetwork.com。利用指导老师提供的注册，登录并学习网站内容。准备一个有用的资料清单。你会收到与网站有关的作业。准备 20 个在网站中你觉得对于你有用的知识清单。
2. 进入 TUN 网站选择"案例、项目和作业"（cases, project, and assignment）。然后选择案例学习："Harrah 从顾客信息中得到的高回报"（Harrah's High Payoff from Customer Information）。回答关于案例的以下问题：
   a. 数据挖掘产生什么样的信息？
   b. 在决策制定的管理中，信息是如何发挥作用的？
   c. 列出被挖掘的数据种类。
   d. 这是 DSS 还是 BI，为什么？
3. 访问 teradatauniversitynetwork.com。找到标题为："Data Warehousing Supports Corporate Strategy at First American Corporation"的文章（Watson、Wixom 和 he Goodhue），阅读文章并回答下面问题：
   a. 公司的 DW/BI 项目的驱动是什么？
   b. 实现了什么战略优势？
   c. 达到了哪些操作和战术优势？
   d. 实施中的关键成功因素（Critical Success factor，CSF）是什么？

小组作业和角色扮演

1. 写一篇 5－10 页的报告，描述一个你熟悉的公司在决策支持中是如何使用计算机和信息系统的，包括网络技术。基于本章的知识，描述如果决策支持系统能够轻易地使用，那么经理应该如何使用这些系统？哪些是你可以得到的，哪些不是？
2. 访问 fico.com、ilog.com 和 pega.com，观看这些网站的演示。准备行业和功能区域的 ADS，列举哪些决策是自动制定的。

网络练习

1. 访问 fico.com。应用网站信息来识别在不同功能区域的 5 个由 ADS 支持的问题。
2. 访问 sap.com 和 oracal.com。找到关于 ERP 如何帮助决策者的信息。另外，检查这些软件产品是如何利用网络技术和网络本身的。基于你的发现写一篇报告。
3. 访问 intelligenterterprise.com。为在本章中引用的每个题目找到一些有趣的开发报告，并写一篇报告。
4. 访问 cognos.com 和 businessobjects.com。比较这两个公司的 BI 产品的性能。
5. 访问 microsoft.com。检查它的 BI 产品。
6. 访问 oracle.com，查看它的 BI 产品。Oracle 的 BI 产品是如何与它的 ERP 产品联系在一起的。
7. 访问 microstrategy.com。找到关于 BI 5 种类型的相关信息。准备一份每种类型总结的表格。
8. 访问 oracle.com，点击应用下的超链接。看看该公司的主要产品是什么。将它们与本章提到的支持技术联系起来。

本章结尾应用案例

Vodafone 利用商务智能实现客户增长和保留计划

问题

Vodafone 新西兰有限公司是电信巨头 UK 的一个子公司，它在新西兰取得了巨大成功。从很小的基础上，公司迅速获得了超过 50% 的市场份额。然而，随着移动电话市场逐渐走向成熟，Vodafone 的市场份额停留在 56% 左右，顾客的总数也没有变化。使情况更加糟糕的是：其他的竞争者进入，遵守政府政策的成本增加了，每个客户的收益也停滞不动。公司不得不重新调整它的从现有顾客保持和增加利润的战略。Vodafone 的顾客分析高级经理 John Stewart 说："既然我们拥有这些顾客，所以我们需要回答新的问题：我

们怎么增加我们的边际利润？我们怎样留住客户？" Vodafone需要基于它的市场、顾客、竞争者的实时知识来做出更好的决策。Cheryl Krivda的一份报告中指出："Vodafone需要使用能够提供基于事实的决策支持的BI，全面转向市场分析。目标就是：利用现有渠道，当顾客需要时，将正确的信息传达给正确的客户。"

解决方案

首先，Vodafone组建了一个顾客知识和分析部门来实施分析、建立模型、市场研究和有竞争性的智能。John Stewart是这个部门的经理。Vodafone利用企业数据仓库（Enterprise Data Warehouse，EDW）来获得组织内所有信息的单方面查看。EDW可以实现组织内所有信息的集中查看，还可以产生事先定义好的查询和报告、在线分析处理和预测分析（见第4章）。公司同时雇佣建模专家来培训自己的分析团队。除了Teradata数据仓库平台外，还有许多其他的软件工具，如KXEN、SAS、SPSS，也被用来建立模型和进行研究。

应用Teradata数据仓库平台和所有相关的工具，Vodafone销售部门的员工现在能够实施分析并获得更好的顾客优化、活动有效性分析和顾客服务。Stewart相信新的工具使得Vodafone能够有整体视角。

> 他说："作为一个团队，通过提问和提供支持，我们相互联接。在这个过程中，我们可以相互学习，这能够使我们对于业务的研究更有价值。当你将所有的信息和知识放在一起时，你能够发现更多关于顾客的信息。

EDW的一个应用就是基于驱动的营销活动。过去，在实施营销活动时人工干预是需要的。通过新的平台，Vodafone能够自动地基于顾客最近的活动发起营销活动。

结果

业界EDW最大的好处就是分析人员能够将大多数时间花在研究数据而不是产生数据上。"现在，我们针对顾客的活动更加有效。" Stewart说。但是，这并不是说我们不停地进行活动。我们对什么样的顾客实施活动更加有目标，与顾客的相关程度也更大。

系统也在决策的制定过程中帮助决策的制定者提供更好的信息。Vodafone正在开发一个应用，这一应用能够优化收入和顾客优先次序。目标就是在活动和与顾客接触的过程中获得最好的收益。不用泄露细节，就可以知道公司正在朝着它的目标迈进。

来源：Compiled from C. D. Krivda, "Dialing up Growth in a Mature Market," *Teradata Magazine*, March 2008, pp. 1–3.

本章结尾案例的问题

1. Vodafone面对的挑战是什么？
2. 他是如何找到问题的？
3. 列出Vodafone的应用使用了哪些工具。
4. 在这些实施中得到的好处是什么？
5. 我们在这个案例中能够学到什么？

参考文献

Acxiom. "Location, Location, Location." **acxiom.com** (accessed March 26, 2009).

Baker, S. (2009, March 9). "The Next Net." *BusinessWeek*, pp. 42–46.

Davenport, T. H., and J. G. Harris. (2009, Winter). "What People Want (and How to Predict It)." *MIT Sloan Management Review*, Vol. 50, No. 2, pp. 23–31.

Davenport, T. H., and J. G. Harris. (2005, Summer). "Automated Decision Making Comes of Age." *MIT Sloan Management Review*, Vol. 46, No. 4, pp. 83–89.

DemandTec. "Giant Food Stores Prices the Entire Store with DemandTec." **demandtec.com** (accessed March 26, 2009).

Gartner, Inc. (2004). *Using Business Intelligence to Gain a Competitive Edge*. A special report. Stamford, CT: Gartner, Inc. **gartner.com**.

Greene, K. (2009, March). "Mapping a City's Rhythm." *Technology Review*, at **technologyreview.com/communications/22286/page1/** (accessed January 2010).

Imhoff, C., and R. Pettit. (2004). "The Critical Shift to Flexible Business Intelligence." White paper, Intelligent Solutions, Inc.

Krivda, C. D. (2008, March). "Dialing up Growth in a Mature Market." *Teradata Magazine*, pp. 1–3.

Krizan, L. (1999, June). *Intelligence Essentials for Everyone*. Washington DC: Joint Military Intelligence College (occasional paper number six). Department of Defense.

Matney, D., and D. Larson. (2004, Summer). "The Four Components of BI Governance." *Business Intelligence Journal*.

Miller, M. L. (2005). *MIS Cases*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Sheridan, B. (2009, March 9). "A Trillion Points of Data." *Newsweek*, pp. 34–37.

Thompson, O. (2004, October). "Business Intelligence Success, Lessons Learned." **technologyevaluation.com** (accessed June 2, 2009).

Watson, H. (2005, Winter). "Sorting Out What's New in Decision Support." *Business Intelligence Journal*.

Ziama, A., and J. Kasher. (2004). *Data Mining Primer for the Data Warehousing Professional*. Teradata, Dayton, OH: Teradata.

数据仓库

学习目标

- 理解数据仓库的基本概念和定义
- 理解数据仓库架构
- 描述数据仓库开发和管理的过程
- 解释数据仓库的具体运作
- 解释数据仓库在决策支持中扮演的角色
- 解释数据集成以及数据提取、转换和加载过程
- 描述实时数据仓库
- 理解数据仓库的管理和安全问题

数据仓库的概念早在20世纪80年后期就提出了。本章介绍了一种重要的基本数据库类型，即数据仓库。数据仓库主要用于决策支持中以改进分析能力。

开篇场景：DirecTV的蓬勃发展与实时数据仓库

首先来看DirecTV公司的例子，它反映了交互式数据仓库和商务智能软件是怎样在企业中传播开来的。通过使用Teradata和GoldenGate公司的解决方案，DirecTV开发了一种可以整合全公司实时数据资产的产品。公司数据仓库总管，Jack Gustafson曾公开宣布此产品在持续使用中已经收回了成本。DirecTV采用实时交易数据管理解决方案，这一技术决定所带来的商业利益已经远远超出了开始预期的技术利益。

DirecTV以电视直播卫星服务著名，由于其先进的高清编程、交互功能、数字录音服务和电子节目指南，DirecTV为电视产业做出了杰出贡献。DirecTV在北美和拉丁美洲拥有超过13 000名雇员，2008年的财政收入达到20亿美元，同时总订户数量接近5 000万。

问题

在持续性的快速增长中，由于客户日电话数量不断增长，所以DirecTV面临着处理高交易数据量的问题。随着市场情况的快速变化，如何存储如此巨大的数据量是DirecTV面临的关键挑战之一。几年前，公司开始寻求一种更好的解决方案，用来给商业方提供呼叫中心的日常报告。管理层期望报告能有多种用途，包括测量和维护客户服务、吸引新客户并防止客户流失。同样重要地，DirecTV的技术团队想要减少当前数据管理系统加载在CPU上的资源工作量。

显然数据仓库的早期实施能很好地满足公司的需求，但随着业务的不断增长，已经快达到这种实施的极限。在动态数据仓库解决方案出现之前，公司每晚用批处理模式将数据从服务器中提取出来，这是一个占用很长时间并耗尽系统资源的过程。上传每日批量数据至数据仓库，早已成为（对很多公司来说，现在仍然是）一种标准程序。如果公司业务竞争不包括数据实时性，那么这样的每日上传程序也许很适合公司的业务。不幸的是，这不是DirecTV的情况。在高度动态的消费市场中，为了管理巨大的呼叫量，DirecTV的业务用户需要从客户呼叫中实时获取数据。

解决方案

首先，新项数据仓库系统的目标是至少每天将最新数据发送到呼叫中心，但是一旦整合解

决方案实现后，目标将下降至每15分钟更新数据。"我们期望不同城市间的广域网中的数据延迟小于15分钟。"Gustafson解释道。

项目的第二个目标是简化改变数据的采集，以减少开发者所需要的维护工作量。尽管多个平台间的数据源不是初始需求的一部分，但当DirecTV看到了GoldenGate集成系统的能力后，这点将会改变。GoldenGate实现了一系列数据管理系统和平台的集成。DirecTV包括Oracle、HP NonStop平台、IBM的DB2系统和Teradata数据仓库，Gustafson说："我们运用GoldenGate不是为了集成为一个系统，这点仍然吸引着我们，我们正在外购通话记录，但是也在外购NonStop和其他数据源。我们认为如果需要购买一种工具来处理这些工作，希望它能在所有我们公司支持的平台上运行。"

结果

随着系统功能的明确，其潜在的商业利益也开始显现。正如Gustafson所说的："一旦建立了数据仓库，我们就会获得了一个能让我们衡量实时流失的巨大商业利益。"他还说道："我们曾说过，既然我们已经有了所有的实时报告，那么我们能用这些报告来做什么呢？"我们曾经的一个做法是利用这些数据来针对特定的客户，以减少流失。借助他们手头的最新数据，呼叫中心的销售人员可以同当天几个小时内刚刚请求取消业务的顾客进行联系，并提供一项新的业务以留住这名顾客。一旦技术组安装好这些必要的报表工具，那么我们就可以针对特定顾客采取促销活动以保留客户，并优先为他们提供特殊服务。而这一活动也开始奏效，"自从我们实施了这一方案后，公司的客户流失情况已经有所降低，"Gustafson说："分析家们已经开始称赞我们在这一领域所取得的成绩，而这一切多半要归功于我们所采取的及时反馈活动，可以在当天为他们发出新的邀请。"

我们建立系统的另一个目的是记录顾客的服务日志，实现对现场报告的重复技术问题的回复报告。这使得管理层能更好地对现场报告做出评估和反馈，从而提高服务质量、减少技术人员开支。实时呼叫中心中生成的报告还可基于日常的呼叫量来控制中心的工作负载量。借助这些数据，管理层可以通过日常呼叫量和历史平均值的对比来进行例外报告。

而系统在另一个业务中心的使用情况，是我们之前未曾预料到的。实时业务报告在此不仅用于订单管理，还用于欺诈检测。通过获取新顾客的实时订单信息，欺诈管理专家可以对这些数据进行检测，以排除欺诈订单。Gustafson指出："这真是帮了我们大忙，它减少了我们的劳动力和生产成本。"

开篇场景的问题

1. 为什么建立一个实时数据仓库对于DirecTV如此重要？
2. DirecTV在建立集成的实时数据仓库时所面临的挑战是什么？
3. 以DirecTV的具体实施为基础，说明传统数据仓库和实时数据仓库的主要区别。
4. 采用实时数据仓库而非传统的数据仓库，DirecTV能获得什么样的战略优势？
5. 你认为什么原因导致像DirecTV这样的大型组织不能建立一个合适的数据仓库。

我们从开篇场景中能够学到什么

开篇场景阐明了实施实时数据仓库的战略价值，以及其对商务智能技术的支持。DirecTV能够将其数据资产分布在整个企业内部，让它的知识人员无论何时何地，只要需要就能使用它。数据仓库将组织内部多个数据库集合成一个整体，形成了单一的公司真实版本，将所有的员工放在了相同的页面上。此外，数据仓库还满足了决策者对实时数据的需求，他们可以在决策中使用数据仓库，提高了公司在产业中的战略竞争优势。这个故事给我们一个重要收获就是——一个实时的、企业级的数据仓库同企业战略上决策支持的结合将会为企业带来显著的利益（财务及

其他）。

来源：L. L. Biggs, "DirecTV Connects with Data Integration Solution," *Business Intelligence Journal*, Vol. 14, No. 1, 2009, pp. 14 – 16; "DirecTV Enables Active Data Warehousing with GoldenGate's Real-Time Data Integration Technology," *Information Management Magazine*, January 2008; directv.com.

2.1 数据仓库的定义和概念

实时数据仓库（Real-time Data Warehousing，RDW）、决策支持系统（Decision Support System，DSS）以及商务智能技术综合运用是一种重要的梳理业务流程的手段。在开篇场景中，我们介绍了实时数据仓库支持决策的一个真实情景，即通过分析来自不同渠道的海量数据来提供支持关键流程的快速解决方案。借助一种易理解的形式存储于数据仓库中的真实数据，扩展了DirecTV的革新业务流程。通过实时数据仓库，DirecTV可以浏览公司的业务实时状态并快速识别问题所在，而这也正是分析解决这些问题的首要步骤。除此之外，客户可以获取他们的实时订阅、电视服务以及其他账户的信息，也就是说，系统同时还具备了显著的竞争优势。

决策的制定需要关于当前运作、趋势和变化的明确、可靠的信息，而数据往往是分散于不同的操作系统下，所以管理者常常是至多基于部分信息来做出所谓的决策。数据仓库通过访问、整合、组织关键业务数据使其一致、可靠、及时和可用，排除了障碍，使得数据实现了随时随地地取用。

2.1.1 什么是数据仓库

简单来说，**数据仓库**（Data Warehouse，DW）就是一个支持决策制定的数据池，它同时还是一个关乎整个组织的所有管理者潜在兴趣的当前数据和历史数据的存储库。数据通常以一种易于分析业务动态的形式来构建，例如在线分析处理、数据挖掘、查询、报表和其他决策支持应用。数据仓库通常是一个面向主题的、集成的、非易失的且随时间而变的数据集合，用来支持管理者的决策。

2.1.2 数据仓库的特点

了解数据仓库的一个基本方法就是了解它的基本特点（Inmon，2005）：

- **面向主题的** 面向主题提供了一种更易理解的数据组织方式，数据以某个具体的主题来组织的，例如销售、生产或者顾客，每个主题下只包括决策支持的相关信息。面向主题使得用户可以决定他们的业务展现形式，以及为什么如此展现。数据仓库与操作性的数据库有很大的不同，后者大多是以产品为导向的，并且常常由于业务的处理而需要更新数据库。
- **集成的** 集成同面向主题密切相关。数据仓库需要将多渠道的数据以一致的形式来存储，并解决由于集成而出现的诸如命名冲突和数据类型差异性的问题。数据仓库是完全集成的。
- **随时间而变的（时间序列）** 数据仓库需要定期维护历史数据。除非是在实时数据仓库中，否则这些数据并不要求提供实时状态。它们检测趋势、偏差以及预报和比较的长期关系，从而支持决策。每个数据仓库都有时态性，时间是所有数据仓库都必须支持的一个重要维度。在数据仓库中，数据分析要包括不同时间点的分析，如日、周以及月等。
- **非易失性** 数据一旦录入数据仓库后，用户就不能对其进行更改和更新。过时的数据将会被丢弃，而变化后的则作为新数据被记录下来。

上述的这些特点极大地提高了数据仓库的数据存取能力，除此之外，数据仓库还有一些别

的特点：

- **基于网络** 数据仓库通常被设计为基于网络应用提供高效的运算环境。
- **关系的/多维的** 数据仓库常常是基于关系架构或是多维架构，最近的一个多维架构的调查是由 Romero 和 Abelló 于2009年发现的。
- **客户端/服务器** 数据仓库运用客户端/服务器架构易于终端用户的访问。
- **实时** 最新的数据仓库已经实现了提供实时或者动态的数据访问和分析能力（Basu 2003，Bonde and Kuckuk 2004）。
- **元数据** 数据仓库通过元数据，即数据的数据，来描述数据的组织方式以及如何有效地使用它们。

尽管数据仓库是数据的集合，但从字面上来说，数据仓库就是一个涉及全过程的东西（Watson，2002）。数据仓库是一门将它的应用结果用于支持决策，并允许随时查询业务信息，同时还具备对于业务的洞察力的学科。数据仓库有3种主要类型，分别是数据集市、业务数据存储（Operational Data Store，ODS）和企业数据仓库（Enterprise Data Warehouse，EDW）。下面我们来介绍这3种类型的数据仓库以及元数据。

2.1.3 数据集市

尽管数据仓库是将数据库与整个企业结合起来，但数据集市相对来说通常较小，并且关注于某一个特别的主题或部门。数据集市是数据仓库的一个子集，通常包括一个单独的主题域，如销售市场、企业运营等。数据集市可以是独立的也可以是非独立的（归属于某一数据仓库）。非独立的数据集市是由数据仓库直接生成的一个子集，它具备稳定的数据模型和提供高效数据的优势。在数据仓库的基础上，非独立的数据集市可以作为一个企业范围内的数据模型而存在，它可以确保数据仓库终端使用者所浏览数据的一致性。数据仓库的高成本限制了它在大公司的应用前景，作为替代，更多的公司开始倾向于选择独立的数据集市，这是一种低成本、低版本的数据仓库。独立的数据集市可以被看做是支持某项业务决策或者某个部门的小型数据仓库，但是它的数据源并不是企业数据仓库。

2.1.4 业务数据存储

业务数据存储提供了一种较新的客户信息存储模式。这类数据库通常作为某一个数据仓库中的临时决策域。不同于数据仓库的静态存储，ODS中的内容在整个业务运营过程中是随时更新的。ODS常常被用于涉及核心应用的短期决策，个别情况下还会与企业数据仓库结合，用于中期或长期决策。举例来说，数据仓库是长期记忆因为它存储的是永久信息，而ODS则是短期记忆，因为它只存储最近的信息。ODS将多个系统中的信息集成起来，提供近实时性的、可变集成的当前数据。而这种数据提取、转换和加载过程和数据仓库的做法是一样的。当操作性数据需要多维化分析时，操作集市（open marts）就会形成，而这些操作性数据则是来源于ODS（Imhoff，2001）。

2.1.5 企业数据仓库

企业数据仓库是支持整个企业决策的大型数据仓库，这也正是我们之前提到的 DirectTV 公司所建立的数据仓库。大范围特性使得它将不同来源的数据标准化，支持商务智能和决策支持的有效运作。EDW通常被用于为多种决策支持系统提供数据，包括客户关系管理、供应链管理、业务绩效管理、业务活动监控（Business Activity Monitoring，BAM）、产品生命周期管理（Product Lifecycle Management，PLM）、收入管理系统以及知识管理系统（Knowledge Management System，

KMS)。应用案例2.1具体介绍了正确设计和实施下的企业数据仓库将为公司带来的巨大利益。

应用案例2.1 企业数据仓库实现成本节约、提高流程效率

NCR成立于1884年俄亥俄州的代顿市，至今已是一家市值56亿美元的纽约证券交易所上市公司。公司为全球零售、金融、保险、通信、制造、旅游及运输各领域提供技术解决方案，包括商店自动化、ATM机、咨询服务、媒体产品以及硬件技术。

1991年NCR被AT&T收购后，NCR致力于开发以区域和产品为中心的自主架构，每个区域可以自行决定产品和服务的提供、市场营销策略、产品定价、业务流程的发展以及报告标准。在这一模式下，整个公司的运营状况被数个财务和运营系统所掌控，而非仅仅针对某一企业。

1997年，NCR脱离AT&T，再次成为一家独立公司，每天公司运营都会有大量的资金损失。资产分析使得NCR注于业务流程重整以维持并增强其在全球市场的竞争地位，逐渐成长为一家真正的全球企业。

全球化的目标使得NCR开始逐步从原来以硬件为核心的区域中心模式向集成化的以解决方案为核心的企业架构转变。为了实现这一转变，NCR就必须使其店面变得全球化、中心化和集成化。只有这样，才能在其业务流程重整的过程中实现有效的控制。企业数据仓库在这一阶段对于NCR极为重要，并对其在之后数年内建立一个新的全球化、单实例的ERP系统也将有重要影响。

企业数据仓库实施的责任由NCR公司的财务部门和全球客户服务部门（Worldwide Customer Services，WCS）共同负责。这两个部门下的业务团队分别负责财务信息的传送和地理信息系统（Global Information System，GIS），它们和EDW团队密切合作，以保证信息技术能够理解公司新架构下的业务需求。公司选取了Teradata数据仓库进行企业数据仓库的构建，一方面是由于它的可扩性，以及对非结构化查询和高并发处理的自适应性，另一方面则是由于其相对低廉的维护费用。

EDW的巨大潜能迅速在公司中体现出来，并带动了NCR财务部门组织架构和业务流程的相应变动，财务部门的财务循环周期从原来的14天减少到了6天，而全球化的报告标准也得以完全建立。在WCS部门中，EDW的建立使得个人客户的收益分析以及随之的计划改善变得可行。不仅是以上两个方面，EDW的巨大潜能还体现在销售和市场、运营和库存管理，甚至是人力资源中。ERP的操作标准化以及客户服务中的动态改善令NCR的未来变得更为明朗，NCR已逐步成为一个全面的全球业务解决方案提供商。

除了上述EDW所带来的丰厚回报外，EDW不仅为公司收回了预期的项目成本，还将带来更多的收入，其中包括每年节约1亿美元的库存成本，2亿美元应收账款的持续减少，每年5 000万美元财务费用的减少，以及在WCS部门最初实施EDW的5年中2 200万美元的成本节约。

除了成本的节约和业务流程的高效化外，还需考虑到更多EDW所能做的和其带来的重要价值，其中最有战略意义的是用EDW来推动增长。

尽管EDW项目并不是创造利润的机会，相反它需要资金的筹集，然而它为公司所节省的成本费用将远远超过实施过程中的花费。而一旦EDW成功实施后，随着公司的成长，它将带来源源不断的利润。以下是EDW为NCR公司带来定量和定性利润的详尽说明：

定性利润

- 财务循环周期从原来的14天减到6天
- 提高了企业报告标准
- 实现个人客户的效益分析以及随之的改善计划
- 提供持续的全球报告流程
- 改进及时运送
- 提高库存管理能力，减少货物过时风险

定量利润

- 每年5 000万美元财务控制成本的减少
- 2亿美元应收账款的持续减少，转化为每年2 000万美元应收账款持有成本的减少
- 1亿美元产成品存货的持续减少，相当于每年1 000万美元库存持有成本的减少
- 在WCS部门最初实施EDW的5年中2 200万美元的成本节约，包括自动向客户提供服务等级协议（Service Level Agreement，SLA），人员减少，以及客户资料维护费用的降低。
- 供应链管理改进带来的1 000万美元的利润
- 由于财务和会计报告中人员的减少，5年内610万美元净现值费用的减少
- 350万美元通信费用的减少
- 由于ERP转型费用的减少节约了300万美元
- 由于Oracle 10.7升级为Oracle11、Oracle 11i，首次实现了报表功能，节约了170万美元的报表开发费用

资源：Teradata，"Enterprise Data Warehouse Delivers Cost Savings and Process Efficiencies，" teradata.com/t/resources/case-studies/NCR-Corporation-eb4455/（accessed June 2009）.

2.1.6 元数据

元数据（metadata）是数据的数据（Sen，2004；Zhao 2005）。元数据描述了数据的结构和部分意义，因此有助于数据的有效或无效使用。Mehra（2005）文献指出，极少的组织真正理解元数据，而极少理解元数据的组织知道如何设计并执行元数据策略。就用法而言，元数据通常被定义为技术元数据或者业务元数据。模式是另外一种浏览元数据的方式，通过模式浏览，可以知道语法元数据（也就是描述数据语法的数据）和结构元数据（也就是描述数据结构的数据）以及语义元数据（也就是描述某个特定域的数据含义的数据）的不同。

接下来，我们将解释传统元数据模式以及如何通过一个全面的元数据集成方法实现有效的元数据策略。这些方法包括本体论和元数据注册，企业信息集成（Enterprise Information Integration，EII），数据提取、转换和加载以及面向服务的架构（Service-Oriented Architecture，SOA）。有效性、可扩展性、重用性、互用性、效率和性能、量化、权限、灵活性、隔离、用户交互、版本、多样性以及低维护成本，这些都是建立一个成功的元数据驱动的企业的成功要素。

Kassam（2002）文献提出，业务元数据包括能提高我们对传统数据（也就是结构数据）理解力的信息。元数据的首要目的是描述数据的内容特征，也即是说，它要提供知识建立所需要的丰富信息。尽管业务元数据的效率较差，但却比结构数据更具潜能。元数据的内容对于所有的用户来说不需是相同的。在某些情况下，元数据有助于数据和信息转换为知识。Bell（2001）的文献认为元数据为元商业架构奠定了基础，Tannenbaum（2002）的文献描述了如何识别元数据的需求，Vaduva and Vetterli（2001）文献概要介绍了数据仓库中的元数据管理，Zhao（2005）文献

描述了元数据管理成熟度的5个阶段，分别是随机状态、发现、管理、优化和自动化。这5个级别有助于组织理解如何使用以及最好地使用元数据。

元数据的设计、建立和使用——即描述和总结数据的数据——以及元数据标准可能涉及的伦理问题。这些问题主要产生于元数据中信息的收集和归属，包括信息的隐私性，以及在设计、收集和分离期中形成的知识产权。这方面更详细的内容可以参见（Brody，2003）。

2.1 节复习题

1. 什么是数据仓库？
2. 数据仓库与数据库相比有什么不同？
3. 什么是业务数据存储？
4. 请说出数据集市、业务数据存储和企业数据仓库的不同。
5. 简述元数据的重要性。

2.2 数据仓库流程概述

不管是私人组织还是公共组织，都会以某种增长速度持续收集数据、信息和知识，并将它们存储于计算机系统中。而这些数据和信息的维护和使用将会变得极为复杂，特别是涉及可扩展性问题。除此之外，由于网络连接尤其是因特网的可靠性和可用性的改善，用户访问信息的需求也在逐步增加。在多个数据库运作的情况下，是否集成为一个数据仓库都变得极为困难，需要相当专业的知识，但其带来的好处将远远超过其花费的成本，具体参见本章开篇场景和应用案例2.2。

应用案例2.2 数据仓库支持First American公司企业战略

随着从传统银行方法向以CRM为中心的战略转型，First American公司逐步从1990年亏损6 000万美元的阴影中走出来，并在10年后成为了创新型金融服务的领军者。这一战略的成功实施离不开VISION数据仓库的帮助，VISION数据仓库中存储着公司客户行为的大量信息，如使用中的产品、购买偏好以及客户价值定位。VISION数据仓库提供了下述功能：

- 识别前20%的有价值客户
- 识别40%～50%的无价值客户
- 客户保有策略
- 低成本分配渠道
- 客户关系扩张策略
- 信息流程新设计

通过数据仓库访问信息，使得渐进性和突进性的改变成为可能。First American公司由此获得了突进性的改变，进入其金融服务的"幸福的第16年"。

来源：Based on D. L. Cooper, H. J. Watson, B. H. Wixom, and D. L. Goodhue, "Data Warehousing Supports Corporate Strategy at First American Corporation," *MIS Quarterly*, Vol. 24, No. 4, 2000, pp. 547 – 567; and D. L. Cooper, H. J. Watson, B. H. Wixom, and D. L. Goodhue, "Data Warehousing Supports Corporate Strategy at First American Corporation," SIM International Conference, Atlanta, August 15 – 19, 1999.

多数组织都需要建立数据仓库，来存储大量时序数据支持决策。这些来自内部、外部不同来源的数据经过数据清洗和组织以满足组织的需要。一旦这些数据存储于数据仓库后，就可以建立服务于某一特定域或部门的数据集市。或者，也可以根据需要，先建立数据集市，然后将其集

成到企业数据仓库中。显然，数据集市不能被二次开发，但是数据已经加载到计算机中或者维持初始状态，用于商务智能工具的操作。

图 2-1 展示了数据仓库的基本框架，以下是数据仓库流程的一些重要概念：

- **数据源** 数据往往来自于多个独立的“遗留”操作系统中，或者一些外部数据提供商，如类国统计局，也有可能来自在线交易处理系统或者 ERP 系统。而来自 Web 日志中的 Web 数据也可以组建数据仓库。
- **数据提取和转换** 使用定制或者商业 ETL 软件实现数据的提取和正确的转换。
- **数据加载** 数据被加载到数据准备区中，进行数据转换和清洗，之后才可以被加载到数据仓库或者数据集市中。
- **综合数据库** 从本质上来说，综合数据库是指由企业数据仓库提供各项决策所需的不同来源的概括和详细数据。
- **元数据** 元数据需要定期维护，以供信息技术人员和用户进行评估。元数据包括数据以及组织规则相关的软件程序，用于组织数据概要，以便于索引和查询，尤其是利用网络工具。
- **中间件** 中间件为数据仓库中的数据访问提供接口。技术用户，如分析师可以通过编写 SQL 查询语句，而其他人则可借助成熟的查询环境，如 Business Objects 来访问数据。业务人员可以使用多种前端应用程序与存储在知识库中的数据进行数据交互，包括数据挖掘、OLAP、报表工具以及数据可视化工具。

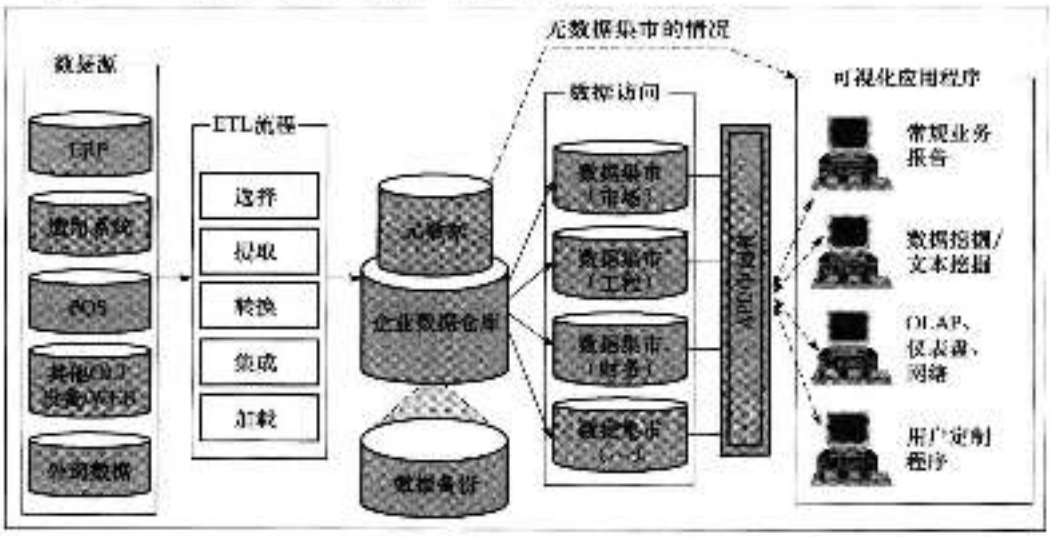


图 2-1 数据仓库框架概览

2.2 节复习题

1. 描述数据仓库流程。
2. 描述数据仓库的重要组件。
3. 阐明中间件在数据仓库中所承担的角色。

2.3 数据仓库架构

数据仓库的基本信息系统架构有很多种。大体来说，这些架构通常是客户/服务器架构或者

多层架构，其中最多见的是二层和三层架构，如图2-2和图2-3所示，但有时也会出现单层架构。多层架构可以满足大规模、高性能要求的信息系统的需求，例如数据仓库。为了了解数据仓库中多层架构的具体应用，Hoffer et al.（2007）区分这些架构，将数据仓库划分为3个部分：

1. 数据仓库本身，包括数据和相关联的软件。

2. 数据采集（后端）软件，用于从遗留系统和外部数据源中提取数据，合并和汇总后，再将它们加载到数据仓库中。

3. 客户端（前端）软件，如决策支持系统、商务智能系统、业务分析引擎，允许用户对数据仓库进行数据存取和数据分析。

在三层架构中，数据和用于数据采集的软件是一层（也就是数据库服务器），数据仓库是另一层，第三层包括决策支持系统、商务智能系统、业务分析引擎等（也就是应用服务器）以及客户端（如图2-2）。数据仓库中的数据被处理2次后，存储于附加的多维数据库中，用于简单的多维分析和数据显示，或者复制到数据集市中。三层架构的优势在于其功能的分离，它消除了资源的限制，使得数据集市的建立变得更为简单。

如图2-3所示，在二层架构中，决策支持系统引擎与数据仓库运行于同一硬件平台上，这比三层架构更经济。但是，当大型数据仓库需要数据密集型应用进行决策支持时，二层架构就会出现性能问题。

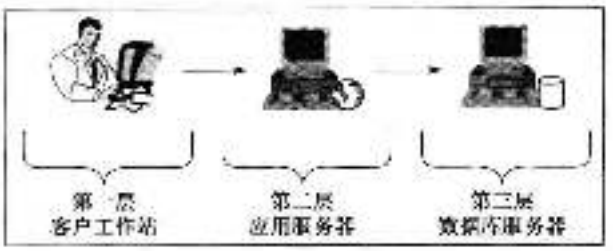


图2-2 三层数据仓库架构

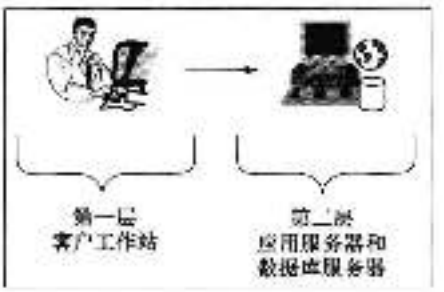


图2-3 二层数据仓库架构

在不考虑组织所处环境和某些特定需求的情况下，大多数人明智地倾向于绝对的方法，即认为某种做法一定要比另一种做法更好。许多咨询顾问和软件供应商只关注系统架构的某一部分，限制了它们的能力和动机，使得它们无法按照组织需要正确地选择架构，这使得架构的选择变得更为复杂化。这些方面早已被提出和研究过。2005年Ball为组织中商务智能的实施提供了决策标准，明确了商务智能实施中多维数据集市的作用，但对于架构的具体层次却未能做出定论。他的标准围绕着数据访问的空间和速度需求的预测问题。

数据仓库和因特网是正确管理企业数据的两大关键技术，它们的结合就是基于WEB的数据仓库。如图2-4所示，基于Web的数据仓库的架构是一种包括PC客户端、Web服务器和应用服务器的三层架构。在客户端，在用户熟悉的图形用户界面（Graphical User Interface，GUI）中，需要因特网连接和最好支持Java应用的网络浏览器，而因特网/内部网/外部网则是客户端和服务器的通信媒介。在服务器端，在数据仓库和应用服务器的支持下，Web服务器对客户端和服务器间的数据流入和流出进行管理。基于Web的数据仓库在数据易于访问、平台独立性和低成本方面的优势极为显著。

（Dragoon，2003）文献指出，美国先锋集团（Vanguard Group）采用基于Web的、三层架构作为企业架构进行数据的集成，向顾客和内部用户提供相同数据。（Anthes，2003）文献指出，希尔顿酒店则借助网络化的企业系统将其所有独立的客户端/服务器（C/S）系统集成为一个二

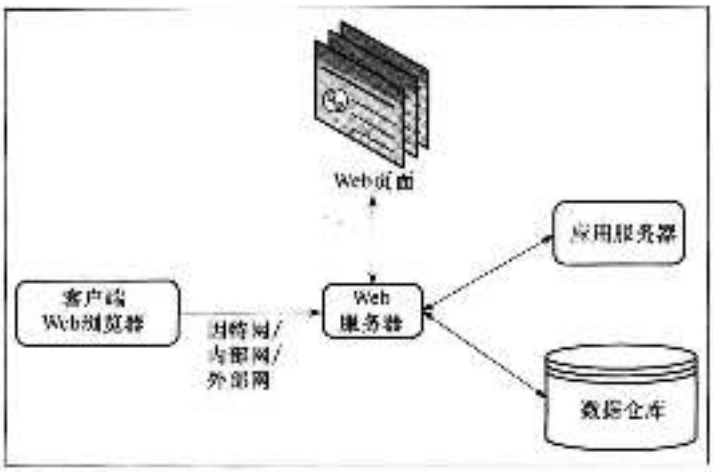


图 2-4 基于 Web 的数据仓库的体系架构

层架构的数据仓库。这一举动为希尔顿酒店带来了 380 万美元的投资（不包括劳动力），波及到 1 500 名用户。公司的处理效率（速度）提高了 6 倍。当数据仓库配置完后，预计每年会为希尔顿酒店节约 450 万美元到 500 万美元。此外，使用 DELL 的聚类技术，也就是并行信息处理技术的辅助下，数据仓库的扩展性以及处理速度都得到了提高。

数据仓库的 Web 架构与其他数据仓库架构的结构是相似的。在数据仓库设计过程中，需要确定 Web 数据仓库到底是安装在交易服务器上还是安装在独立的服务器上。在基于 Web 应用的实际过程中，页面的载入速度极为重要，因此就要仔细计算服务器的承载能力。

当决定使用何种架构时，还需考虑以下几点：

- **使用何种数据库管理系统**（Database Management System，DBMS）？大多数的数据仓库是基于关系数据库管理系统（Relational Database Management System，RDBMS）建立的。Oracle（oracle. com）、SQL Server（microsoft. com/sql/）和 IBM 的 DB2（306. ibm. com/software/data/db2）都是著名的关系型数据库。这些产品都支持 C/S 架构和 Web 架构。
- **是否使用并行处理和分区**？并行处理使得多 CPU 可同时处理数据仓库查询请求，并提高数据仓库的可扩展性。数据仓库设计过程中要考虑到数据分区和划分标准的问题，也就是将数据库中的表拆分为更小的表，以提高数据访问的效率。这对于典型的大数据量存储的数据仓库来说极为重要。2009 年，Furtado 对数据仓库的并行和分区进行了最新研究，而 Teradata 就数据仓库的并行和处理进行了新奇的尝试。
- **是否使用数据迁移工具进行数据加载**？数据从现行系统转移到数据仓库中的过程极为繁琐和耗费人力。依赖于数据资产的多样性和存储位置，数据迁移可能是个简单的过程，或者相反，需要 1 个月时间。迁移工具的使用与否要由现行数据资产的全面评估结果决定，同时还要考虑到这些商业工具的具体性能。
- **使用何种工具进行数据检索和数据分析**？定期使用特定工具对数据进行定位、访问、分析、提取和转换，并最终加载到数据仓库中通常是重要的。但需明确数据迁移工具究竟是自行开发还是从第三方购买，或者直接使用数据仓库系统中的自带工具。而一些过于复杂和实时的迁移需求则需要使用到特定的第三方 ETL 工具。

2.3.1 可选的数据仓库架构

文献（Golfarelli and Rizzi，2009）认为，从设计角度来看，数据仓库架构的最高级别可以分

为企业数据仓库设计和数据集市设计两种。图2-5介绍了除了单纯的企业数据仓库和单纯的数据集市之外，一些介于或超越传统架构的基本数据仓库架构。其中最值得注意的是集中星形拓扑架构（Hub-and-Spoke）和联合架构。这5种架构分别是在由 Ariyachandra and Watson（2005，2006a，and 2006b）中提出的。而在此之前，Sinha（2005）已经研究出了15种不同的数据仓库开发方法。这些方法分别来自核心技术供应商、基础架构供应商和信息建模公司。

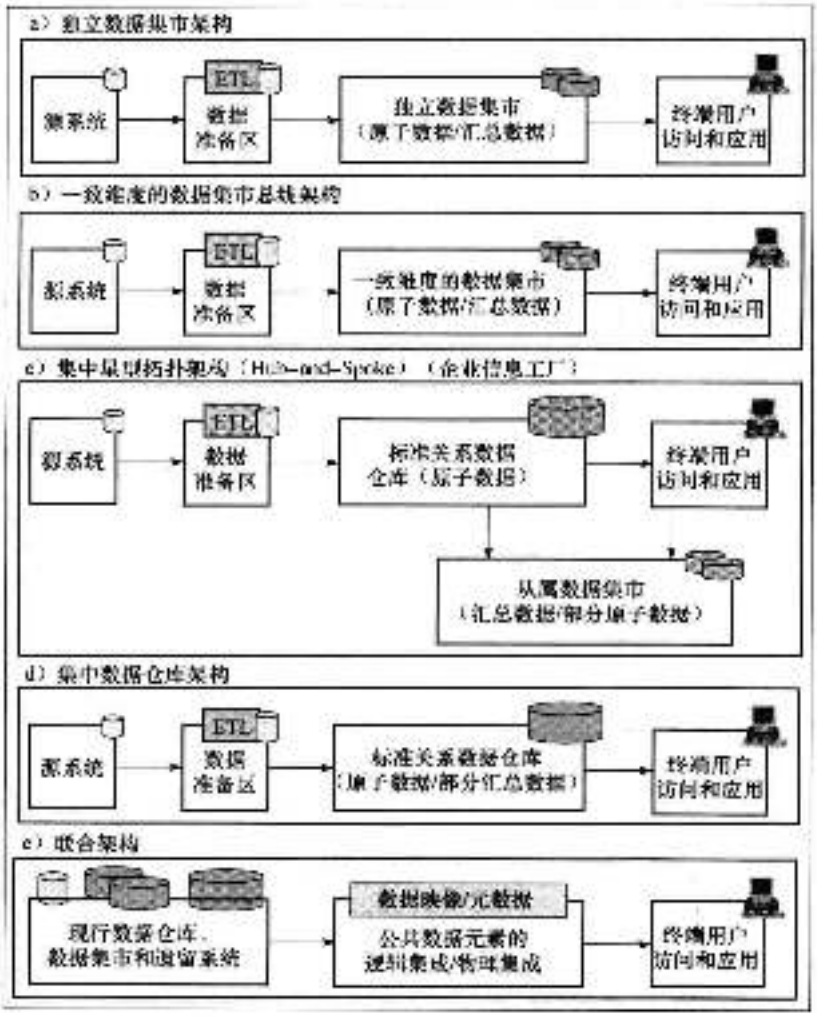


图2-5 可选的数据仓库架构

来源：Adapted from T. Ariyachandra and H. Watson, "Which Data Warehouse Architecture Is Most Successful?" *Business Intelligence Journal*, Vol. 11, No. 1, First Quarter, 2006, pp. 4 – 6.

a. 独立数据集市架构 这一架构被认为是最简单和最低成本的数据仓库架构。数据集市独立运作，为组织的各个单元提供服务。由于其独立性，因此会存在不一致数据定义及不同的维度和度量值，而这使得跨数据集市的数据访问很难实现，原因是数据的唯一性和真实性无法保证。

b. 数据集市总线架构 数据集市总线架构是独立数据集市的可行替代品，适合于多个数据集市被中间件连接的情况。由于在各个单独的数据集市之间数据是相互关联的，因此至少在元数据的级别上，更便于维护整个企业数据的一致性。虽然这种架构允许数据集市间的复杂数据

查询，但是其分析结果的展示却并不令人满意。

c. **星形拓扑架构** 这可能是目前最流行的数据仓库架构。它关注于建立一个包括集中数据仓库和一些服务于专门组织单元的从属数据集市的可扩展、可维护架构，通常用于一个主题接一个主题的迭代式开发。这种架构考虑到了用户接口和报表的简易型和定制化。缺点是缺乏企业全局观，容易形成数据冗余和数据延迟。

d. **集中数据仓库架构** 集中数据仓库架构和星形拓扑架构很相似，不同之处在于，它没有非独立的数据集市，却用一个巨型企业数据仓库服务于所有的组织机构。集中的方法使得用户不再受限于数据集市，可以对数据仓库中的所有数据进行访问。这缩减了技术团队所需转换和更改的数据量，使得数据的管理和监控简单化。如果这种架构设计和实施正确的话，那么只要是在企业内部，无论是谁、无论什么时间、什么地点、都可对企业运行及时和全面的了解。Teradata 公司主张集中数据仓库架构，建议使用没有任何数据集市的数据仓库，如图 2-6 所示。

e. **联合数据仓库架构** 联合数据仓库架构是对自然力量的妥协，是开发一个完美系统的最优方法。它从不同渠道集成分析资源来满足业务的变化。从本质上来说，联合方法需要不同系统的集成。在联合架构中，现行的决策支持架构将被取消，通过需要的数据源访问数据。联合方法需要中间件供应商提供分布查询和连接功能。用户通过使用基于可扩展标记语言（Extensible Markup Language，XML）工具，如数据仓库、数据集市、网站、文档和操作系统等，可以对分布的数据源进行全球监控，当用户选取查询目标并按下查询按钮时，这些工具会对分布的数据源进行自动查询，并将查询结果关联起来，最终展现给用户。大多数专家（Eckerson，2005）认为在性能和数据质量方面，联合方法对数据仓库是一种补充而不是替代。

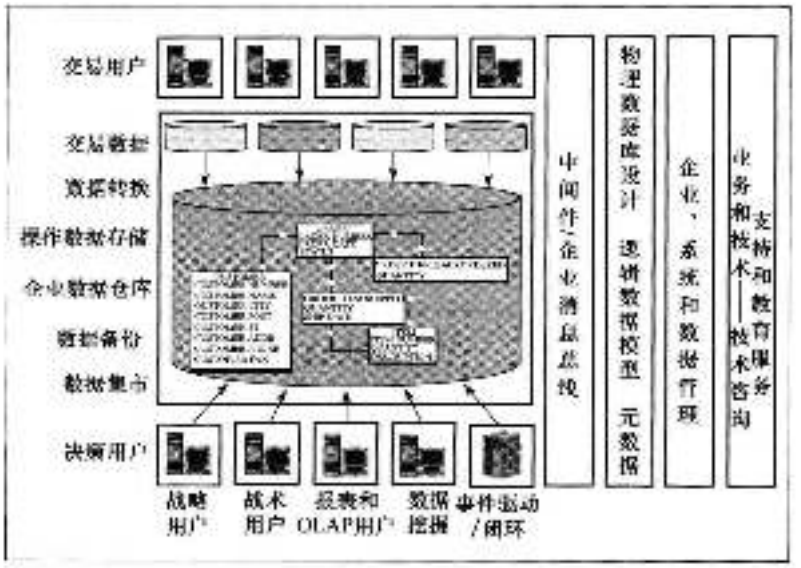


图 2-6 Teradata 公司的企业数据仓库

来源：Teradata 公司（teradata.com），已授权使用。

Ariyachandra and Watson（2005）提出了 10 种影响架构选择的潜在因素，它们分别是：

1. 组织单元间的信息独立性
2. 上级管理对信息的需要
3. 对数据仓库的紧急需要
4. 终端用户任务的性质

5. 资源限制
6. 数据仓库实施前的战略考虑
7. 与现行系统的兼容性
8. 内部员工的理解能力
9. 技术性问题
10. 社会因素和政治因素

这些因素同信息系统项目以及决策支持系统和商务智能项目实施的成功因素相似。通常满足上层管理的信息需求以及设计开发过程中的社会因素或者政治因素这类行为问题，要比技术支持更为重要。尽管我们说技术也是重要的。虽然，每种数据仓库架构都可以实现最有效的应用，并能为组织带来最大的效益，但数据集市却是在实践中效果最差的。有关这方面更详细的信息可以参考 Ariyachandra and Watson（2006a）的相关文献。

2.3.2 哪种架构是最好的

自从数据仓库成为现代企业必不可少的一个部分后，哪种数据仓库架构是最好的就摆上了议题。数据仓库领域的两大学者对其有各自的看法，Bill Inmon 认为星形拓扑架构最好，而 Ralph Kimball 则主张一致维度的数据集市总线架构更为优秀。其他的架构也可能是最优的，但是这两种观点是完全不同的，每种都有强烈的支持者。为了证实究竟哪种架构是最好的，Ariyachandra and Watson（2006b）进行了一项基于经验的研究。他们通过对参与数据仓库实施过程的人员，借助网络调查的方式来收集数据。网络调查的具体内容包括对反馈者的相关信息、反馈者的公司、反馈者的公司所使用的数据仓库以及数据仓库架构的成功与否。

Ariyachandra 和 Watson 最后总计收回了 454 份调查问卷。调查涵盖了从年收入小于 1 000 万美元的小型企业到年收入超过 100 亿美元的大型公司。调查中有 60% 的公司位于美国且从事不同的行业，其中金融服务产业的反馈最多，达到了 15%。调查结果显示，最优秀的数据仓库架构是星形拓扑架构，支持率达到了 39%；接下来是总线架构，支持率为 26%；以及集中架构，支持率为 17%；独立数据集市，支持率为 12%；最后是联合架构，支持率为 4%。主流数据仓库平台的普及率分别是 Oracle 为 41%、Microsoft 为 19% 以及 IBM 为 18%。每种数据架构的平均毛利润也从独立数据集市的 37 亿美元到联合架构的 60 亿美元不等。

Ariyachandra 和 Watson 使用 4 种指标来衡量数据仓库架构的成功与否：（1）信息质量；（2）系统质量；（3）对个人的影响；（4）对组织的影响。每个问题满分为 7 分，分数越高，架构的成功性也越高。表 2-1 表示了每种架构 4 种指标的平均得分。

表 2-1 各种架构成功性的平均评价得分

	独立数据集市	总线架构	星形拓扑架构	集中架构（没有非独立的数据集市）	联合架构
信息质量	4.42	5.16	5.35	5.23	4.73
系统质量	4.59	5.60	5.56	5.41	4.69
对个人的影响	5.08	5.80	5.62	5.64	5.15
对组织的影响	4.66	5.34	5.24	5.30	4.77

Ariyachandra 和 Watson 的研究表明，独立数据集市在所有的指标中都得分最低，这一结论印

证了独立数据集市实践性较差的论点。接下来最低的是联合架构。当公司由于收购和合并存在不同决策支持平台时，至少短期内公司会选取联合数据仓库架构。而这一结论也说明了联合架构并不是一项最优的长期选择。有趣的是，总线架构、集中星形拓扑架构和集中架构的得分却没有太大的差距，因此，在这种评价指标的简单比较下，我们无法证明某种架构比某种架构更为优越。

Ariyachandra 和 Watson 同时也收集了一些关于数据仓库作用域，包括从最小的子单元到全企业范围，以及数据仓库的大小，也就是数据存储量的相关信息。他们发现大多数企业级实施中，以及大型数据仓库会选取星形拓扑架构。此外，他们也收集了不同架构实施所需的成本和时间信息。其中，星形拓扑架构成本最高，且费时最长。

2.3 节复习题

1. 二层架构和三层架构的相同点和不同点是什么？
2. Web 如何影响数据仓库的设计？
3. 列出本章所提到的可选的数据仓库架构。
4. 在开发数据仓库中选取数据仓库架构应注意什么？列出最重要的 10 点。
5. 哪种数据仓库架构最好？为什么？

2.4 数据集成以及提取、转换和加载的过程

在全球化竞争的压力下，对投资回报率（ROI）、管理和投资咨询以及政府法规的要求，使得管理者们开始重新思考如何集成和管理业务。决策者们普遍需要访问整合后的来源不同的数据。在数据仓库、数据集市和商务智能套件产生之前，数据源的访问是一项重大而艰苦的工作。即使在当今基于 Web 的现代数据管理工具下，访问什么数据以及如何将数据展示给决策者也是需要数据库专业人员才能解决的非平凡工作。随着数据仓库容量的增加，数据集成也在逐步发展。

业务分析需要进一步发展。合并和收购的发生、对监管的要求以及新渠道的引入都将驱动商务智能需求的改变。除了历史的、清洗后的、合并的以及时间点的数据外，业务用户对即时的、非结构化的和远程数据的需要也在逐渐增加，而这些数据都要与数据仓库中的内容相集成。（Edward，2003）文献指出通过 PDA、语音识别和语音合成进行数据访问越来越普遍，这使得数据集成变得更为复杂。越来越多的集成问题开始出现在企业系统中。Orovic（2003）文献列出了集成项目中作用因素和非作用因素。在不同数据库间或者不同数据源间实现正确的数据集成是困难的。（Nash，2002）文献指出，一旦数据集成失败，就会给企业系统带来灾难，这些系统包括 CRM、ERP 和供应链系统。

2.4.1 数据集成

数据集成包括 3 个重要阶段。一旦数据集成成功，数据和 ETL、分析工具和数据仓库环境均可被访问。这 3 个阶段分别是：数据访问，也就是从数据源中访问和提取数据的能力；数据合并，也就是不同数据源间的业务集成；变化捕捉，即基于企业数据源变化的识别、捕捉和传送。应用案例 2.3 为我们讲解了 BP 海湾油公司如何从数据仓库实施中的数据集成获得好处。此外，像 SAS 软件公司这样的产品供应商也已开发出强大的数据集成软件。SAS 企业的数据集成服务器包括在集成过程中能够提高数据质量的客户数据集成工具。Oracle 的商务智能套件也同样支持数据集成。

应用案例2.3 BP润滑油商务智能和全球标准化项目的巨大成功

为了实现全球发布信息的一致性和信息管理的透明性，BP润滑油公司在并购后开始实行其商务智能和全球标准化（Business Intelligence and Global Standards，BIGS）项目。与即时商务智能一样，BIGS为诸如财务、市场、销售以及供应和物流等环节提供了一致和详细的信息展示。

BP润滑油是世界上最大的石油石化集团之一，作为BP公共有限公司集团的一部分，在全球汽车润滑油市场中名列前茅。BP公司最出名的润滑油品牌是Castrol，它的业务范围超过了100个国家，雇员人数高达1万。在战略上，BP润滑油采取以客户为中心，致力于提高其在机动车市场的效率。在最近的并购活动后，BP公司抓住了其快速成长的机会，公司的效能和灵活性都得到了进一步的提高。

挑战

并购后，BP润滑油希望提高其信息管理和商务智能的一致性、透明性和可访问性。要实现这一目标，公司就必须对其不同源系统中的数据进行集成，以避免ERP标准化系统引入的延迟。

解决方案

由于信息管理和商务智能的战略考虑，BP润滑油开始率先实施BIGS。BIGS的核心是Kalido，一种能够实现准备、实施、运营和管理数据仓库的自适应企业数据仓库解决方案。

Kalido的合并企业数据仓库解决方案支持这项BIGS所需的复杂的数据集成以及多变的报表需求。为了适应项目对报表的需求，这一软件同时还允许在所有信息完全保存的前提下，轻松地实现信息架构基础的快速修改。系统集成和存储多种源系统中的信息，为以下各环节提供了一致的数据支持：

- **市场** 通过深度探讨发票层详细信息，可以观察到客户收益以及细分市场利润
- **销售** 销售发票报告提高了关税成本和实际支付额
- **财务** 具备审计能力的全球标准化的损益表、资产负债表和现金流量表；客户负债管理供应和物流；订单同步以及跨多个ERP平台的动态流程

收益

通过提高数据的可见性和即时性，BIGS可以提供大量信息支持业务机会的识别，以实现公司效益的最大化，并关联成本的管理。由于BIGS项目中的数据同步，而为BP公司带来的好处主要有：

- 提高业务数据的一致性和透明性
- 轻松、快速和更灵活的报表
- 全球标准和当地标准的适应
- 快捷、低成本以及灵活的实施周期
- 现行业务流程和日常业务中断的最小化
- 识别数据质量问题，并做出解决方案
- 提高对新业务机会的智能反应能力

来源：Kalido，"BP Lubricants Achieves BIGS，Key IT Solutions，" keyitsolutions.com/asp/rptdetails/report/95/cat/1175/（accessed August 2009）；Kalido，"BP Lubricants Achieves BIGS Success，" kalido.com/collateral/Documents/English-US/CS-BP%20BIGS.pdf（accessed August 2009）；and BP Lubricant homepage，bp.com/lubricanthome.do（accessed August 2009）.

数据仓库的主要目的是集成不同系统中的数据。提供数据和元数据集成的集成技术有：

- 企业应用集成
- 面向服务的架构
- 企业信息集成
- 数据提取，转换和加载

企业应用集成（Enterprise Application Integration，EAI）是从源系统向数据仓库中推送数据的媒介，具备集成应用的功能。EAI关注系统间功能的共享，而不是数据，使系统变得更为灵活且具有重用性。传统上，EAI解决方案主要关注应用程序接口层的应用重用问题；而现在，由于使用定义和文档良好的粗粒度SOA架构，即业务流程或者功能集成的引入，EAI也得到了进一步的完善。WEB服务就是一种实施SOA架构的专门手段。EAI可以被用于在准实时数据仓库中进行数据查询，或者将决策信息传送至OLTP系统中。EAI的实施手段和工具有很多种。

企业信息集成（Enterprise Information Integration，EII）允许，如关系数据库、Web服务以及多维数据库之类的多个数据源间的实时数据集成。EII是一类从源系统中提取数据以满足信息需求的机制。EII工具使用预定义的元数据以视图的模式将集成后的数据展现给终端用户。XML（Kay，2005）文献指出，XML是EII最为重要的一部分，XML使数据在创建和之后的使用中都被标注，这些标注可以被扩展和修改以适应任何知识领域。

物理数据的集成已经成为在数据仓库和数据集市中建立数据集成视图的惯例做法。XML（Kay，2005）文献指出，随着EII工具的出现，虚拟数据的集成也变得可能。Manglik and Mehra讨论了新数据集成类型的好处和不足，这种类型将传统物理方法扩展到一种全面的企业视角。

接下来我们将讨论向数据仓库中加载数据的方法：ETL。

2.4.2 提取、转换和加载

数据仓库的核心技术流程是：**提取、转换和加载**（Extraction，Transformation and Load，ETL）。ETL技术已经存在了一段时间了，对数据仓库的流程和使用有帮助。ETL流程是任何以数据为中心项目的集成组件。ETL通常需要占用以数据为中心的项目中70%的时间，这对于任何一个IT管理人员来说都是一种挑战。

ETL流程包括提取（也就是从一个或多个数据库中读取数据）；转换（即将提取后的数据由一种数据类型转换为另一种所需的数据类型，以便于存储于数据仓库或者其他简单的数据库中）以及加载（也就是将数据存入数据仓库中）。转换通常发生于规则使用、表格查询或者数据合并中。这3种数据库功能被集成于一类工具中，用于将数据从一个或多个数据库中提取出来并加载入另一个数据库或者数据仓库中。

ETL工具常常在不同的源和目标间进行数据传送，并记录在源和目标间移动时的数据元素（比如元数据）变化，在必要时与其他的应用交换元数据，并监控所有运转的流程和操作（比如调度计划、错误管理、检查日志和统计数据等）。ETL对数据集成和数据仓库同样重要。ETL的目的是向数据仓库中加载集成和清洗后的数据。ETL流程中使用的数据可以来自不同的数据源：大型机应用、ERP应用、CRM工具、平面文件、Excel电子数据表，甚至是消息队列。图2-7描绘了ETL流程。

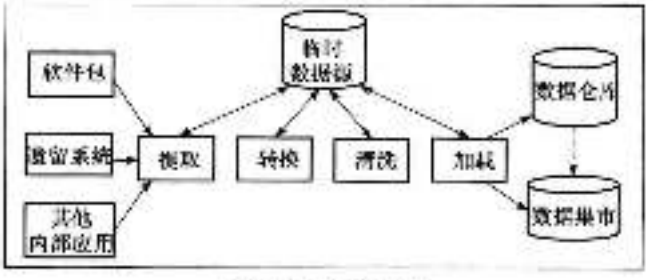


图2-7 ETL流程

将数据迁移到数据仓库中，需要从关联数据源中提取数据。数据源包括从 OLTP 数据库、电子数据表、个人数据库（如 Microsoft Access），以及外部文件中提取的文件。通常，所有的输入文件先被写到一个预先设计的、用于加速加载流程的临时表中。数据仓库包括大量的业务规则，这些规则定义了数据如何使用、概括规则、编码属性的标准化以及计算规则。在数据加载到数据仓库之前，源文件的数据质量必须是准确无误的。良好定义的数据仓库的优点之一就是这些规则能存储在元数据仓库中，并且可以直接用于数据仓库中。这一点与 OLTP 的做法不同，OLTP 中的数据和业务规则是分散于整个系统中的。数据仓库中数据加载的过程既可以在业务规则的开发和维护中，借助于提供图形用户界面（Graphical User Interface，GUI）的数据转换工具来实现，也可以通过诸如 PL/SQL、C++、.Net 等编程语言，自行编写软件或实用工具来加载数据仓库这种传统方法来实现。这一决定对组织来说通常是艰难的，当组织确定购买数据转换工具或者自行编写数据转换程序时，会受到以下很多因素的影响：

- 数据转换工具很贵
- 学习数据转换工具费时较长
- 在学会使用数据转换工具前，无法估量 IT 组织做得如何

从长期来看，转换工具应该简化数据仓库的维护，同时使检测和纠错更有效率（也就是将数据中异常部分去除）。OLAP 和数据挖掘工具将依赖于数据转换的效果。

（Songini，2004）文献指出，作为 ETL 使用的成功范本，Motorola 公司借助 ETL 工具从 30 个不同的采购系统中收集数据并将其传送到其全球 SCM 数据仓库中，进行公司总支出的分析。

Solomon（2005）将 ETL 技术划分为 4 大类：复杂的、可使用的、简单的和基础的。通常，我们认为复杂的 ETL 技术会使数据仓库项目文档完备且管理精确。

尽管自行开发 ETL 工具是可行的，但是使用现行的 ETL 工具更为简单。根据文献（Brown，2004），以下是一些选择 ETL 工具的重要标准：

- 多个数据源架构下的数据读写能力
- 自动捕获和传输元数据
- 符合开放标准的历史
- 是否为开发者和用户提供了简单易用的界面

ETL 的广泛使用标志着数据管理的匮乏以及相关数据管理策略的缺少。Karacsony（2006）曾提出冗余数据的扩大化和 ETL 流程的数量之间存在着某种直接的关系。当数据作为一项企业资产被正确管理时，ETL 的效果大大地降低，同时冗余数据也会全部消除。这导致了在数据质量改善的同时，维护过程中大量数据的存储以及新开发的高效执行。ETL 设计的不合理将会大大增加维护、转换以及更新的成本。因此，正确选择开发和维护 ETL 过程所使用的技术和工具是非常重要的。

但一定数量的 ETL 软件包也是有用的。数据库供应商目前提供的 ETL 软件包与独立的 ETL 工具相比功能上略有提高。SAS 意识到数据质量的重要性，提出了工业界第一个完全集成的解决方案，它将 ETL 和数据质量两方面因素结合起来，使得数据真正转换为有战略价值的资产。除了 SAS 外，市场上还有一些其他的 ETL 软件商，如微软、Oracle、IBM、Informatica、Embarcadero 和 Tibco。想知道 ETL 更详细的信息，请参见 Golfarelli and Rizzi（2009）、Karacsony（2006）以及 Songini（2004）相关文献。

2.4 节复习题

1. 数据集成是什么？
2. 描述 ETL 流程的 3 个阶段。

3. 为什么ETL过程对于数据仓库的结果如此重要？

2.5 数据仓库的开发

数据仓库对于任何组织来说都是一个大工程，它要比一个简单的主机选择和项目实施要复杂得多，它不仅涉及和影响到组织内的许多部门以及输入输出接口，同时还要作为CRM商业战略的一部分。数据仓库所带来的好处可以分成直接利益和间接利益两类，其中直接利益包括以下几方面：

- 终端用户可以以多种方式进行广泛的分析。
- 实现企业数据的一致性，也就是事实的单一版本。
- 更好和更及时的信息，数据仓库允许从高成本的操作系统转向低成本的服务器的信息处理，因此终端用户的信息请求得以快速执行。
- 提高系统性能。由于一些操作系统的报表请求被转向DSS，因此数据仓库可以没有生产处理。
- 数据访问更为简便。

间接利益来源于终端用户享用直接利益的过程。总的来说，这些利益提高了企业的业务知识，展现了企业的竞争优势，提升了客户的服务及其满意度，促进了决策的制定，并且有助于业务流程的改进，这些即是对提升企业竞争优势最强有力的帮助。（想要了解数据仓库如何为企业提升竞争优势的具体讨论，可以参见（Parzinger and Frolick，2001）；想要知道组织如何获取额外回报的详细讨论，可以参见Waston et al.（2002）。）考虑到数据仓库可以带来的潜在利益以及一个项目在时间和金钱两方面所需要的大量投资，组织如何构建数据仓库使得成功的机会最大化是非常重要的。除此之外，组织必须显而易见地考虑成本。Kelly（2001）文献中描述了一个考虑管理者（也就是通过改进传统决策支持功能来节约成本）、采集者（通过自动化进行信息采集和传播来节约成本）和使用者（通过使用数据仓库制定决策来节约成本或获利）这三类人利益的投资回报率方法。这些成本包括硬件、软件、网络带宽、内部开发、内部支持、培训、外部咨询等方面，而净现值（Net Present Value，NPV）则以超过数据仓库的预期使用年限来进行计算。这些利益被上述三方瓜分，其中管理者占将近20%，采集者占30%，而使用者则占到50%。Kelly认为随着组织的改变，使用者是否参与到数据仓库开发过程中，将作为衡量系统是否成功的要素。

应用案例2.4介绍了日本Hokuriku Coca-Cola瓶装公司（HCCBC）的数据仓库开发过程及其为公司带来的巨大竞争优势。这个系统如此成功，按照计划Coca-Cola自动贩卖机在日本超过了100万台。

应用案例2.4 Coke的数据仓库让事情变得越来越好

面对竞争压力和消费需求，一个成功的瓶装公司如何确保自动贩卖机带来收益？对于HCCBC来说，这个问题的答案就是数据仓库和Teradata的分析软件。HCCBC建立数据仓库的主要原因是为了向其竞争对手——Mikuni美国公司建立数据仓库系统进行反击。其数据仓库不仅收集历史数据，同时还从各个自动贩卖机中收集近实时数据。它将每一个自动贩卖机看为一个门店，自动贩卖机中的数据通过无线网络传送至总部。这一项目开始于2001年，数据仓库提供了详细的产品信息，包括每次销售发生的具体时间和日期、某种产品卖出的时间、某个顾客是否少找钱了以及某个机器是否发生了故障。在任何一种情况下，都会触发警报，自动贩卖机通过

无线传输系统将报告直接传送到数据中心。美国可口可乐公司曾使用调制解调器进行自动贩卖机和经销商间的信息传送长达10余年。

2002年，HCCBC进行了一项初步试验，将其所有位于长野县的全部自动贩卖机连接到无线网络收集每台的近实时销售网点的数据。结果令人震惊，所有的贩卖机都可以精确预测需求和快速识别问题，销售总额也增加了10%。除此之外，由于贩卖机服务的精准化，延时和其他费用减少了46%。另外，每个店员能够服务的自动贩卖机的数目提高了42%。

由于这次试验的圆满成功，HCCBC计划采用实时数据仓库，将这一改动扩大到全公司范围，将近6万台机器。最终，这一数据仓库解决方案将跨过企业边界的界限，进入到整个可口可乐瓶装网络中。这样，全日本超过100万台的自动贩卖机都将接入网络，这一切将极大地缩减公司的成本，并为公司带来更大的收益。

来源：Adapted from K. D. Schwartz, "Decisions at the Touch of a Button," *Teradata Magazine*, teradata.com/t/page/117774/index.html (accessed June 2009); K. D. Schwartz "Decisions at the Touch of a Button," *DSS Resources*, March 2004, pp. 28-31, dssresources.com/cases/coca-colajapan/index.html (accessed April 2006); and Teradata Corp., "Coca-Cola Japan Puts the Fizz Back in Vending Machine Sales," teradata.com/t/page/118866/index.html (accessed June 2009).

对于一个成功的数据仓库项目来说，业务对象的清晰定义、管理层对项目的支持、合理的时限和预算以及管理期望都是必不可少的。数据仓库战略是数据仓库成功引入的蓝图，这一战略需要明确企业的预期目标、动机以及实现目标后的进一步计划。这就需要考虑到组织的前景规划、架构和文化。Matney（2003）文献中提出了一系列可以帮助企业进行数据仓库开发的灵活高效战略。一旦建立数据仓库的计划和相应支持到位后，企业就需要仔细核查数据仓库的供应商。表2-2是目前市面上数据仓库供应商的一部分，更多的可以参见数据仓库协会（twdi.com）和《DM Review》（dmreview.com）。多数供应商会提供其数据仓库和商务智能产品的软件演示。

表2-2 部分数据仓库供应商

供应商	产品
Computer Associates（cai.com）	数据仓库工具和产品的综合套件
DataMirror（datamirror.com）	数据仓库管理和性能产品
Data Advantage Group（dataadvantage.com）	元数据软件
Dell（dell.com）	数据仓库服务器
Embarcadero Technologies（embarcadero.com）	数据仓库管理和性能产品
Business Objects（businessobjects.com）	数据清洗软件
Harte-Hanks（harte-hanks.com）	CRM产品和服务
HP（hp.com）	数据仓库服务器
Hummingbird有限公司（hummingbird.com）	数据仓库引擎和探索仓库
Hyperion Solution（hyperion.com）	数据仓库工具、产品和应用的综合套件
IBM（ibm.com）	数据仓库工具、产品和应用
Informatica（informatica.com）	数据仓库管理工具和产品
Microsoft（Microsoft.com）	数据仓库工具和产品
Oracle（包括PeopleSoft和Siebel）（oracle.com）	数据仓库、ERP和CRM工具、产品和应用
SAS协会（sas.com）	数据仓库工具、产品和应用
Siemens（Siemens.com）	数据仓库服务器
Sybase（Sybase.com）	数据仓库产品和应用的综合套件
Teradata（teradata.com）	数据仓库工具、产品和应用

2.5.1 数据仓库供应商

McCloskey（2002）列举了选择供应商时所应遵循的6项原则，分别是：经济实力、与ERP系统的

关联性，合格的咨询顾问、市场份额、行业经验以及之前建立的合作关系。企业可以通过展会和公司网站获取供应商的具体信息，也可直接向供应商询问具体的产品信息。Van den Hoven（1998）文献区分了3种数据仓库产品的不同之处。第一类是用于处理像是数据定位、数据提取、数据转换、数据清洗、数据传输以及数据加载之类的功能。第二类则是一种类似数据库引擎的数据管理工具，用于数据仓库和元数据的存取和管理。第三类是一种数据访问工具，可供终端用户在数据仓库中进行数据分析。这类数据仓库产品包括查询生成器、可视化、EIS、OLAP以及数据挖掘。

2.5.2 数据仓库开发方法

很多组织都需要建立数据仓库进行决策支持，而它们所采用的方法有两种。第一种方法是数据仓库之父Bill Inmon提出的，他主张由上向下的开发方法，使传统的关系数据库能够适应整个企业范围内数据仓库开发的需要，也就是EDW开发方法。第二种方法是Ralph Kimball提出的，他主张运用维度建模由底向上的开发方法，这也是数据集市的开发方法。

Breslin（2004）文献认为，知道这两种方法的相同和不同之处，有助于我们理解数据仓库的基本概念。表2-3对两种方法进行了详细的对比，下面将具体介绍这两种方法。

Inmon的模型：EDW方法 Inmon的方法强调由上向下进行开发，并使用实体关系图（Entity-Relationship Diagram，ERD）和螺旋式开发等数据库开发方法和工具。EDW的开发方法并不排斥建立数据集市。EDW是一种理想的开发方法，它提供了一致和全面的企业观。Murtaza（1998）文献提出了开发EDW的框架。

Kimball模型：数据集市方法 Kimball的数据集市策略是一种"大计划、小实施"的方法。数据集市是一类面向主题或者面向部门的数据仓库，它是数据仓库的缩小版，主要关注某个具体部门的应用请求，例如市场或者销售部门。这一模型采取了由数据表入手的维度建模技术。Kimball提倡由底向上的开发方法，以便于在数据仓库建立的同时完成数据集市的建立。

表2-3 EDW和数据集市开发方法的对比

评价指标	数据集市方法	EDW方法
范围	单主题域	多主题域
开发时间	数月	数年
开发成本	1万美元－10万美元以上	100万美元以上
开发难度	低或中等	高
开发前的数据准备	业务知识	企业知识
数据源	少数操作系统和外部系统	多数操作系统和外部系统
大小	兆字节－吉字节	吉字节－太字节
时间范围	五至十年历史数据	历史数据
数据转换	低或中等	高
更新频率	每小时、每天、每周	每周、每月
技术硬件	工作站和部门服务器	企业服务器和大型计算机
操作系统	Windows和Linux	Unix，Z/OS，OS/390
数据库	工作站或标准数据库服务器	企业数据库服务器
用法		
并发用户数	10	100－1 000
用户类型	业务层分析师和管理人员	企业分析师和高级管理人员
商业焦点	业务领域活动的最优化	组织能最优化，支持企业决策

来源：Based on J. Van den Hoven, "Data Marts: Plan Big, Build Small," in *IS Management Handbook*, 8th ed., CRC Press, Boca Raton, FL, 2003; and T. Anjanchandra and H. Watson, "Which Data Warehouse Architecture Is Most Successful?" *Business Intelligence Journal*, Vol. 11, No. 1, First Quarter 2006, pp. 4-6.

哪种模型更好 没有通用的、一成不变的策略适用于所有的数据仓库。随着用户需求、企业业务需求以及企业在数据源管理方面的成熟度的变化，企业的数据仓库策略可以从简单的数据

集市发展到复杂的数据仓库。对于许多企业来说，除了向业务用户展现更好的访问数据所能带来的好处外，数据集市还是获取数据仓库实现和管理经验的首要一步。除此之外，数据集市往往还能显示数据仓库的商业价值。最终，获得 EDW 是理想的（参见应用案例 2.5）。然而按照开发 EDW 的方法开发独立数据集市通常会为组织带来更大的好处，特别是在组织不能或者不愿开发大规模项目的时候。数据集市证明了可行性及其所带来的种种好处，这些将带来对 EDW 的投资。表 2-4 总结了这两类模型的本质特征的不同。

表 2-4 Inmon 模型和 Kimball 模型的本质区别

特征	Inmon	Kimball
方法与架构方法	自顶向下	自底向上
架构结构	企业数据仓库支持部门数据库	数据集市对一个单独的业务活动建模，通过一致维度的数据总线实现企业数据的一致性
方法的复杂度	相当复杂	相当简单
与开发方法相比	源于螺旋形方法	4 步流程：是关系数据仓库管理的一个分支
物理设计的考虑	较完全	不完全
数据建模		
数据定位	面向主题或是数据驱动	面向业务流程
工具	传统的 ER 图、数据流图	多维建模；是关系建模的分支
终端用户的可访问性	低	高
主要用户	IT 专家	终端用户
组织中的定位	企业信息工厂的集成部分	操作数据的转换和保留
目标	基于已被印证的数据库方法和技术而实现的一种可行的技术解决方案	出于优化终端用户在一定的响应时间内直接查询数据的目的而实现的解决方案

来源：Based on M. Breslin, "Data Warehousing Battle of the Giants: Comparing the Bases of Kimball and Inmon Models," *Business Intelligence Journal*, Vol. 9, No. 1, Winter 2004, pp. 6-20; and T. Ariyachandra and H. Watson, "Which Data Warehouse Architecture Is Most Successful?" *Business Intelligence Journal*, Vol. 11, No. 1, First Quarter 2006.

应用案例 2.5 HP 将数百个数据集市合并为一个企业数据仓库

2005 年 12 月，Hewlett-Packard 决定将其全球 762 个数据集市合并为一个企业数据仓库。HP 意图通过这一做法获得超前的商业意识，明确如何更好地为客户提供服务。HP 的总裁兼首席执行官 Mark Hurd 声称，企业内部对于分析数据的"如饥似渴"错误地导致了大量数据集市的建立。这些数据孤岛的设计和维护都极为昂贵，同时也没有为 HP 带来其所需的企业内部信息和客户信息。2006 年中旬，HP 开始将其数据集市中的数据合并到一个新的数据仓库中，而所有的这些数据集市将被完全清除。

来源：Based on C. Martins, "HP to Consolidate Data Marts into Single Warehouse," Computerworld, December 13, 2005.

2.5.3 数据仓库开发的其他思考

一些组织想要外包数据仓库。他们既不想处理硬件和软件的请求，也不想管理信息系统。这类问题的一个办法就是使用托管数据仓库。在这种情景下，另外一家公司会拥有数据仓库开发和维护的丰富经验和专业知识。但是，这种方法还要考虑到数据的安全和隐私问题。技术前沿

2.1 为我们介绍了托管数据仓库的更详细内容。

技术前沿 2.1 托管数据仓库

一个托管数据仓库拥有并不亚于现场数据仓库，并具有与其几乎相同的功能，但是它不消耗客户端的计算机资源。托管数据仓库提供了计算机升级、网络升级、软件认证、内部开发、内部支持和维护的成本优势，并且提供商务智能服务。

托管数据仓库具有如下优点：

- 需要最小的基础设施投资
- 内部系统的开放能力
- 释放现金流
- 提供强有力的解决方案
- 实施强有力的解决方案以支持增长
- 提供高质量的设备和软件
- 实现快速的连接
- 实现远程获取数据
- 帮助公司专注于其核心业务
- 满足大数据量的存储需求

尽管具有以上优点，但托管数据仓库却并不一定适合每个组织。一个财政收入超过 500 万美元的大公司在没有充分利用互联网基础设施和 IT 员工的情况下会损失资金。其次，公司认为引入外包应用程序会导致他们失去对数据的控制，那么就不会依赖于商务智能服务提供商。最后，不利于托管数据仓库的最重要的主要最普遍的因素是外包敏感应用程序是不明智的，因为在安全和隐私方面存在隐患。

来源：Based on M. Thornton and M. Lampa, "Hosted Data Warehouse," *Journal of Data Warehousing*, Vol. 7, No. 2, 2002, pgs. 27-34; and M. Thornton, "What About Security? The Most Common, but Unwarranted, Objection to Hosted Data Warehouses," *DM Review*, Vol. 12, No. 3, March 18, 2002, pgs. 30-43.

2.5.4 数据仓库中的数据表示

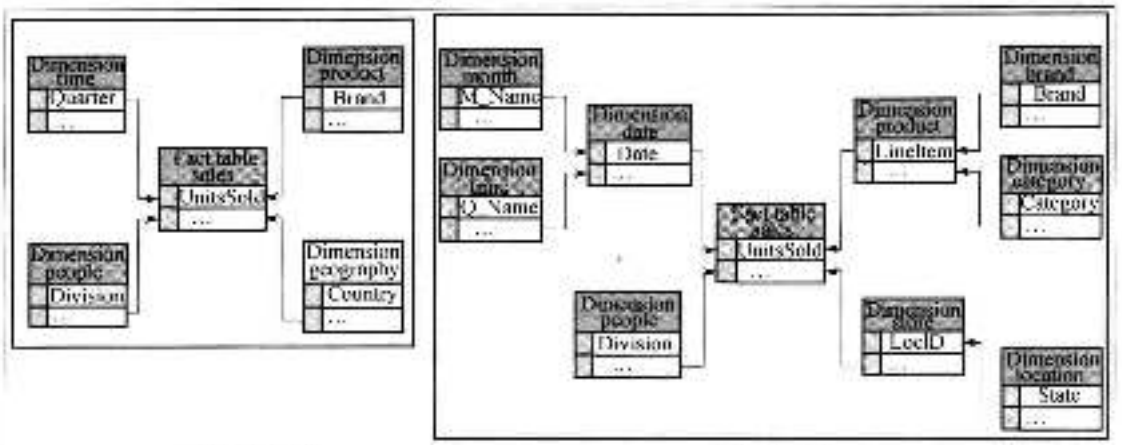
图 2-1 表示了一种典型的数据仓库结构。数据仓库架构也有很多改进版（见图 2-5）。不论是何种架构，数据仓库中的数据表示一直都是基于维度建模的理念。**维度建模**是一个支持大量查询访问的基于检索的系统。数据仓库中数据的存储和表示不仅要适合并且能提高复杂多维查询的处理能力。通常，星形模式和雪花模式是数据仓库中实现维度建模的方法。

星形模式（有时被称做星形关联模式）是最普遍使用和最简单的维度建模。一个**星形模式**包含一个中心事实表和多个相关的**维度表**（Adamson，2009）。事实表包含了大量与观测事实和外部链接（例如，外键）相对应的行数据。事实表包含了用来进行决策分析和查询报表的描述属性，外键用来链接维度表。决策分析属性包括性能测量指标、操作指标、聚集度量值（例如，销售数据、客户保留率、毛利润、产品成本、废品率）和其他所有指标，这些指标用来分析企业的业绩。换句话说，事实表主要解决了数据仓库用什么来支持决策分析的问题。

围绕在中心事实表周围（通过外部键相连）的是维度表。维度表包含了中心事实表列数据的分类和聚合信息。维度表包含用以描述事实表数据的属性，并对数据进行分析和总结。维度表与中心事实表的行具有一对多的关系。在查询中，维度可以对事实表中的数据值进行切片和切块，以满足特定的信息需求。星形模式使得只读数据库结构具有如下特点：快速的查询响应时间、简易化和维护简单。图 2-8a 展示了一个简单的星形模式。星形模式是雪花模式的一种特殊情形。

雪花模式是多维数据库中表的逻辑排列，其实体关系图表现为雪花状。与星形模式相似的

是，雪花模式由中心事实表（通常只有一个）表示，中心事实表与复杂维度相连。在雪花形模式中，维度被标准化为多张维度表，而星形模式中的维度被非规范化为单个维度，单个维度由单张表表示。图 2-8b 展示了一个简单的雪花模式。



a）星形模式　　b）雪花模式

图 2-8　多维表

2.5.5　数据仓库中的数据分析

当数据被正确地存储存在数据仓库后，可以采用多种方法使用数据支持组织的决策制定。毫无疑问，OLAP 是数据仓库领域中应用最广的数据分析技术，并且由于数据容量的指数增长以及数据驱动分析的商业价值逐渐得到了认可，OLAP 已日益普及。通过对组织数据资源库（例如数据仓库、数据集市）的多维分析查询，OLAP 能够快速解决特定的问题。

2.5.6　OLAP 与 OLTP

OLTP 是用于描述交易处理系统的专业术语，这些交易处理系统主要负责获取和存储与日常业务相关的数据，如 ERP、CRM、SCM、POS 等。OLTP 系统处理关键业务需求，使日常业务交易自动进行并产生实时报表和常规分析。但是，OLTP 系统不能进行大数据量的特定分析和复杂查询。另一方面，通过对组织数据更高效率的特定分析，OLAP 能够满足以上需求。OLAP 与 OLTP 紧密联系：OLAP 通过 OLTP 获取数据，而 OLTP 使业务交易自动进行，OLAP 制定的决策管理着这些业务交易。表 2-5 展示了 OLTP 和 OLAP 之间的区别。

2.5.7　OLAP 操作

OLAP 中最主要的操作结构是基于称为立方体的概念。OLAP 中的立方体是一种支持快速数据分析的多维数据结构（实际的或虚拟的）。它也可以被定义为能够进行多维度高效率操作和数据分析。立方体中的数据结构旨在克服关系型数据库的局限性：关系型数据库不适合大数据量的实时分析。相反，关系型数据库更适合一系列的交易操作（增加、删除和修改数据）。尽管关系型数据库中有很多报表生成工具，但是执行涉及多张数据库表的多维查询却很慢。

通过改变数据定位和定义分析计算，分析人员运用 OLAP 可以通过数据库和计算机屏幕浏览数据的一个特定子集（和其随时间不断发展的子集）。这些由用户发起的、通过规范切片（经过旋转）和上钻/钻取（经过聚集和分类）获取数据的动作有时被称做“切片和切块”。通常使用的 OLAP 操作包括切片和切块、钻取、上卷和旋转。

表2-5 OLTP与OLAP的对比

标准	OLTP	OLAP
用途	执行日常业务功能	决策支持，提供业务和管理查询功能
数据源	交易数据库（专注于效率和连贯性的标准化数据资源库）	数据仓库和数据集市（专注于准确性和完整性的标准化数据资源库）
报表	常规、定期、集中关注的报表	特定的、多维的、广泛关注的报表和查询
资源需求	普通关系型数据库	多处理器、大存储量的专业数据库
处理速度	快速（记录业务交易和常规报表）	缓慢（密集的、复杂的、大规模的资源查询）

- **切片**：切片是多维数组的子集（通常是二维表示），它与一个或多个不属于本子集的维度数值相对应。图2-9展示了对一个三维立方体进行的简单切片操作。
- **切块**：切块操作是对多维数据立方体，按二维以上进行的切片操作。

钻取/上钻：钻取或上钻是一种特定的OLAP技术，借此用户可以获取最概括到最详细的数据。

上卷：一个上卷动作包括计算所有一维或多维的数据关系。为此，应该定义一个计算关系或公式。

旋转：旋转是用来改变报表或特定查询的维度方向。

OLAP的种类 OLAP分为几类；其中，ROLAP、MOLAP和HOLAP使用得最为普遍。

关系型在线分析处理（Relational Online Analytical Processing，ROLOP）是多维在线分析处理（Multidimensional OLAP，MOLAP）技术的替代。ROLAP和MOLAP分析工具都是使用多维数据模型分析数据，ROLAP的不同之处在于它不需要进行预先计算和信息存储。相反，当终端用户需要时，ROLAP工具将获取关系型数据库中的数据并生成SQL查询来计算适当级别的信息。ROLAP可以产生附加数据库表（汇总表或聚集），这些表可以总结在任意维度组合下的数据。应当谨慎设计ROLAP使用的关系型数据库。与OLTP使用的数据库相比，ROLAP使用的数据库性能良好。因此，ROLAP还创建数据的额外备份。

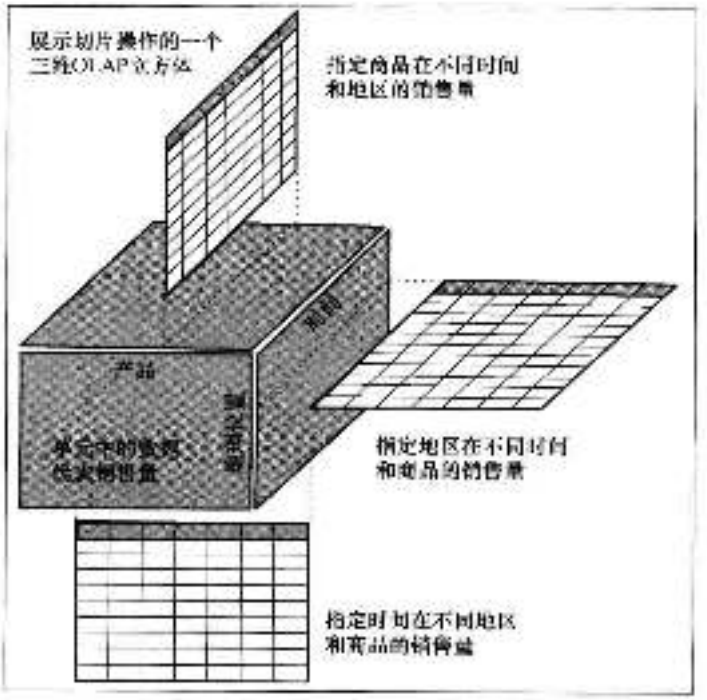


图2-9 简单三维数据立方体的切片操作

MOLAP 是 ROLAP 技术的替代。MOLAP 与 ROLAP 的显著差异在于它需要在多维数据立方体中进行预先计算和信息存储，这些操作被称作散处理。MOLAP 将数据存储在优化后的多维数组仓库中，而不是关系型数据库中（通常 ROLAP 使用关系型数据库）。

MOLAP 和 ROLAP 之间的不良交易需要使用 ETL，并使得查询速度缓慢。人们创建了更好的查询方法使以上两种方法的优缺点都得到优化。这些查询方法形成了混合型在线分析处理（Hybrid Online Analytical Processing，HOLAP），它结合了 ROLAP 与 MOLAP 两者的特性。HOLAP 可以将部分数据存储在 MOLAP 存储器中，将另一部分数据存储在 ROLAP 存储器中。立方体设计者对此分区的控制程度因产品的不同而不同。技术前沿 2.2 介绍了运用 MicroStrategy BI 工具进行简单分析的案例。

技术前沿 2.2 MicroStrategy 的数据仓库

MicroStrategy 是商务智能、数据仓库管理系统和商业报表解决方案领域的主流独立供应商。近来，此市场中的其他大型供应商纷纷被大型 IT 公司并购重组：Oracle 收购了 Hyperion，IBM 收购了 Congos，SAP 收购了 Business Objects。尽管存在这些并购，但商务智能和数据仓库领域仍然是积极的、充满活力与机遇的。

下面是运用 MicroStrategy 软件分析假设的商业问题。TDUN 网站展示了一个更加全面的案例。设想这样的场景：你（一个全球电信公司的销售副总）将去欧洲出差。在星期一会见地区销售人员之前，你想知道上季度（2004 年第 4 季度）销售代表的业务情况。你可以访问 MicroStrategy 网站创建一个特定报表。为了创建这份报告和其他的 OLAP 报告，你需要 TeradataStudentNetwork.com 网站的登录密码。这个网站的教育用途是免费的，只有你的教授可以为你获取登录密码，从而你可以使用此网站的 MicroStrategy 软件和其他一系列的商务智能资源。

当你登录了 TeradataStudentNetwork.com 网站后，首先登录到“Apply & Do”并在“Software”部分选择“MicroStrategy BI”选项。在“MicroStrategy/BI”页面，按以下步骤操作：

1. 点击“MicroStrategy Application Modules”链接，你将登录到一个页面，此页面显示先前生成的 MicroStrategy 应用程序列表。
2. 选择“Sales Force Analysis Modules”。此模块支持整个销售过程的深度分析。该分析增加了你的领导力，优化了产品线，利用了组织中最成功的销售经验，并提升了销售组织的效率。
3. 在“Sales Force Analysis Modules”页面，你会看到 3 个选项：View、Create 和 Tools。在“View”部分，点击“Shared Reports”链接，你将登录一个具有大量已生成的共享报表的页面。
4. 在“Shared Reports”页面，点击“Pipeline Analysis”文件夹。渠道分析报表分析了销售渠道中所有开放性机会与交易。这些报表可以衡量销售渠道的当前状态，观测其改变趋势和关键事件，并识别关键开放机会。你可以检查每个销售代表的销售渠道以及他们是否完成了上季度的销售指标。
5. 在“Pipeline Analysis”页面，点击名为“Current Pipeline vs. Quota By Sales Region and District”报表。这份报表展示了每个销售地区销售渠道的当前状态。它也能反映此季度的目标配额是否能完成。
6. 在“Current Pipeline vs. Quota By Sales Region and District”页面选择（单击）“2004 Q4”作为报表参数，这表示你想查看销售人员上季度配额的完成情况。
7. 点击页面底部的“Run Report”按钮，运行报表。你将登录一个销售报表页面，此处计算了 3 个欧洲销售区域的所有指标值。在这份互动报表中，通过下拉组合框可以很简单地选择从欧洲到美国或者加拿大的区域，或者你可以点击进入 3 个欧洲区域中的一个，来查看此区域的详细分析。

2.5 节复习题

1. 列举数据仓库的优点。
2. 列举选择数据仓库供应商的标准，并说明它们为什么是重要的。
3. 什么是 OLAP，如何与 OLTP 进行区分？
4. 什么是立方体？什么是钻取/上钻/切片和切块？

5. 什么是 ROLAP、MOLAP 和 HOLAP？与 OLAP 有何不同？

2.6 数据仓库的实施问题

实施数据仓库项目需要投入大量精力，还必须根据已建立的方法来计划和执行。然而，这个项目的生命周期涉及很多方面，没有人能够成为各个领域的专家。这里我们将讨论与数据仓库相关的具体概念和案例。Inmon（2006）文献提供了一系列的方法，数据仓库项目的策划人可以用这些方法来实施项目。

Reeves（2009）和 Solomon（2005）文献给出了与关键问题相关的准则，指出了应当衡量的风险，并提供了可以保证数据仓库成功实施的流程。他们汇总了 11 项可并行执行的重要任务列表：

1. 建立服务级别协合同和数据更新需求
2. 识别数据源和政府政策
3. 数据质量计划
4. 数据模型设计
5. ETL 工具的选择
6. 关系型数据库软件和平台的选择
7. 数据传输
8. 数据转换
9. 协调流程
10. 执行和完成计划
11. 终端用户支持

参考以上准则可增加组织成功的机会。假设一个企业级数据仓库的大小和规模，未预料到这些问题会极大增加项目失败的风险。

Hwang and Xu（2005）文献对数据仓库成功案例进行了深入研究。结果显示数据仓库项目的成功是多方面努力的结果，Hwang 和 Xu 提出，建立一个数据仓库应以改善用户效率为目标。这样做的显著好处是即时的检索信息和信息质量的提升。研究结果也显示项目的成功取决于多方面因素。

人们期望了解他们的 BI 和数据仓库项目与其他公司的项目相比到底有多成功。Ariyachandra and Watson（2006a）文献提出了 BI 和数据仓库项目成功的衡量标准。Watson et al.（1999）文献研究了数据仓库的失败案例。他们的研究成果显示人们对“失败”有不同的定义，这一结果也得到了 Ariyachandra and Watson（2006a）文献的证实。数据仓库协会（tdwi.org）提出了数据仓库成熟度模型，可供企业描述数据仓库的发展历程。该模型提供一种便捷方法来衡量组织机构的数据仓库的实施动力，现在在哪，下一步该怎么做。该成熟度模型由 6 个阶段组成：孕育期、初创期、发育期、成长期、成熟期、衰退期。商业价值随着模型中的各个阶段而增长。这些阶段依据一系列的特性而划分，包括系统范围、分析结构、管理层观念、分析类型、领导力、资金投入、技术平台、变革管理、行政管理。详情请参阅 Eckerson et al.（2009）和 Eckerson（2003）文献。

Saunders（2009）文献提供了一种易于理解的开发数据仓库方法。Weir（2002）专门给出了实施数据仓库解决方案的最佳实践。下面是最明确的实施准则列表：

- 解决方案必须符合企业战略和商业目标。
- 行政人员、管理人员和用户必须全力投入。
- 管理用户对整个解决方案的期望很重要。
- 数据仓库必须逐步建立。
- 项目初期必须考虑适应性和可扩展性。
- 解决方案必须由 IT 和商业人员共同管理（这样一个良性的业务——供应关系才能建立）。

- 只有加载数据与决策分析有关的，已经清理的，来源于已知/可靠的数据源（组织的内部及外部）的数据。
- 不要忽视培训需求（目标用户可能并不精通电脑）。
- 选择与现有设备相符的可靠工具和方法。
- 注意组织人员、政策和地盘之争。

数据仓库解决方案存在众多风险。其中大部分风险也存在于其他IT项目中，但数据仓库项目的风险更严重，因为数据仓库项目的成本高、需要时间和资源、规模巨大。在项目启动时就应该评估各种风险。实施一个成功的数据仓库项目，应当谨慎衡量各种风险和避免以下问题：

- **错误的项目发起** 你需要一个拥有所需资源的执行赞助商以支持和赞助数据仓库项目。你也需要一个执行项目驾驭者，他能赢得其他执行人员的尊重，对技术抱有良性的质疑态度，果断并且灵活。同时还需要一个IS/IT经理来主持项目。
- **制定不可能完成的目标** 你不想在关键时刻让管理层失望。每个数据仓库项目都有两个阶段：第一阶段是销售阶段，这个阶段是向需要访问资源的人销售能带来的利益，说服他们实施你的数据仓库计划。第二阶段是努力实现第一阶段中许诺的目标。比如仅仅是从1～700万的利益，你很有希望实现这个目标。
- **从事与政治不相关行为** 不要宣传数据仓库能帮助管理人员更好地制定决策，这样说容易暗示你认为目前他们的决策做得不好。应该告诉他们将从数据仓库中获得有用的信息来帮助决策制定。
- **将能利用的数据加载到数据仓库中** 不要让数据仓库成为一个数据垃圾堆。这将导致系统运行速度缓慢。实时计算和分析逐渐成为趋势。实时加载数据时，数据仓库必须关闭。
- **相信数据仓库设计与传统数据库设计相同** 一般来说，不是这样的。数据仓库的目标是访问全部记录，而传统数据库访问一个或一些记录。存储内容也不同，在数据组织方式上差异尤为明显。DBMS趋向于非冗余的、标准化的和关系型，而数据仓库是冗余的、非标准化的和多维度的。
- **选择一个面向技术而非面向用户的数据仓库管理员** 成功实施数据仓库的关键之一在于理解用户的需求，而不是为了技术追逐先进的技术。
- **专注传统的内部关系型数据，忽视外部数据、文档、图片甚至音频和视频的价值** 数据有很多种格式，同时必须在正确的时间，以正确的格式提供给正确的人员。它们必须被合理分类。
- **用重复而且冲突的数据定义交付数据** 数据清理是数据仓库中的关键问题。它包括协调冲突的数据定义和对组织内的数据进行格式化。政策上来说，这可能很困难，因为它通常需要在行政级别上改变。
- **相信性能、能力和可扩展性的承诺** 通常，数据仓库需要比开始预算时具有更强的性能和速度。项目初期的计划需要升级。
- **相信一旦数据仓库建立并运行起来，你的问题就多了** DSS/BI项目倾向于持续性展开。每次部署都是对原型过程的迭代。这将一直需要给数据仓库增加更多不同的数据，同时也要给现有和新增的决策制定者提供附加分析工具。高能力和每年预算必须完成，因为成功将带来成功。数据仓库项目是一个持续性的过程。
- **专注于特殊数据挖掘、定期生成报表而不是预警** 数据仓库中信息的发展过程如下：（1）从旧系统中提取数据，清理数据并添加到数据仓库中；（2）理解用户需求才能支持特殊报表；（3）将特殊报表转换成定期计划报表。理解并满足客户需求看起来很容易，但实际情况却不乐观。管理者业务繁忙，又需要花费时间来阅读报表。预警系统比某一时期的报表系统更好，它使数据仓库任务成为关键。预警系统监测数据流入数据仓库的

过程，一旦关键事件发生就产生预警，通知相应的关键人员。

在大多数组织中，只有在高级管理层对项目开发的强烈支持下并且项目经理在组织结构中拥有较高职位时，数据仓库项目才能成功。上述因素对大规模IT项目可能是正确的，但对数据仓库的实现更加重要。数据仓库的成功实施建立了一个支持组织决策分析的结构性框架，在某些情况下也提供可以访问组织客户和供应商信息的综合性SCM。网络数据仓库的实施（有时又称为网络仓库）使得访问大量数据更加便捷，但是却很难衡量数据仓库的硬效益。硬效益的定义是可表示为货币形式的组织效益。很多公司的IT资源是有限的，必须优先安排一些项目。管理层的支持和优秀的项目经理可以保障数据仓库项目拥有成功实施所必需的资源。数据仓库资源需付出巨额成本，在某些情况下，还需要高端处理器和大量可直接访问的存储设备。网络数据仓库还有特殊安全要求来确保仅授权用户能访问数据。

用户参与到数据和访问建模开发中，这是数据仓库开发的关键成功因素。在数据建模中，需要专门知识来判断所需要的数据，定义与数据有关的业务规则，并决定需要哪种聚类和计算。访问建模可用来决定如何从数据仓库中提取数据，通过确定哪些数据需要索引来协助定义仓库的物理定义。它也指出是否需要依赖数据集市来促进信息检索。开发和实施数据仓库需要一系列团队技能。这些技能包括数据仓库技术和开发工具的深层知识。如上所述，源系统和开发技术涉及大量成本投入和开发进程，它们可被用来加载和维护数据仓库。

应用案例2.6展示了保险行业大规模实施集成数据仓库的典型案例。

应用案例2.6 一个大型保险公司运用AXIS集成企业数据

美国的一个大型保险公司开发了一个集成数据管理和报表系统，用来提供企业业绩和风险的统一视图，该系统在大量的业务单元计划和管理活动中成为新的战略性角色。

XYZ保险公司（不显示实名）和它的附属公司组成了世界上最大的金融服务组织。一个世纪前，XYZ保险公司已经成长并将实施业务多元化，成为了家庭财产保险、意外保险、人寿保险、退休保险、资产管理和战略投资服务领域的主流供应商。如今，该公司成为了一个工业巨头，拥有超过1 500亿美元的法定资产、年收入超过150亿美元、员工超过20 000名并拥有超过1 000家公司在其保护伞下运行。

问题

对于它的大部分业务，投保该公司的家庭和当分散并且独立。随着时间的迁移，企业逐渐壮大，这种分散管理方式使得公司业绩报表和决策制定截然不同。由于对企业业绩看法不一致，公司报表是无远见的、分散的、缺乏的、并且通常是不准确的。获取、巩固、清除和验证基础财务信息的重担使得企业不能运用有效分析和深刻见解来支持管理活动。

为了解决集成的迫切需要，XYZ保险公司于2004年1月发起了需求分析倡议，产生了统一数据管理系统。此集成系统称做AXIS。预计它能够提供实时、准确和有效的顶尖水平的报表工具和分析服务，可以支持企业规计划、资金管理、风险评估和管理决策制定。

解决方案

XYZ保险公司决定运用最佳组合方法开发AXIS。与从单一供应商购买所有组件不同，它将选择最适合各模块分析需求的组件。下列工具/供应商可供选择：

- 数据仓库：AXIS系统具有集中星形拓扑架构，Teradata数据仓库处于中心位置。
- 提取、传输、集成和元数据管理：Information Powercenter处理从源系统到AXIS环境（和AXIS的中间系统）的所有数据传输。

- 报表和分析：AXIS 系统中所有可视化的报表和分析功能是由一整套的 Hyperion 工具提供的，其中包括 Essbase、Planning、Reporter、Analyzer 和 Intelligence。
- 元数据管理：运用 Kalido 主数据管理系统（Master Data Management，MDM）开发和维护引用数据的层次结构、维度和界面翻译和转换的业务规则。

结果

即使一个拥有 XYZ 保险业资源的企业，实施 AXIS 系统也绝非易事。它需要生成超过 200 个业务源系统界面。顶峰时，开发团队雇佣了 280 名员工（60% 来源于内部 IT 和业务部门，40% 来源于外部承包商），他们对项目投入了 600 000 工时。具备完整功能的系统在 2006 年 4 月发布。

拥有技术和支持流程的标准化企业信息资产，XYZ 保险公司能够加强大部分劳动密集型业务和报表生成活动。这释放了人力资本和企业资源，用于更具战略意义和更高价值的企业活动。另一个益处是企业的业务单元拥有了一致和准确的业务信息可供决策者参考。AXIS 系统最重要的好处可能是它将 XYZ 保险公司变成了敏捷企业。因为企业管理人员可以及时地访问企业级数据，业务单元可以准确并迅速地应对变化的情况（解决问题并利用机会）。

来源：Based on Teradata, "A Large US-based Insurance Company Masters Its Finance Data," Teradata Industry Solution, teradata.com/t/WorkArea/DownloadAsset.aspx?id=4858(accessed July 2009).

大型数据仓库和可扩展性

除了动态性之外，数据仓库需要支持可扩展性。与扩展性相关的最主要问题有数据仓库的数据量、数据仓库预计增长的速度、并发用户的数量、用户查询的复杂度。数据仓库必须可以水平和垂直地扩展。由于数据量增长和支持新业务功能的需求，数据仓库需要扩展。数据量增长也许是当前周期的数据（例如当月的数据）或者历史数据增加的结果。

Hicks（2001）文献描述了大型数据库和数据仓库。Wal-Mart 不断增加它的大型数据仓库的规模。Wal-Mart 被认为可以运用数百太字节（TB）的数据仓库来研究销售趋势，追踪库存和其他任务。IBM 最近公布了其 50TB 数据仓库基准（IBM，2009）。美国国防部门正在使用一个 5PB（Peta byte）数据仓库和存储库以存储 900 万军事人员的医学记录。因为需要存储新闻素材，所以 CNN 也有一个规模达 PB 级的数据仓库。

如果一个数据仓库的大小呈指数级增长，那么它的可扩展性将成为重要问题。高可扩展性意味着查询和其他数据访问功能，将随着数据仓库大小呈线性增长（理想化的）。请查看 Rosenberg（2006）文献关于提高查询性能的方法。实际上，人们已开发出专门方法来创建可扩展数据仓库。当管理数百太字节 TB 或更多的数据时，可扩展性将很难实现。TB 级数据具有相当大的惯性，占用大量物理空间，同时需要功能强大的计算机。有些公司使用并行处理器，另一些公司运用灵活的索引和搜索来管理数据。有些公司在不同物理数据存储之间传输数据。当越来越多的数据仓库达到 PB 级别时，将会继续研制出越来越好的可扩展性解决方案。

Hall（2002）文献也可解决可扩展性问题。AT&T 是大型数据仓库部署和应用领域的行业领导者。运用其 26TB 级的数据仓库，AT&T 能检测电话卡冒用，调查有关绑架和其他罪行的电话。它也能计算电视观众选择下一个美国偶像的百万个电话投票。

列举一个数据仓库成功实施的案例，见 Edwards（2003）文献。Jukic and Lang（2004）文献调查了数据仓库的发展趋势，并指出了数据仓库和 BI 应用的开发和支持与离岸外包资源使用的特殊问题。Davison（2003）指出了 IT 离岸外包一直以每年 20% 到 25% 的速度增长。当考虑离岸外包数据仓库项目时，必须认真考虑文化和安全因素，请参阅（Jukic and Lang，2004）。

2.6 节复习题

1. DW 实施过程中可并行执行的主要任务是什么？

2. 列举并讨论反映稳健的 DW 实施准则。
3. 当开发一个成功的数据仓库时，需要考虑和避免的最重要的风险和问题是什么？
4. 什么是可扩展性？它在 DW 中是如何应用的？

2.7 实时数据仓库

传统的数据仓库和 BI 工具专注于辅助管理者制定战略和战术决策。增加的数据量和加快的更新速度，从根本上改变了数据仓库在现代企业中的角色。对于许多企业来说，制定快速和一致的企业决策不仅仅需要一个传统数据仓库或者数据集市。传统数据仓库不再是商业的关键。数据一般每周更新一次，这不能应对近实时的业务。

越来越多的数据快速进入数据仓库，并要求立即转换成决策，这意味着组织需要实时数据仓库。因为决策支持已成为操作性的，集成 BI 需要闭环分析，之前的 ODS 将不再支持现在的需求。

2003 年，实时数据仓库诞生，并将这些技术用来支持运营决策。**实时数据仓库**（Real-time Data Warehousing，RDW），也称为**动态数据仓库**（Active DataWare housing，ADW），是通过数据仓库加载和提供数据的过程。它是从 EDW 概念演变而来的。RDW/ADW 的动态特征补充和扩展了传统数据仓库，实现了战术决策功能。企业中直接与客户和供应商接触的员工有权很容易地制定基于信息的决策。当 ADW 直接给客户和供应商提供信息时，甚至能产生更大的效益。获取决策制定所需的信息能积极促进大多数客户服务、SCM、物流及其他服务。电子商务已成为动态数据仓库需求（Armstrong 2000）的主要催化剂。例如，网上零售商 Overstock. com 公司（overstock. com）将数据用户连接到实时数据仓库。在 Egg plc，世界最大的网上银行，客户数据仓库（Customer Data Warehouse，CDW）进行近实时更新（详见应用案例 2.7）。

因为业务需要发展，所以数据仓库的需求也在发展。基层的数据仓库在基本层面简单地报告发生的事件。在下一层面，数据仓库进行一些分析。随着系统的发展，它能提供预测功能，这将导致下一层面的运作。发展到最高层面，ADW 能够让事件主动发生（例如，创建销售和营销活动、识别和利用机会）。请看图 2-10 对该演变过程的图形描述。Wiesobel（2009）文献介绍了一项对管理数据仓库演变的最新研究。

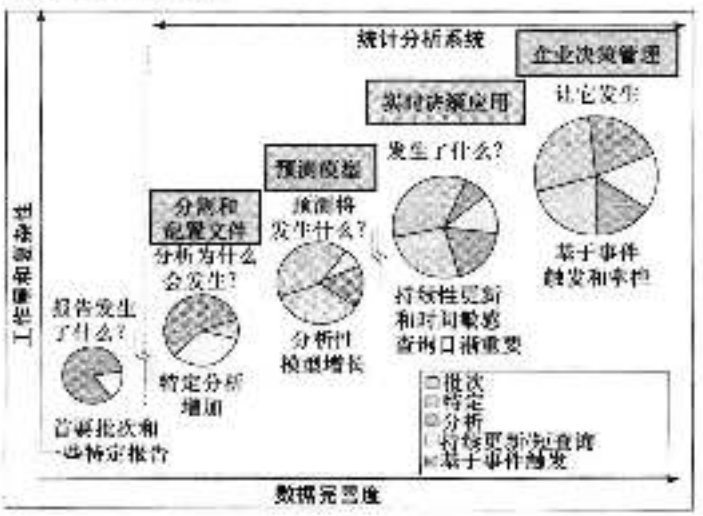


图 2-10 企业决策演变

来源：Courtesy of Teradata Corporation

Teradata 公司提供了支持 EDW 的基本要求。它也使动态数据仓库具有提交数据更新、性能、有效性和支持企业决策管理的新特征（见图 2-11 中的例子）。

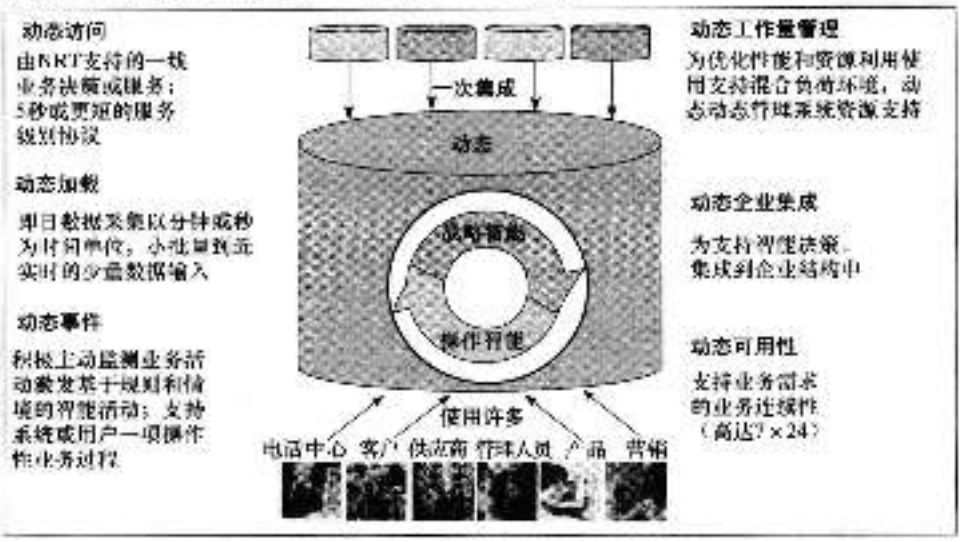


图 2-11 Teradata 动态 EDW

来源：Courtesy of Teradata Corporation

ADW 提供了综合信息库来建设企业战略和战术决策支持。运用实时数据仓库，一旦业务发生，数据就从 OLTP 系统中集合并立即被移入数据仓库中，而不是以夜间模式从 OLTP 系统中提取操作数据到 ODS 中。这将支持数据仓库的实施更新同时取消了 ODS。基于这个特点，运用实时和历史数据进行战术和战略查询成为可能。

根据 Basu（2003），传统数据仓库和实时数据仓库之间最显著的区别是数据获取模式的转变。下面是一些具有实时数据需求的业务案例和企业需求：

- 一个企业不能用一整天的时间等待它的业务数据加载到数据仓库以供分析。
- 目前，数据仓库可以获取组织固定状况的快照，取代了说明每次更新和类似模式的增量实时数据。
- 在传统的星形拓扑架构中，很难保持元数据的同步性。开发、维护和确保许多系统安全性的成本很高，相反开发、维护和确保一个大型数据仓库安全性的成本却较低，因为数据被 BI/BA 工具集中了。
- 在大量夜间批处理的情况下，系统需要 ETL，对大量夜间数据仓库的处理能力要求很高，处理过程也许会占用很长时间。一个实时采集数据的 EAI 可以减少或取消夜间批处理。

尽管 RDW 具有优势，但是开发 RDW 会产生一系列问题。这些问题与架构、数据模型、物理数据库设计、存储和可扩展性、可维护性相关。另外，访问数据的精确时间甚至减少到微秒，系统也可能会提取和产生不同版本的事实信息，这将使团队成员感到迷惑。详情请查阅 Basu（2003）和 Terr（2004）的文献。

实时解决方案给 BI 提出了一系列挑战。尽管实时数据仓库项目并不适用于所有解决方案，但是如果企业运用合理方法来处理项目风险，采用合理计划方案并专注于质量安全工作，那么实时数据仓库项目有可能成功。在实施带有 BI/BA 方法的复杂数据仓库时，了解常见困难和采用最佳实践能降低项目问题的严重程度。Burdett and Singh（2004）和 Wilk（2003）文献讨论了详细情况并介绍了真实的实施案例，也可以查阅 Akbay（2006）和 Ericson（2006）文献。

请看技术前沿2.3中对实时概念演变的详细介绍。大陆航空公司的航线管理仪表盘应用应用案例（请看本章末应用案例）展示了实时BI在面对面客户交互情况下访问数据仓库的能力。柴作人员运用实时数据仓库来识别大陆航线网络中的问题。另一个例子，UPS投资了6亿美元来运用实时数据和程序。（Malykhina，2003）文献指出，通过管理实时包流技术，预计该投资每年将减少1亿英里运输成本并节约1 400万加仑燃油。表2-6对传统数据仓库和实时数据仓库环境进行了比较。

表2-6 传统数据仓库环境和动态数据仓库环境的对比

传统数据仓库环境	动态数据仓库环境
只有战略决策	战略和战术决策
结果与时间以及量	不操作及量结果
只存储每日、每周、每月数据，通常数据汇总是适当的	仅存储几分钟之内的全面详细数据
中等程度的用户并发	大量（1 000或更多）用户同时访问和查询该系统
使用高度限制性的报告或检查来确认现有的流程和模式；经常使用已形成的汇总表和数据集市	使用灵活特定的报表和计算机辅助业报（例如数据挖掘）发现新假设和关系
高级用户、知识工作者、内部用户	业务人员、呼叫中心、外部客户

来源：Based on P. Coffee, "'Active' Warehousing," *e Week*, Vol. 20, No. 25, June 23, 2003, p. 36; and Teradata Corp., "Active Data Warehousing," teradata.com/t/page/87127/index.html (accessed April 2006).

应用案例2.7 Egg PLc点燃了近实时领域的竞争

Egg Plc（egg.com）是世界上最大的网上银行。它通过因特网网站向超过360万的客户提供金融、保险、投资、抵押贷款业务。1998年，Egg选择了Sun Microsystems创建了一个可靠、可扩展、安全的基础设施平台来处理超过250万的日常交易。2001年，为消除延迟问题，该系统升级了。这个新的CDW使用Sun、Oracle和SAS软件产品。源数据仓库拥有大约10TB的数据，使用16-CPU服务器。该系统支持近实时数据访问。它向内部用户提供数据仓库和数据挖掘服务，并且向客户提供所需的客户数据集。运用实时数据构造数以百计的销售和营销活动（在几分钟内）。最大的好处是此系统能更快制定出关于特殊客户和客户群的决策。

来源：Compiled from "Egg's Customer Data Warehouse Hits the Mark," *DM Review*, Vol. 15, No. 10, October 2005, pp. 24-28; Sun Microsystems, "Egg Banks on Sun to Hit the Mark with Customers," September 19, 2005, sun.com/emi/Press/sunflash/2005-09/sunflash.20050919.html (accessed April 2006); and ZD Net UK, "Sun Case Study: Egg's Customer Data Warehouse," whitepapers.zdnet.co.uk/0,39025945,60159401p-39000449q,00.htm (accessed June 2009).

技术前沿2.3 动态数据仓库的实时实现

2003年数据仓库在实践中的作用迅速增长。实时系统虽然新奇，但最近的负面传言是其向需求者即时提供混乱的数据和信息。许多专家，包括《eWeek》的技术主编Peter Coffee，都认为实时系统必须提供实时决策制定过程。NCR公司数据仓储事业部的CTO Stephen Brobst，认为实时数据仓库应用是企业运用数据的演进过程。动态性意味着数据仓库也可以被用作业务和战术工具。Brobst提出了5阶段模型，与Coffee提出的企业在数据应用领域是如何增长的（Brobstet et al.，2005）相一致。这些阶段（他们打算回答的问题）分为：报告（什么发生了）、分析（为什么会发生）、预测（什么将发生）、运作（什么正在发生），和动态数据仓库（我希望发生什么）。最后一个阶段，动态数据仓库是企业能获取最大利益的地方。许多组织正在增强中央数据仓库的功能来服务于运作和战略决策

支持。

来源：Based on P. Coffee, "'Active' Warehousing," eWeek, Vol.20, No.25, June 23, 2003, p.36; and Teradata Corp. "Active Data Warehousing," teradata.com/t/page/87127/index.html (accessed April 2006).

实时数据仓库、近实时数据仓库、意近数据仓库和动态数据仓库是实践中用来描述同一概念的不同名称。Gonzales（2005）文献描述了对 ADW 的不同定义。根据 Gonzales 的定义，ADW 是可以提供所需混合战术和战略数据的唯一选择。ADW 的构建架构与 Bill Inmon 开发的企业信息工厂架构非常相似。两者之间的唯一区别是在单一环境中实施了数据存储。然而，一个基于 XML 的 SOA 和 Web 服务为所需混合战术和战略数据提供了另一个选择。

实时数据仓库的一个关键问题是不能持续更新所有数据。这在实时生成报表时肯定会产生问题，因为一个人的结果与另一个人的不符。例如，公司运用网络智能业务对象发现了实时智能的一个重要问题。（Peterson，2003）文献指出，不同时间内生成的实时报表内容各不相同。而且，也许没有必要持续更新某种数据（例如，3 年或者 3 年以前的课程成绩）。

实时需求改变了我们对数据库的设计、数据仓库、OLAP 和数据挖掘工具的传统观点。因为动态查询时，它们需要同时逐字更新。但是动态查询的重大商业价值已得到证实，所以企业在业务流程中采用动态查询是至关重要的。认真规划是实施的关键。

2.7 节复习题

1. 什么是 RDW？
2. 列举 RDW 的好处。
3. 传统数据仓库和实时数据仓库间的重要区别是什么？
4. 列举 RDW 的驱动力。

2.8 数据仓库管理系统、安全问题和未来发展趋势

企业可以有效创建和使用数据仓库，从而使其具有明显的竞争优势。由于规模巨大和内在的特性，数据仓库需要特别强大的监测管理来保持其令人满意的效率和生产力。成功的数据仓库管理员需要拥有比传统数据库管理员（Database Administrator，DBA）更多的技巧和经验。**数据仓库管理员**（Data Warehouse Administrator，DWA）应该熟悉高性能软件、硬件和网络技术，还需要具备坚实的业务洞察力。由于数据仓库提供了 BI 系统和 DSS 来帮助管理者进行决策制定活动，所以为了合理设计和维护数据仓库结构，DWA 应该熟悉决策制定过程。对于 DWA 来说，在使数据仓库具备应对快速改进的灵活性的同时，将现行需求和数据仓库的能力维持在一个稳定水平是特别重要的。最后，DWA 必须具备卓越的沟通能力。请查看 Bernander et al.（2000）文献中对 DBA 和 DWA 关键区别的描述。

信息的安全性和保密性是数据仓库领域主要和密切关注的问题。美国政府已经通过了法规（例如，Gramm-Leach-Bliley 隐私和保护措施、1996 年的健康保险可移植性和问责性法案[HIPAA]），对客户信息管理实行强制性要求。因此，为遵守众多隐私条例，企业必须创建有效但灵活的安全程序。根据 Elson and LeClerc（2005），数据仓库安全性应关注 4 个主要领域：

1. 建立有效企业和安全性政策和程序。有效安全政策应从顶层开始，贯彻执行管理，并应传达到组织内的每个人。
2. 通过实施逻辑安全性程序和技术来限制访问。这包括认证、访问控制和加密技术。
3. 限制对数据中心环境的物理访问。
4. 建立安全性和保密性的有效内部控制审核程序。

请查阅技术前沿 2.4 中对 Ambeo 重要软件工具的介绍，该工具可监测数据仓库安全和隐私。最后，记住应谨慎地通过移动设备访问数据仓库。在这种情况下，数据仅能被只读访问。

短期内，数据仓库开发决定于最显著因素（例如，数据存储量、对延迟的容忍、数据类型

的多样性和复杂性）和次显著因素（例如，未满足终端用户对仪表盘的需求、平衡记分卡、元数据管理、信息质量）。基于这些因素，Moseley（2009）和 Agosta（2006）文献认为数据仓库的发展趋势将倾向于简单性、实用性和性能。

技术前沿 2.4 Ambeo 提供成熟的数据访问审计解决方案

从 1997 年开始，Ambeo（ambeo.com；现在的 Embarcadero 技术有限公司）已部署了可以提供以下服务的技术：性能管理、数据使用情况跟踪、数据隐私审查和对财富 1 000 强公司的监测。这些公司具备一些大型数据库环境。Ambeo 数据访问审计解决方案在企业信息安全基础设施中起着主要作用。

Ambeo 技术是相对简单的解决方案，它记录了数据库的一切交易，并且成本较低或无需成本。此外，它提供数据访问审计来准确识别谁在查看数据，他们什么时候查看的，他们用数据做了什么。这种实时监测有助于快速和有效识别危安全漏洞。

来源：Based on "Ambeo Delivers Proven Data Access Auditing Solution," *Database Trends and Applications*, Vol. 19, No. 7, July 2005; and Ambeo, "Keeping Data Private (and Knowing It): Moving Beyond Conventional Safeguards to Ensure Data Privacy," am-beo.com/why_ambeo_white_papers.html (accessed May 2009).

数据仓库的未来发展趋势

数据仓库领域正在/已经成为近几十年来信息技术中的活跃领域，同时 BI 也证实了该领域的重要性将会日渐增加。下面是近来流行的、将在定义未来数据仓库中起着重要作用的概念和技术。

- **来源**（从不同和分散的来源获取数据）
- **开源软件** 在数据仓库、商业智能和数据集成领域中，开源软件工具的运用正以空前的水平增长。数据仓库领域开源软件的增长基于很好的理由（Russom，2009）：（1）经济衰退带动了人们对低成本开源软件的兴趣；（2）开源工具正在进入新的成熟阶段；（3）开源软件发展了传统企业软件而不是代替它们。
- **软件即服务（Software as a Service，Saas），"扩展的 ASP 模型"** SaaS 是部署信息系统应用的一种创造性方法，应用提供商向用户提供面向需求的服务许可应用程序（通常在因特网上）。SaaS 软件供应商使用自己的服务托管应用程序或者上传应用程序至客户端。本质上，SaaS 是 ASP 模型的全新和改进版本。数据仓库用户很难发现能满足特殊需求的基于 SaaS 应用程序和资源。随着这些软件的提供变得越来越便捷，作为数据仓库的应用平台，SaaS 的吸引力和实际使用也会增加。
- **云计算** 云计算也许是近年来最新和最具创新性的平台，其中汇聚和虚拟化了大量硬件和软件资源，因此当需要时，它们可以被自由分配给应用和软件平台。随着工作量的增加，信息系统应用程序也按比例增加。虽然云计算和类似的虚拟技术程序是为业务应用程序而建立的，但数据仓库应用平台也开始运用这些技术。当数据仓库中的数据量变化不可预测，决策规划能力变得困难时，云计算中的动态分配可发挥重大作用。
- **数据仓库应用程序** 近年来，数据仓库领域中讨论最广泛的问题之一就是数据仓库应用程序。该问题的初始定义与一个整体性解决方案有关，该解决方案仅给数据仓库提供全面的技术堆栈（软件、硬件等）。从那时起，它的定义就被修改了，它开始提供局部技术匹配服务以满足用户的特殊需求。未来发展趋势是基于最佳组合管理。
- **基础设施**（系统架构-软件和硬件-系统升级）
- **实时数据仓库** RDW 意味着现行数据仓库的数据更新周期更加频繁（几乎同步于操作型数据库中的数据更新）。实时数据仓库系统能达到近实时数据更新，其数据延迟通常从几分钟到几小时不等。随着延迟的减少，数据更新的成本看似是成倍增长。未来的许多技术进步（从自动数据获取到智能软件代理）使得实时数据仓库的价格可以负担得起。
- **数据管理技术和实践** 下一代数据仓库平台最迫切的需求包括技术和实践，一般我们并

不认为这是平台的一部分。特别地，许多用户需要更新数据管理工具，这些工具可以通过数据仓库处理数据。未来主数据管理技术将迅速发展。这一全新且极其重要的概念正在日渐流行，原因如下：（1）与业务系统更紧密的集成需要 MDM；（2）大多数数据仓库仍缺少 MDM 和数据质量功能；（3）常规和财务报表必须完全清晰和准确。

- **在内存中处理（64 位计算能力）或"超级计算能力"** 64 位系统通常可以提供比旧系统更快的 CPU 和更省电的硬件设施。但是，对于数据仓库来说，64 位系统最显著的好处是它具有可寻址存储器的超大空间，可以部署内存数据库以支持报表或分析应用程序的运行，它们需要快速查询的反应速度。内存数据库可提供这样的速度，因为它们没有磁盘输入/输出。内存数据库通常是 DBMS 的功能，但有些 BI 平台也提供内存数据存储和相关处理服务。

ETL 工具通常支持 64 位环境的内存处理，因此可以在大存储空间下执行复杂连接和转换，不需要将数据下载到磁盘的临时表中。这使得 ETL 数据流入正确的管道，这意味着 ETL 工具可以升级，在较短时间内处理大量数据。

- **新 DBMS** 数据仓库平台由多种基本组件组成，其中最关键部分是数据库管理系统（Database Management System，DBMS）。这是理所当然的；事实上，DBMS 是数据仓库平台中的组件，该平台需要执行大量工作来建立数据模型和优化查询性能。因此，人们必然期望对新一代的 DBMS 进行创新。

高级分析 当用户舍弃基于 OLAP 的基本方法并开始高级分析时，数据仓库将给用户提供不同的分析方法。有些用户选择基于数据挖掘、预测分析、统计、人工智能等的高级分析方法。同时，大多数用户将选择基于 SQL 的方法。是否基于 SQL，高级分析方法看来都将会是下一代数据仓库的发展趋势。

数据仓库的未来开始来充满希望和挑战。当世界商业开始全球化和复杂化时，对商业智能和数据仓库工具的需求也日渐突出。快速发展的信息技术工具和技术正朝着可以满足商业智能系统的未来需求的正确方向发展。

2.8 节复习题

1. 为确保数据仓库中客户数据的安全性和保密性，企业可以采取什么措施？
2. DWA 应具备什么性能？为什么？
3. 可以创建未来数据仓库的最新技术是什么？为什么？

2.9 相关资源、链接和 Teradata 大学网络的连接

使用下列资源，加深对本章和其他各章的理解。

2.9.1 资源和链接

我们推荐你进一步阅读和查看下列资源和链接：

- 数据仓库协会（tdwi.com）
- 《DM Review》（dereview.com）
- DSS 资源（dssresources.com）

2.9.2 案例

所有的大型 MSS 供应商（例如 MicroStrategy、Microsoft、Oracle、IBM、Hyperion、Cognos、Essys、Fair Isaac、SAP 和 Information Builders）均提供有趣的客户成功案例。学术导向案例可以在以下网站找到：哈佛商学院案例集（harvardbusinessonline.hbsp.harvard.edu），商业绩效供

高资源中心（bpir. com），思想集团出版社（idea- group. com），常春藤联合出版社（ivylp. com），管理研究 ICFAI 中心（icmr. icfai. org/casestudies/icmr_case_studies. htm），知识风暴（knowledgestorm. com）和其他网站。寻找更多的案例资源，请查看 Teradata 校园网（teradatauniversitynetwork. com）："大陆航空公司应用实时商业智能腾飞"，"北部卡罗来纳州蓝十字和蓝盾的数据仓库治理"，"运用全球数据仓库，3M 转向以客户为中心"，"数据仓库支持公司战略"，"Harrah 从客户信息中获得高回报" 和 "旋涡"。同时也推荐数据仓库失败案例集，由 8 个简短的数据仓库失败案例组成。

2.9.3 供应商、产品和演示

《DM Review》（dereview. com）上刊登了供应商、产品和样品程序的完整列表。表 2-2 列出了供应商信息。也可登录 techonologyevaluation. com 查看相关信息。

2.9.4 期刊

本文推荐下列期刊：

- Baseline（baselinemag. com）
- Business Intelligence Journal（商业智能杂志）（tdwi. org）
- CIO（cio. com）
- CIO insight（cioinsight. com）
- Computerworld（计算机世界）（computerworld. com）
- Decision Support Systems（决策支持系统）（elsevier. com）
- DM Review（dereview. com）
- eWeek（eWeek. com）
- InfoWeek（infoweek. com）
- InfoWorld（infoworld. com）
- InernetWeek（internetweek. com）
- Management Information Systems Quarterly（管理信息系统季刊）（MIS Quarterly，misq. org）
- Technology Evaluation（技术评估）（technologyevalution. com）
- Teradata Magazine（teradata. com）

2.9.5 其他参考文献

关于数据仓库更多信息，请查看下列内容：

- C. Imhoff，N. Galemmo，and J. G. Geiger.（2003）. *Mastering Data Warehouse Design：Relational and Dimensional Techniques*. New York：Wiley.
- D. Marco and M. Jennings.（2004）. *Universal Meta Data Models*. New York：Wiley.
- J. Wang.（2005）. *Encyclopedia of Data Warehousing and Mining*. Hershey，PA：Idea Group Publishing.

更多关于数据库和数据仓库开发架构的信息，请查看下列内容：

- R. T. Watson.（2006）. *Data Management*，5th ed.，New York：Wiley.

2.9.6 Teradata 大学网络的连接

TUN（teradatauniversitynetwork. com）提供了大量数据仓库信息和案例。最佳案例之一是大陆航空公司案例。本章中前面部分提到了其他的推荐案例。在 TUN 中，如果你点击了"课程"（Courses）标签并选择"数据仓库"（Data Warehousing），你将看到许多相关文章、作业、著作

专章、课程网站、PPT、项目、研究报告、教学大纲和网络课程的链接。你也将看到动态数据仓库软件产品展示的链接。

最后，你将看到 Teradata 的链接（teradata.com），在这里你会发现其他信息，包括优秀的数据仓库成功案例、白皮书、网络课程和 Teradata 杂志的网络版本。

本章重点

- 数据仓库是为组织数据而专门构建的数据存储库，终端用户可以很容易地使用应用程序访问数据。
- 数据集市中的数据主题唯一（例如营销）。数据集市是数据仓库中数据子集的复制。数据集市是成本相对较低的解决方案，可以被数据仓库替代或者作为数据仓库的补充。数据集市可以依赖或独立于数据仓库。
- ODS 是一种客户信息档案数据库，该数据库通常被用做数据仓库架构的中间层。
- 数据集成包含 3 个主要进程：数据访问、数据联合和变化捕获。3 个进程正确执行后，就可以访问数据，还可以访问 ETL 数组、分析工具和数据仓库环境。
- ETL 技术从众多来源中提取数据，清理数据并加载数据到数据仓库中。ETL 是数据中心项目的集成过程。
- 通过实时加载数据和向用户提供数据，支持动态决策制定，实时或动态数据仓库补充并扩展了传统数据仓库，进入了制定业务和战术决策领域。
- 数据和信息的安全性和隐私性是数据仓库领域的关键问题。

关键术语

Active Data Warehousing（ADW，动态数据仓库）
Decision Support Systems（DSS，决策支持系统）
Extraction，Transformation，and Load（ETL，提取、转换和加载）
parallel processing 并行处理
ad hoc query 特定查询
dependent data mart 非独立的数据集市
grain 粒度
prototyping 原型法
best practices 最佳实践
dimensional modeling 维度建模
Graphical User Interface（GUI，图形用户界面）
Real-time Data Warehousing（RDW，实时数据仓库）
cloud computing 云计算
dimension tables 维度表
independent data mart 独立数据集市
relational database 关系型数据库
cube 立方体
drill down 钻取
metadata 元数据
Relational Online Analytical Processing（ROLAP，关系型在线分析处理）
data cube 数据立方体
Enterprise Application Integration（EAI，企业应用集成）
multidimensional analysis 多维分析
risk 风险
data integration 数据集成
Enterprise Data Warehouse（EDW，企业数据仓库）
multidimensional database 多维数据库
scenario 场景
data mart 数据集市
enterprise decision management 企业决策管理
multidimensional OLAP（MOLAP，多维在线分析处理）
software agent 软件代理
data quality 数据质量
Enterprise Information Integration（EII，企业信息集成）
oper marts 操作集市
speech recognition 语音识别
Data Warehouse（DW，数据仓库）
expert 专家
Operational Data Store（ODS，业务数据存储）
SQL 结构化查询语言
Data Warehouse Administrator（DWA，数据仓库管理员）
extraction 提取
snowflake schema 雪花形架构
Database Management System（DBMS，数据库管理系统）
star schema 星形架构

讨论题

1. 比较数据集成和 ETL。它们之间有什么样的关系？
2. 什么是数据仓库？它的好处是什么？为什么网络可访问性对数据仓库很重要？
3. 数据集市可以代替数据仓库或补充它。比较并讨论这两种选择。
4. 讨论数据仓库给终端用户带来的好处。
5. 列举数据库管理员和数据仓库管理员之间的区别和相似之处。
6. 阐述数据集成是如何提升数据质量的。
7. 比较 Kimball 和 Inmon 数据仓库开发方法。说明它们分别在何时使用最有效。
8. 讨论创建数据仓库时的安全问题。
9. 研究当前高毕外包的数据仓库的开发实施。撰写关于此问题的报告。课堂上讨论此问题的利润、成本和社会因素。

练习

Teradata 大学和其他动手练习

1. 描述日本可口可乐公司数据仓库项目的开发和应用情况（应用案例 2.4 中的总结）。可以在 DSS 资源网站上找到此案例，dssresources.com/cases/cocacolajapan/index.html。阅读此案例并回答进一步分析和讨论的 9 个问题。
2. 阅读 Ball（2005）文献并对其中介绍的标准进行排序（最好对一个真实企业而言）。撰写报告阐述每条标准的重要性并说明原因。
3. 在考虑数据仓库项目开发时，说明什么时候需要实施二层架构或者三层架构。
4. 阅读 teradatastudentnetwork.com 网站的大陆航空公司案例（在本章案例应用的最后部分做了总结）并回答问题。
5. 登录 teradatastudentnetwork.com 网站，阅读案例“Harrah's High Payoff from Customer Information（Harrah 从客户信息中获取的高回报）”并回答问题。将 Harrah 的结果与航空公司和赌场如何运用客户数据联系起来。
6. 登录 teradatastudentnetwork.com 网站，阅读“Data Warehousing Failures（数据仓库失败案例）”并回答问题。因为其中介绍了 8 个案例，所以课堂应被分成 8 组，每组分配一个案例。另外，阅读资料 Ariyachandra 和 Watson（2006a），分析每个案例的失败原因，不要专注于资料中一个或多个的成功因素。
7. 登录 teradatastudentnetwork.com 网站，阅读“Ad-Vent Technology：Using the MicroStrategy Sales Analytic Model（Ad-vent 技术：使用 Microstrategy 销售分析模型）”并回答相关问题。Microstrategy 软件可以从 TUN 网站下载。同时，你也许想使用 Barbara Wixom 关于 Microstrategy 软件的 PPT（关于 Microstrategy 教程的演示幻灯片），这也可以从 TUN 网站下载。
8. 登录 teradatastudentnetwork.com 网站，观看名为“Real-Time Data Warehousing：The Next Generation of Decision Support Data Management（实时数据仓库：下一代决策支持数据管理）”和“Building the Real-Time Enterprise（建立实时企业）”的网络研讨会。阅读文章“Teradata's Real-Time Enterprise Reference Architecture：A Blueprint for the Future of IT（Teradata 的实时企业相关架构：IT 的未来蓝图）”，这也可以从此网站下载。介绍实时概念和技术是如何发挥作用的，企业如何运用它们来扩展现行数据仓库和 BI 架构以支持日常决策制定的。撰写一份报告，阐述实时数据仓库如何使企业具备竞争优势。详细阐述项目实施和操作中的困难，并说明在实际中如何解决它们。
9. 登录 teradatastudentnetwork.com 网站，观看网络研讨小组“Data Integration Renaissance：New Drivers and Emerging Approaches（数据集成复兴之路：新驱动和新兴方法）”和“In Search of Single Version of the Truth：Strategies for Consolidating Analytic Silos（寻找唯一的事实：巩固分析的战略）”和“Data Integration：Using ETL，EAI 和 EII Tools to Create an Integrated Enterprise（数据集成：运用 ETL、EAI 和

EII 工具来创建集成企业)"。同时阅读"Data Integration(数据集成)"研究报告。比较这些报告。这些报告中介绍的最重要的问题是什么?将数据集市和电子表格集成为一个数据仓库架构具有挑战性。应对挑战的最佳方法是什么?网络搜索相关领域的最新进展。将上述报告与你所发现的最新进展进行比较。

10. 研究数据仓库领域的未来发展趋势。对此主题进行网络搜索。同时,阅读2篇文章:L. Agosta,"Data Warehousing in a Flat World: Trends for 2006(扁平世界中的数据仓库:2006年的发展趋势)",DM Direct Newsletter,2006年3月31日;J. C. Geiger,"CIFs: Evolving With the Times(CIFs:随着时间的不断演进)",DM Review,2005年11月:38-41。比较你在这2篇文章中的发现。
11. 登录 teradatastudentnetwork.com 网站。阅读最新的文章、研究报告和数据仓库案例。介绍此领域中的最新进展。撰写报告说明数据仓库如何在 BI 和 DSS 中应用。

小组作业和角色扮演

1. 在过去的6年里,Kathryn Avery 担任了全国性连锁零售企业(Big chain)的 DBA。最近,她被任命主持 Big chain 的第一个数据仓库项目。该项目得到了高级管理人员和 CIO 的大力支持。开发数据仓库的目标是为了改善报表系统,特别是销售和营销领域的报表系统。从长远来看,项目目标是为改善 Big chain 的 CRM。Kathryn 曾参与数据仓库协会的会议,并阅读过相关文章。但是,她对开发方法仍有迷惑。她知道有2大集团:EDW(Inmon)和已架构的数据集市(Kimball)。

   最初,她认为两种方法完全不同,但是当她认真研究时,开始抱怀疑态度。Kathryn 有众多问题需要解答:

   a. 两种方案之间的不同之处在哪?

   b. 选择一个特定方案的重要原因是什么?

   c. 她下一步该采取什么措施?

   帮助 Kathryn 回答以上问题。(此练习基于此文献:K. Duncan,L. Reeves 和 J. Griffin,"BI Experts' Perspective(BI 专家观点)",Business Intelligence Journal(商务智能杂志),2003年,8(4):14-19。)
2. Jeet Kumar 是一个大型区域银行的数据仓库管理员。5年前他被任命去实施一个支持银行 CRM 业务战略的数据仓库项目。运用此数据仓库,银行能够成功地集成客户信息、分析客户利润、吸引客户、增强客户关系和保留客户。

   几年之后,由于更新数据的频率更加频繁,银行数据仓库逐渐发展为实时数据仓库。现在,该银行想实施客户自助服务和呼叫中心应用系统,这需要更快地更新数据。

   Jeet 希望让数据更新。一个替代选择是实施实时数据仓库项目。他的 ETL 供应商准备支持这个项目。然而,Jeet 已了解了 EAI 和 EII 技术,并想知道如何将这些技术运用到项目中。

   特别的,Jeet 有如下问题:

   a. 到底什么是 EAI 和 EII 技术?

   b. EAI 和 EII 技术与 ETL 技术有什么关系?

   c. EAI 和 EII 技术与实时数据仓库有何关系?

   d. EAI 和 EII 技术是实时数据仓库所需要的技术还是补充,或者是对实时数据仓库的替代?

   帮助 Jeet 回答以上问题。(此练习基于此文献:S. Brobst,F. Levy 和 C. Muxlla,"Enterprise Application Integration and Enteprise Information Integration(企业应用集成和企业信息集成)",Business Intelligence Journal(商务智能杂志),2005,10(2):27-32。)
3. 采访你大学的管理员或者企业的管理人员,判断数据仓库是如何协助他们工作的。撰写一份报告描述你的发现。报告中应包括成本估计和项目收益。
4. 浏览本章中的数据仓库风险列表,在实际项目中找出其中两个风险。
5. 访问 teradata.com 网站,阅读白皮书"Measuring Data Warehouse ROI(衡量数据仓库 ROI)"和"Realizing ROI: Projecting and Harvesting the Business Value of an Enterprise Data Warehouse(实现 ROI:发现并收获数据仓库的商业价值)"。同时,观看网络课程"The ROI Factor: How Leading Practitioners Deal With

the Tough Issue of Measuring DW ROI（影响 ROI 的因素：怎样引导开发人员处理衡量 DW 的 ROI 的难题）"。描述其中最重要的问题，并对这些问题与 Ariyachandra 和 Watson（2006）中介绍的成功因素进行比较。

6. 阅读 K. Liddell Avery 和 Hugh J. Watson 文章的"Training Data Warehouse End-users（培训数据仓库的终端用户）"，Business Intelligence Journal（商务智能杂志），2004，9（4）：40～51。（该文章可见于 teradatastudentnetwork. com 网站）。将不同的小组看成不同的终端用户，说明他们遇到的难题，并讨论对不同小组进行合理培训的好处。让小组中的每个成员扮演一个角色，讨论什么类型的数据仓库培训适合你。

网络练习

1. 上网搜索关于数据仓库的信息。识别出对此概念感兴趣的新闻。在图书馆、电子图书馆和 Google 上搜索关于 ABI/INFORM 的最新文章。登录 tdwi. com、technologyevaluation. com 和主要供应商：teradata. com、sas. com、oracle. com 和 ncr. com。也可访问 cio. com、dmreview. com、dssresources. com 和 db2mag. com。
2. 研究 ETL 工具和供应商。登录 fairisaac. com 和 egain. com。也可访问 dmreview. com。
3. 联系一些数据仓库供应商，获取他们产品的信息。特别关注提供复杂功能产品的供应商，例如 Cognos、Software A&G、SAS Institute 和 Oracle。可以从这些供应商那么获取免费的在线演示程序。下载一个或两个程序并运行它们。撰写一份实验报告。
4. 登录 teradata. com，下载数据仓库项目开发的成功案例。写一份报告介绍你的学习心得。
5. 登录 teradata. com 下载数据仓库的白皮书和网络课程。阅读白皮书并观看网络课程（将班级分组以完成所有任务）。写一份报告介绍你的学习心得。
6. 寻找数据仓库项目的最新成功案例。登录数据仓库供应商网站并寻找案例或者成功的故事。选择其中一个并向班级同学做简要介绍。

本章结尾应用案例

大陆航空公司因使用实时数据仓库而腾飞

当商务智能成为日常业务系统的关键组件时，提供用户快速更新和预警的实时数据仓库项目也在日渐增加。实时数据仓库和 BI 支持制定积极进取的"前行商业计划"，帮助改变了大陆航空公司的经营状况，将最差的经营状况转变为最好的，将最好转变为最喜爱的。大陆航空公司是实时 BI 方面的领军企业。2004 年，它赢得了数据仓库协会颁发的最佳实践和领导奖项。

问题

大陆航空公司成立于 1934 年，在美国西南部拥有一架单引擎洛克希德飞机。2006 年，大陆航空公司成为全美第 5 大航空公司和世界第 7 大航空公司。大陆航空公司拥有全美最广泛的全球航线网络，拥有通往超过 227 个目的地的超过 2 300 条航线。

回到 1994 年，大陆航空公司陷入了重大的财务危机。它两次申请美国破产法第 11 章的保护，并准备申请第 3 次，最终很可能破产，机票销量下跌，因为客户看重的因素表现都很差，包括飞机的准时起飞率很低、行李托运问题频繁、许多客户由于超重而离开。

解决方案

大陆航空公司崛起于 1994 年，当时 Gordon Bethune 担任公司 CEO 并发起了前行计划，该计划被分成 4 个部分立即执行。通过更好地理解客户需求和客户对公司服务的意见，Bethune 致力于提升客户价值绩效考核。财务管理活动也成为业务重组的一个目标。早在 1998 年，航空公司拥有分散的营销和业务系统，由外部供应商托管和管理。查询处理过程和向高价值客户宣传营销计划需要花费大量时间，且无效率。另外，工作人员制定决策时很难获取相关信息。1999 年，大陆航空公司选择了将营销系统、IT 系统、财务系统和业务数据源系统集成为一个内部 EDW。数据仓库在此过程中起了主要作用。

不久以后，大陆航空公司开始处于盈利状态，而且各项业绩指标均在航空领域排名第 1。Bethune 和他的管理团队提升了公司目标。除了业绩最好之外，他们期望大陆航空公司成为最受客户欢迎的航空公司。

前行计划采取了更多可行的方法，使得大陆航空公司从排名第1到最受客户欢迎的公司。技术成为支持该新行动方案的关键因素。早期，访问历史的、集成数据就可以满足公司需求。这产生了可观的战略价值。但是，数据仓库对实时的、可提起诉讼信息的需求日渐迫切，用于支持企业级业务决策制定和商业过程。

幸运的是，数据仓库开发团队已经预料到并安排了实时数据仓库项目。在开始时，团队就创建了一个架构，用来处理实时数据进入数据仓库的过程，从遗留系统中提取数据到数据仓库中，进行实时业务查询。2001年，可以从数据仓库中获取实时数据，存储数量迅快速增长。大陆航空公司将主要业务系统中的实时数据（从几分钟到几小时）转移到数据仓库中，这些数据是关于客户、机票预订、登机、运作和航线等方面的。大陆航空公司的实时数据仓库包括下列内容：

- 财务管理和会计
- 客户关系管理
- 机组人员运作和工资
- 系统安全与漏洞
- 飞行作业

结果

就第一年而言，在部署数据仓库项目后，大陆航空公司识别并消除了超过700万美元的诈骗，节约了4 100万美元的成本，伴随着6年内3 000万美元的软硬件投资，大陆航空公司在财政收入增加、营销成本节约、欺诈发现、需求预测和追踪和改进数据中心管理方面收益50 000万美元。管理人员对业务有着统一、集成、可靠的认识，因此可制定更好的、更快的决策。

大陆航空公司现在已成为实时BI领域的领军者，因为其系统具有以下功能：可扩展延伸的架构，实时捕获何种数据的实践决策、与终端用户的良好关系、精干的数据仓库成员、能够明智衡量战略和战术决策支持的需求、理解决策支持与业务之间的协同、运用实时数据改变商业流程。（请登录 teradata. com/t/page/139245/，查看大陆航空公司的样本系统输出屏幕。）

本章结尾应用案例的问题

1. 介绍大陆航空公司实施前行计划的好处。
2. 说明为什么航空公司运用实时数据仓库是重要的。
3. 浏览 teradata. com/t/page/139245/，的样本系统输出屏幕。说明它是如何帮助用户识别问题和机会的。
4. 识别传统数据仓库和大陆航空公司实施的实时数据仓库之间的主要区别。
5. 与传统数据仓库相比，大陆航空公司可从实时系统项目中获取什么战略优势？

参考文献

References

Adamson, C. (2009). *The Star Schema Handbook: The Complete Reference to Dimensional Data Warehouse Design*, Hoboken, NJ: Wiley.

Agosta, L. (2006, January). "The Data Strategy Adviser: The Year Ahead—Data Warehousing Trends 2006." *DM Review*, Vol. 16, No. 1.

Akbay, S. (2006, Quarter 1). "Data Warehousing in Real Time." *Business Intelligence Journal*, Vol. 11, No. 1.

Ambeo. "Keeping Data Private (and Knowing It): Moving Beyond Conventional Safeguards to Ensure Data Privacy." **am-beo.com/why_ambeo_white_papers.html** (accessed May 2009).

Ambeo. (2005, July). "Ambeo Delivers Proven Data Access Auditing Solution." *Database Trends and Applications*, Vol. 19, No. 7.

Ambeo, G. H. (2003, June 30). "Hilton Checks into New Suite." *Computerworld*, Vol. 37, No. 26.

Ariyachandra, T., and H. Watson. (2006a, January). "Benchmarks for BI and Data Warehousing Success." *DM Review*, Vol. 16, No. 1.

Ariyachandra, T., and H. Watson. (2006b). "Which Data Warehouse Architecture Is Most Successful?" *Business Intelligence Journal*, Vol. 11, No. 1.

Ariyachandra, T., and H. Watson. (2005). "Key Factors in Selecting a Data Warehouse Architecture." *Business Intelligence Journal*, Vol. 10, No. 2.

Armstrong, R. (2000, Quarter 3). "E-nalysis for the E-business." *Teradata Magazine Online*, **teradata.com**.

Ball, S. K. (2005, November 14). "Do You Need a Data Warehouse Layer in Your Business Intelligence Architecture?"

datawarehouse.ittoolbox.com/documents/industry-articles/do-you-need-a-data-warehouse-layer-in-your-business-intelligence-architecture-2729 (accessed June 2009).

Basu, R. (2003, November). "Challenges of Real-Time Data Warehousing." *DM Review*.

Bell, L. D. (2001, Spring). "MetaBusiness Meta Data for the Masses: Administering Knowledge Sharing for Your Data Warehouse." *Journal of Data Warehousing*, Vol. 6, No. 2.

Berander, A., B. Berander, A. Fadlalla, and G. James (2000, Winter). "Data Warehouse Administration and Management." *Information Systems Management*, Vol. 17, No. 1.

Bonde, A., and M. Kuckuk. (2004, April). "Real World Business Intelligence: The Implementation Perspective." *DM Review*, Vol. 14, No. 4.

Breslin, M. (2004, Winter). "Data Warehousing Battle of the Giants: Comparing the Basics of Kimball and Inmon Models." *Business Intelligence Journal*, Vol. 9, No. 1.

Briggs, L. L. "DirecTV Connects with Data Integration Solution." *Business Intelligence Journal*, Vol. 14, No. 1, 2009, pp. 14–16.

Brobst, S. E. Levy, and C. Morelli. (2005, Spring). "Enterprise Application Integration and Enterprise Information Integration."*Business Intelligence Journal*, Vol. 10, No. 2.

Brody, R. (2003, Summer). "Information Ethics in the Design and Use of Metadata." *IEEE Technology and Society Magazine*, Vol. 22, No. 2.

Brown, M. (2004, May 9–12). "8 Characteristics of a Successful Data Warehouse." *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference* (SUGI 29). Montreal, Canada.

Burdett, J., and S. Singh. (2009). "Challenges and Lessons Learned from Real-Time Data Warehousing." *Business Intelligence Journal*, Vol. 9, No. 4.

Colter, P. (2003, June 23). "'Active' Warehousing." *eWeek*, Vol. 20, No. 25.

Cooper, B. L., H. J. Watson, B. H. Wixom, and D. L. Goodhue. (2000). "Data Warehousing Supports Corporate Strategy at First American Corporation."*MIS Quarterly*, Vol. 24, No. 4, pp. 547–567.

Cooper, B. L., H. J. Watson, B. H. Wixom, and D. L. Goodhue. (1999, August 15–19). "Data Warehousing Supports Corporate Strategy at First American Corporation." SIM International Conference. Atlanta.

Davison, D. (2003, November 14). "Top 10 Risks of Offshore Outsourcing." META Group (now Gartner, Inc.) Research Report, Stamford, CT.

Dragoon, A. (2003, July 1). "All for One View." *CIO*.

Eckerson, W. (2005, April 1). "Data Warehouse Builders Advocate for Different Architectures." *Application Development Trends*.

Eckerson, W. (2003, Fall). "The Evolution of ETL." *Business Intelligence Journal*, Vol. 8, No. 4.

Eckerson, W., R. Hackathorn, M. McGlynn, C. Twogood, and G. Watson. (2009). "Data Warehousing Appliances." *Business Intelligence Journal*, Vol. 14, No. 1, pp. 40–48.

Edwards, M. (2003, Fall). "2003 Best Practices Awards Winners: Innovators in Business Intelligence and Data Warehousing." *Business Intelligence Journal*, Vol. 8, No 4.

"Eggs Customer Data Warehouse: Hits the Mark." (2005, October). *DM Review*, Vol. 15, No. 10, pp. 24–28.

Elson, R., and R. LeClerc. (2005). "Security and Privacy Concerns in the Data Warehouse Environment." *Business Intelligence Journal*, Vol. 10, No. 3.

Erickson, J. (2006, March). "Real-Time Realities." *BI Review*.

Furtado, P. (2002). "A Survey of Parallel and Distributed Data Warehouses." *International Journal of Data Warehousing and Mining*, Vol. 5, No. 2, pp. 57–78.

Golfarelli, M., and Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies*. San Francisco, CA: McGraw-Hill Osborne Media.

Gonzales, M. (2005, Quarter 1). "Active Data Warehouses Are Just One Approach for Combining Strategic and Technical Data." *DB2 Magazine*.

Hall, M. (2002, April 15). "Seeding for Data Growth." *Computerworld*, Vol. 36, No. 16.

Hicks, M. (2001, November 26). "Getting Pricing Just Right." *eWeek*, Vol. 18, No. 46.

Hoffer, J. A., M. B. Prescott, and F. R. McFadden. (2007). *Modern Database Management*, 8th ed. Upper Saddle River, NJ: Prentice Hall.

Hwang, M., and H. Xu. (2005, Fall). "A Survey of Data Warehousing Success Issues." *Business Intelligence Journal*, Vol. 10, No. 4.

IBM. (2009). *50 TB Data Warehouse Benchmark on IBM System Z*. Armonk, NY: IBM Redbooks.

Imhoff, C. (2001, May). "Power Up Your Enterprise Portal." *E-Business Advisor*.

Inmon, W. H. (2006, January). "Information Management: How Do You Tune a Data Warehouse?" *DM Review*, Vol. 16, No. 1.

Inmon, W. H. (2005). *Building the Data Warehouse*, 4th ed. New York: Wiley.

Jukic, N., and C. Lang. (2004, Summer). "Using Offshore Resources to Develop and Support Data Warehousing Applications." *Business Intelligence Journal*, Vol. 9, No. 3.

Kalido. "BP Lubricants Achieves BIGS Success." **kalido.com/collateral/Documents/English-US/CS-BP%20BIGS.pdf** (accessed August 2009).

Kalido. "BP Lubricants Achieves BIGS, Key IT Solutions." **keyitsolutions.com/asp/cpdetails/report/95/cat/1175/** (accessed August 2009).

Karacsony, K. (2006, January). "ETL Is a Symptom of the Problem, not the Solution." *DM Review*, Vol. 16, No. 1.

Kassam, S. (2002, April 16). "Freedom of Information." *Intelligent Enterprise*, Vol. 5, No. 7.

Kay, R. (2005, September 19). "ETL." *Computerworld*, Vol. 39, No. 38.

Kelly, C. (2001, June 14). "Calculating Data Warehousing ROI." *SearchSQLServer.com Tips*.

Malykhina, E. (2008, January 3). "The Real-Time Imperative." *InformationWeek*, Issue 1208.

Manglik, A., and V. Mehra. (2005, Winter). "Extending Enterprise BI Capabilities: New Patterns for Data Integration." *Business Intelligence Journal*, Vol. 10, No. 1.

Martins, C. (2005, December 13). "HP to Consolidate Data Marts into Single Warehouse." *Computerworld*.

Matney, D. (2003, Spring). "End-User Support Strategy." *Business Intelligence Journal*, Vol. 8, No. 2.

McCloskey, D. W. (2002). *Choosing Vendors and Products to Maximize Data Warehousing Success*. New York: Auerbach Publications.

Mehra, V. (2005, Summer). "Building a Metadata-Driven Enterprise: A Holistic Approach." *Business Intelligence Journal*, Vol. 10, No. 3.

Moseley, M. (2009). "Eliminating Data Warehouse Pressures with Master Data Services and SOA." *Business Intelligence Journal*, Vol. 14, No. 2, pp. 53–65.

Murtaza, A. (1998, Fall). "A Framework for Developing Enterprise Data Warehouses." *Information Systems*

Management, Vol. 15, No. 4.

Nash, K. S. (2002, July). "Chemical Reaction." *Baseline*.

Orovic, V. (2008, June). "To Do & Not to Do." *eAI Journal*.

Parzinger, M. J., and M. N. Frolick. (2001, July). "Creating Competitive Advantage Through Data Warehousing." *Information Strategy*, Vol. 17, No. 4.

Peterson, T. (2003, April 21). "Getting Real About Real Time." *Computerworld*, Vol. 37, No. 16.

Reeves, L. (2009). *Manager's Guide to Data Warehousing*. Hoboken, NJ: Wiley.

Romero, O., and A. Abelló. (2009). "A Survey of Multidimensional Modeling Methodologies." *International Journal of Data Warehousing and Mining*, Vol. 5, No. 2, pp. 1–24.

Rosenberg, A. (2006, Quarter 1). "Improving Query Performance in Data Warehouses." *Business Intelligence Journal*, Vol. 11, No. 1.

Russom, P. (2006). "Next Generation Data Warehouse Platforms." TDWI Best Practices Report, available at tdwi.org/research/reportseries/reports.aspx?pid=842 tdwi.org (accessed January 2010).

Saunders, T. (2009). "Cooking up a Data Warehouse." *Business Intelligence Journal*, Vol. 14, No. 2, pp. 16–22.

Schwartz, K. D. "Decisions at the Touch of a Button." *Teradata Magazine*. **teradata.com/t/page/117774/index.html** (accessed June 2009).

Schwartz, K. D. (2001, March). "Decisions at the Touch of a Button." *DSS Resources*, pp. 28–31. dssresources.com/cases/coca-colajapan/index.html (accessed April 2006).

Sen, A. (2004, April). "Metadata Management: Past, Present, and Future." *Decision Support Systems*, Vol. 37, No. 1.

Sen, A., and P. Sinha (2005). "A Comparison of Data Warehousing Methodologies." *Communications of the ACM*, Vol. 48, No. 3.

Solomon, M. (2005, Winter). "Ensuring a Successful Data Warehouse Initiative." *Information Systems Management*, Vol. 22, No. 1, 26–36.

Songini, M. L. (2004, February 2). "ETL Quickstudy." *Computerworld*, Vol. 38, No. 5.

Sun Microsystems. (2005, September 19). "Egg Banks on Sun to Hit the Mark with Customers." **sun.com/smi/Press/sunflash/2005-09/sunflash.20050919.1.xml** (accessed April 2006; no longer available online).

Vassiliadis, A. (2002, Spring). "Identifying Meta Data Requirements." *Journal of Data Warehousing*, Vol. 7, No. 2.

Teradata Corp. "A Large US-based Insurance Company Masters Its Finance Data." **teradata.com/t/WorkArea/DownloadAsset.aspx?id=4858** (accessed July 2009).

Teradata Corp. "Active Data Warehousing." **teradata.com/t/page/87127/index.html** (accessed April 2006).

Teradata Corp. "Coca-Cola Japan Puts the Fizz Back in Vending Machine Sales." **teradata.com/t/page/118866/index.html** (accessed June 2009).

Teradata Corp. "Enterprise Data Warehouse Delivers Cost Savings and Process Efficiencies." **teradata.com/t/resources/case-studies/NCR Corporation-eb4455/** (accessed June 2009).

Terr, S. (2004, February). "Real-Time Data Warehousing: Hardware and Software." *DM Review*, Vol. 14, No. 2.

Thornton, M. (2002, March 18). "What About Security? The Most Common, but Unwarranted, Objection to Hosted Data Warehouses." *DM Review*, Vol. 12, No. 3, pp. 30–43.

Thornton, M., and M. Lampa. (2002). "Hosted Data Warehouse." *Journal of Data Warehousing*, Vol. 7, No. 2, pp. 27–34.

Vaduva, A., and T. Vetterli. (2001, September). "Metadata Management for Data Warehousing: An Overview." *International Journal of Cooperative Information Systems*, Vol. 10, No. 3.

Van den Hoven, J. (1996). "Data Marts: Plan Big, Build Small." *Information Systems Management*, Vol. 15, No. 1.

Watson, H. J. (2002). "Recent Developments in Data Warehousing." *Communications of the ACM*, Vol. 8, No. 1.

Watson, H. J., D. L. Goodhue, and B. H. Wixom. (2002). "The Benefits of Data Warehousing: Why Some Organizations Realize Exceptional Payoffs." *Information & Management*, Vol. 39.

Watson, H. J., J. Gerard, L. Gonzalez, M. Haywood, and D. Fenton. (1999). "Data Warehouse Failures: Case Studies and Findings." *Journal of Data Warehousing*, Vol. 4, No. 1.

Weir, R. (2002, Winter). "Best Practices for Implementing a Data Warehouse." *Journal of Data Warehousing*, Vol. 7, No. 1.

Wilk, L. (2003, Spring). "Data Warehousing and Real-Time Computing." *Business Intelligence Journal*, Vol. 8, No. 2.

Wiemhöfer, R. (2009). "A Survey of Managing the Evolution of Data Warehouses." *International Journal of Data Warehousing and Mining*, Vol. 5, No. 2, pp. 24–56.

ZD Net UK. "Sun Case Study: Egg's Customer Data Warehouse." **whitepapers.zdnet.co.uk/0,39025945,60159401p-39000449q,00.htm** (accessed June 2009).

Zhao, X. (2005, October 7). "Meta Data Management Maturity Model." *DM Direct Newsletter*.

业务绩效管理

学习目标

- 全面理解 BPM
- 理解闭环过程如何将战略与实施相结合
- 描述计划和管理报告中出色的实践案例
- 描述绩效管理和指标的区别
- 理解 BPM 中各种方法论的作用
- 描述平衡记分卡和六西格玛原理
- 理解记分卡和仪表盘的区别
- 理解基本的仪表盘设计

业务绩效管理（Business Performance Management，BPM）是决策支持系统（Decision Support System，DSS）、企业信息系统（Enterprise Information System，EIS）、商务智能（Business Intelligence，BI）的进一步发展。从进入市场开始，它经过了25年的发展。由于融入了决策支持，所以BPM不仅仅是一项技术。它将过程、方法、度量和应用设计融为一体，覆盖了整个企业全面的财务和业务绩效管理。它能够帮助企业将其战略和目标转换成计划，监控违反计划的绩效，分析计划的结果和实际结果之间的差异，并通过调整企业的目标和行动来对分析结果进行反馈。

本章主要介绍BPM的基本过程、方法、度量和系统。由于BPM注重战略和方法，所以，它区别于DSS和BI，本章将从探寻企业的战略及其执行的概念，以及它们之间存在的差距开始谈起。

开篇场景：Harrah 公司加倍下注

从1937年开始运营的Harrah娱乐公司是世界上最大的博彩公司。在它运营的大部分时间里，其财务表现卓越并得到了前所未有的扩张。在2000年，Harrah公司旗下17个市场的21家赌博旅店遍布全美，员工超过40 000人，累计为1 900万名顾客提供服务。到了2008年，这些数据已经变为遍布6大洲的51家赌博酒店，85 000万员工，4 000万名顾客。Harrah公司的大部分成就归功于其精明的市场营销运作和优质的服务以及并购战略。

问题

除了在博彩业成为领军企业外，Harrah公司也一直是商务智能和绩效管理领域的领头羊。与竞争者不同的是，Harrah公司还常避免向旅店、购物中心和旅游景点无节制地投资。它的运营都是基于一个基本的商业战略："深入了解顾客的需求，向他们提供优质的服务，用他们的忠诚作为回报，这样无论何时何地当他们想玩的时候就会想到Harrah公司（Weston and Volonino，2001）。"这一战略的执行，得益于创意营销、善于应用信息技术和出色的经营。

这一战略在20世纪90年代后期，由时任Harrah公司首席运营官的Gary Loveman提出。现在，Loveman是Harrah娱乐公司的主席、董事长和首席执行官。在进入Harrah公司之前，Loveman是哈佛大学商务管理研究生院的副教授，他在零售市场营销和服务管理方面有着丰富的经验。当他进入Harrah公司时，分配给他的任务是将Harrah公司转换成"建立顾客品牌忠诚度的

市场导向的企业”（Swabey，2007）。当时，Harrah 公司几乎没有选择，没有足够的资金向它的竞争对手 Bellagio 那样建立新的豪华赌场和娱乐场中心。相反，它决定通过了解顾客的行为和表现，将投资回报率最大化。这是因为在高度竞争的博彩娱乐市场，吸引并留住顾客是一个企业成功的关键，因为顾客忠诚度和满意度可以成就也可以毁灭一个公司。吸引并满足顾客需求比豪华的住所或环境更加有效。但是，这个目标必须要通过说服赌徒们更多地购买 Harrah 公司的资产来实现。

因为 Harrah 公司的会员卡制度已经实行很多年了，所以它的顾客对它已经了解得很透彻（Swabey，2007）。但是，分组座谈会证实了管理层们的疑虑——顾客们也许有卡，但是他们并不忠诚于 Harrah 公司。他们博彩的将近 65% 的资金花在了别处。第一步就是要找出企业的顾客是谁。经过分析揭示两个事实：（1）超过 80% 的收益来自超过 25% 的顾客；（2）大部分顾客都是“普通人”（中老年人），而且并不是被奢侈的场所吸引来的（Shill and Thomas，2005）。Harrah 公司怎样收集、利用和阐释这些数据，分析和发现客户类型，以使终生价值最大化呢？

解决方案

Harrah 公司的答案是一个称为“Total Gold”的解决方案，“Total Gold”是一项申请了专利的客户忠诚度解决方案，现在又叫“Total Rewards”解决方案。这项方案不仅通过现金和赠券的方式回报顾客在 Harrah 的任何一家娱乐场所进行的博彩或者其他任何活动，更重要的是，这项计划向企业提供了广泛收集有关顾客及其行为的大量的、实时的交易信息。信息通过记录顾客所有行为（例如，在饭店消费、饮酒情况、在博彩中的损失情况等）的“Total Rewards”卡收集。

这些信息提供给中央数据仓库。世界各地的 Harrah 员工都可以访问这些数据。这个数据仓库构成了“闭环”的市场营销系统的基础，该系统使得 Harrah 公司在营销竞争中清晰地界定自己的目标，执行和监控这些活动，从中了解哪种特定类型的顾客在什么种类的活动中可以为企业带来最大的收益。综合的结果是 Harrah 公司建成了一个“可以不断提高客户服务交互和业务成果的、差别化的忠诚度和服务框架”（Stanley，2006），同时这一系统也向 Harrah 的运营系统提供实时信息，该运营系统可以对顾客博彩或参加 Harrah 公司的其他活动产生很大影响。

结果和产生的新问题

Harrah 公司的 Total Rewards 会员卡方案和闭环市场营销系统在过去的几十年里为它带来了可观的回报，包括（Watson and Volonino，2001）：

- Harrah 赌场的品牌认定
- 增加了价值几百万美元的忠诚于 Harrah 公司的顾客
- 增加了参加多个 Harrah 公司娱乐活动的顾客数量，增加了数百万美元的盈利能力
- 提高了企业信息技术投资方面的内部回报率

总之，相对于竞争对手，顾客在 Harrah 公司的任意消费每年都有可观的增长，这样的结果就是给企业增加了数亿美元的收益。

这一系统获过很多奖项（例如 TDWI 最佳实践奖），并且成为许多案例研究的课题。它曾经被评价为“当今最伟大、最成功的指导行动的案例”（Swabey，2007）。当然，奖品和荣誉不能成为将来成功的保障，尤其是面对全球经济不景气的时候。

到 2007 年年底的这 10 年中，美国娱乐业的每股收益明显高于其他行业（Knowledge@W. P. Carey，2009）。但是过去的两年发生了变化。虽然被认为是不受经济低迷的影响，但娱乐业实质上正在遭受资本市场和世界经济崩溃的影响。如拉斯维加斯等城市，不仅酒店入住率下降，平均每个游客的消费水平也在下降。很多赌场的情况并不确定，因为他们花费巨额的债务修建新的更大更豪华的酒店赌场，没有足够的资金储备摆脱收益下降的困境。

与它的竞争对手不同，Harrah 公司没有高大的建筑物（Shill and Thomas，2005）。但是，与

它的竞争对手一样，Harrah 也面临着大量的经济问题。在 2009 年的前 3 个月，它宣布的运营亏损为 12 700 万美元，虽然与上年同期相比有所减少。在 2008 年的前 3 个月，它的运营损失为 27 000万美元。在 2008 年，在 Harrah 公司从 Apollo 管理公司和 TGP 资本公司私下借贷之后，它的债务负担翻了一番（达到了 240 亿美元），如今它的高额贷款将它推到了破产的边缘。

所以，即使 Harrah 公司实施备受赞赏的绩效管理系统很多年了，被公认为数据使用和预测分析的带头人，但它仍然免不了有和它的“小型竞争者”一样的战略上的问题和经济问题。

Harrah 公司仍然依赖它的市场营销活动来增加需求。另外，它采取了一系列的举措来减少债务和花费。在 2008 年 12 月，Harrah 公司完成了债务交换方案，使其债务减少了 11.6 亿美元，并且正在实施另一项债务削减和成熟度延伸计划，这为它节省了 280 万美元。就像其他赌博公司一样，它在经济衰退期间解雇了拉斯维加斯的 1 600 名员工，削减管理人员的薪酬，暂停了 401K 捐献。尽管它延迟了在恺撒皇宫 660 多间房间的建设，但却一直在恺撒皇宫建设新的会议中心，预订很火爆。

管理层也受到了来自“效应-管理”流程的激励，这一流程由丰田公司提出，被称为精益运营管理。精益运营管理是注重效率而不是效力的绩效管理框架。Harrah 公司首先使用了这个框架的几个性能，并于 2009 年在全企业内推广。

开篇场景的问题

1. 描述 Harrah 公司的营销战略。Harrah 公司与它的竞争对手存在哪些不同？
2. Harrah 公司的 Total Rewards 项目指的是什么？
3. Harrah 公司的闭环营销系统的基本因素是什么？
4. Harrah 公司营销战略的结果是什么？
5. 现在 Harrah 公司面临什么样的经济问题？Total Rewards 系统可以在一定程度上解决这些问题吗？

我们从开篇场景中能够学到什么

在过去的几年中，Harrah 公司的闭环市场营销系统使得它可以实施明显区别于它的竞争对手的战略。这一系统同样提供了管理业务和战术关键监控指标的手段。这一系统的问题是建立在经济增长的假设上的，或者至少有稳定的需求。它做不到或在短时间内难以实现的是预测急剧减少的或不存在的需求，或者经济上的基本变化。就像 Harrah 的首席执行官 Loveman 所说的，“我们对经济衰退没有经验，我们习惯了过去增长一段时间里的基本或近期交易，现在还不是很清楚它的走向。”

来源：Compiled from Knowledge @ W. P. Carey, “High- Rolling Casinos Hit a Lose Streak,” March 2, 2009, knowledge. wpcarey. asu. edu/article. cfm? articleid = 1752#(accessed January 2010); S. Green, “Harrah’s Reports Loss, Says LV Properties Hit Hard,” *Las Vegas Sun*, March 13, 2009, lasvegassun. com/news/2009/mar/13/harrahs-reports-losss-says-lv-properties-hit-hard(accessed January 2010); W. Shill and R. Thomas, “Exploring the Mindset of the High Performer,” *Outlook Journal*, October 2005, accenture. com/Global/Research_and_Insights/Outlook/By_Issue/Y2005/ExploringPerformer. htm(accessed January 2010); T. Stanley, “High- Stakes Analytics,” *Information Week*, February 1, 2006, informationweek. com/shared/printableArticle. jhtml? articleID = 177103414(accessed January 2010); P. Swabey, “Nothing Left to Chance,” *Information Age*, January 18, 2007, information-age. com/channels/information- management/features/2772256/nothing- left- to- chance. thtml(accessed January 2010); and H. Watson and L. Volonino, “Harrah’ High Payoff from Customer Information,” *the Data Warehousing Institute Industry Study 2000—Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions*, January 2001, terry. uga. edu/ ~ hwatson/Harrahs. doc(accessed January 2010).

3.1 业务绩效管理概述

如同这章将要表述的那样，Harrah 公司的闭环市场营销系统具有绩效管理系统的所有特点。更重要的是，这一系统将 Harrah 公司的战略、计划、分析系统和行动贯穿一线，使得它

可以稳定地提高企业绩效。Harrah 公司最近的经验也告诉我们，相对于某个方面而言（例如仅仅关注市场或顾客忠诚度），成功的绩效管理应该是多方面的，同时要具备对假设提出质疑和探索的能力，特别是在不稳定的时期。组织如果想获得长期的成功，就需要坚持不懈地调整自己（Axson，2007）。企业的绩效管理进程是评估企业为存活和发展而进行改变和调整的有效途径。

3.1.1 BPM 定义

在商业和贸易领域中，绩效管理有许多名称，包括企业法人绩效管理（Corporate Performance Management，CPM）、企业绩效管理（Enterprise Performance Management，EPM）、战略企业管理（Strategic Enterprise Management，SEM）和业务绩效管理。CPM 是由市场分析公司 Gartner 提出的（gartner.com）。EPM 是 Oracle 公司旗下的仁科（PeopleSoft）公司提出的有相同含义的术语。SEM 是 SAP（sap.com）使用的术语。在这章中，使用的是 BPM 而不是其他的术语，因为这一术语最早由 BPM 标准协会提出，并且在 BPM 论坛中仍在沿用。术语**业务绩效管理**（Business Performance Management，BPM）是指企业用于计量、监控和管理业务绩效的业务流程、方法、指标和技术。它有 3 个主要的组成部分（Colbert，2009）：

1. 相关技术支持下的闭环管理和分析过程的整合，用于指导财务和运营活动
2. 用来在业务上定义战略目标，并计量和管理针对目标绩效的工具
3. 一系列核心的过程，包括财务和运营计划、合并和报表、建模、分析和监控关键绩效指标（Key Performance Indicator，KPI），并与企业战略紧密相连

3.1.2 比较 BPM 和 BI

BPM 是 BI 发展的产物，它融合了很多 BI 的技术、应用和技能。当 BPM 第一次作为独立的概念被提出时，人们对 BPM 和 BI 之间的区别感到疑惑。这会不会仅仅是相同概念的不同术语？或者 BPM 是 BI 的新一代，再或者这两者之间真的存在实质性的不同？因为以下的各种原因，这些疑惑今天依然存在：

- BPM 和 BI 工具和套件的推销和销售是同一家公司
- BI 也在不停地演变，导致两者之间原本存在的差异逐渐消失（例如 BI 曾经专注于部门内部而不是整个企业）
- BI 是 BPM 的关键要素

目前，*BI* 这个术语是用于描述访问、分析和报告企业相关数据的技术。它包括一系列连续的软件，如特定查询、报表、在线分析处理、仪表盘、记分卡、搜索和可视化等。这些软件产品开始是独立的工具，但是 BI 软件提供商已经将它们整合为 BI 套件。

BPM 被认为是"BI + 计划"，意思是 BPM 是 BI 和同一平台上计划的聚合，即计划、监控和分析整个周期（Calumo Group，2009）。BPM 包括的过程并不是全新的。事实上，每个大中型企业在对全局的战略计划及运营计划有反馈的地方（例如，预算、具体的计划、执行和测评）都存在进程。BPM 增加的是集成这些流程、方法、指标和系统，从而成为一个整体的解决方案。

BI 实践和软件可以说就是 BPM 解决方案的一部分。然而，BPM 不仅仅是软件。BPM 是企业级的战略，以防止企业牺牲总体的绩效来达到局部业务的最优化。BPM 不是一次性的方案或者只关注部分的方案。相反，BPM 是一系列不断发展的流程，如果使用得当，将对企业产生巨大影响。BPM 成功的关键就是将整个企业贯穿起来。它"帮助用户通过行动来实现他们'共同的事业'：达到绩效目标、执行企业战略、给相关利益者传递价值"（Tucker and Dunon，2009）。

这并不意味着BI项目不能有明确的战略、集中的控制，或者从根本上影响企业。例如，运输安全机构（Transportation Security Administration，TSA）使用称为绩效信息系统（Performance Information System，PIMS）的商务智能系统来跟踪客户流量、屏幕绩效（损耗、旷工、加班费和伤病）、危险项目和总客户吞吐量（Henschen，2008）。这个系统是以MicroStrategy（microstrategy.com）的商务智能软件为基础构架，系统每天的高级用户使用量达到2 500人，每周的临时用户达到9 500人。PIMS中的信息对TSA的运营十分重要，并且在某些情况下接受国会指令。TSA从最高层到最底层的员工都在使用这个系统，并在2007—2008年的财务年中，成功地节省了大约1亿美元的代理费用。很明显，这个系统具有战略和运行的价值。然而，它不是BPM系统。

图3-1 BPM周期

来源：W. Eckerson, "Performance Management Strategies: How to Create and Deploy Performance Management Strategies," TDWI Best Practices Report, 2009.

最基本的区别在于BPM是战略驱动的。它包含一系列从战略到行动的闭环过程，目的在于使企业的经营绩效达到最优（见图3-1）。这个周期暗示，达到最佳绩效要从确定目标和方向开始（也就是战略），制定达到这些目标的举措和计划（也就是计划），控制偏离目标和方向的真实绩效（也就是监控），采取改正的行动（也就是行动和调整）。3.3－3.6节将详细研究这些主要步骤。

3.1 节复习题

1. 定义BPM。
2. BPM和BI有何不同？它们的相同点有哪些？
3. 简要描述TSA的PIMS。
4. 列举BPM的主要步骤。

3.2 制定战略：我们想到哪里去

暂时将你想象为一个长跑者，正在为即将到来的比赛训练。在准备时，假如教练对你说："我对这个比赛不是很了解，也不确定距离是多少，但是我认为你应该出去，每天跑8个小时，直到比赛那天。最后就可以成功了。"如果教练这么说，你肯定认为教练在胡说。很明显，为了使训练计划有意义，你需要知道将要参加的是什么类型的比赛（例如，马拉松、半程马拉松，还是10英里），你期望的完成时间是多少（例如，取得前5名的成绩需要2小时10分钟）。你还应该知道自己的优势和劣势，以确定目标是否能够实行，为了达到这一目标需要做哪些准备（例如，在比赛的最后阶段的冲刺有困难）。

像上面的教练一样运营管理的公司数量十分惊人，特别是在企业不稳定或者困难时期。通常，反对的声音认为："制定战略，形成正式的计划太慢并且十分不灵活，你需要的是采取针对我们企业特殊时期的更醒目、更协调的行动，如果花费大量的时间定义目标、明确重点、形成战略、管理结果，必定有人在最后将你打败。"然而，没有明确的目标或目的，在行动的过程中很

难取舍。没有明确的优先次序，根本无法在这些挑选出的选择中决定怎样分配资源。没有计划，就无法对工作任务进行指导。没有分析和评价，就不能决定哪个机会将成功或者失败。目标、目的、优先次序、计划和批判性思维组成了意义明确的战略。

3.2.1 战略规划

战略这一术语有很多定义。由于它经常和其他术语混用，所以人们对此很容易感到困惑，例如战略愿景、战略重点。除去这些含义上的差异外，它们都关注同一个问题："未来我们想到哪里去？"对于大部分企业来说，战略规划中提供了这个问题的答案。你可以将战略规划想象成一幅地图，这幅地图详细描述了一个企业从它现在的状态到实现未来愿景要采取的一系列行动。

通常情况下，战略规划是从企业的上层开始，并且着眼于整个企业。由此，创建战略规划是为了企业的业务单位或职能部门。如果不考虑规划是为企业的哪一个层次制定的——企业全局、业务单位或是职能部门，战略规划接下来的工作都是很常见的过程。

1. 进行现状分析 现状分析回顾了企业目前的状况，（"我们在哪儿？"）为财务绩效和运营绩效建立底线和关键趋势。

2. 决定规划周期 传统上，企业制订计划都是以年为单位，规划周期为3～5年。规划时间很大程度上取决于市场的可变性及可预测性、产品的生命周期、企业的规模、技术革新率和行业的资本密集程度。市场环境越易变化，则越不容易预测；生命周期越短，企业规模就越小；技术更新越快，资本越不密集，规划的周期越短。

3. 企业环境分析 环境分析是评估企业的优势、劣势、机会及威胁。它用来定义和区分对企业产生潜在影响或直接影响的关键顾客、市场、竞争者、政府、人口、利益相关者和行业等因素。

4. 识别关键成功因素 关键成功因素（Critical Success Factor，CSF）描述企业如果想在自己的细分市场取得成功必须要擅长的东西。对制造业企业而言，关键成功因素的例子是产品质量和产品创新。对于提供低成本的企业，如沃尔玛，分销能力则是关键成功因素。

5. 完成差距分析 像环境分析一样，差距分析用于定义和排序企业内部在流程、结构、技术及应用方面的优势和劣势。这些差距反映了决策实际需求什么和企业真正能够提供什么。

6. 创建战略愿景 企业的战略愿景提供企业在未来应该发展成什么模样的景象——产品和市场的转变。通常，愿景表达了企业目前的状况和期望达到的状况。

7. 提出商业决策 这一步骤在于制定的计划基于前几步得到的数据和信息，并且与企业的愿景一致。常识告诉我们，制定战略应该利用企业的优势，利用机会，规避劣势，应对威胁。企业需要确定战略在企业内部的一致性，制定的战略与企业文化相吻合，企业所拥有的资源和资金能保证这一战略实现。

8. 确定战略目的和目标 不能为企业的财务和运营计划的制订过程明确指引方向的战略规划是不完整的。在运营或财务计划制定之前，必须先制定战略目标并且将其精炼并转换为明确的目标或目的。战略目的是一种规范企业目标的大概说明或宏观的行动步骤。在企业将战略目标转换为财务目标或运营目标之前，应该先将其转换为明确的目标或目的。战略目标定义了在一段确定的时间内的一定量的目标。例如，企业想要提高其资产回报率（Return On Asset，ROA）或者提高综合收益率，在企业细化运营计划之前需要将这些目标定量化（例如，将资产回报率从10%提高到15%，或者将利润率从5%提高到7%）。战略目的和目标指导企业的实际运营活动，并且能够根据整体目标追踪进展。

3.2.2 战略差距

制定长远战略是一回事，执行这一战略又是另一回事。在过去的几十年里，大量的研究表明，许多企业战略规划和执行这些战略规划之间的差距很突出。Monitor Group（Kaplan and Norton，2008）和 Conference Board（2008）指出，最近对高级主管的一项调查表明，准确描述"战略执行"成为企业最应该优先考虑的事。类似地，根据 Palladium Group（Norton，2007）的调查数据表明，90%的企业不能成功地执行它们的战略。尽管许多研究将"战略差距"的原因归为以下4种之一，但是造成这一现象的原因却是多样的。

1. **交流** 在很多企业中，只有很少一部分员工理解企业的战略。Palladium Group（Norton，2007）将这一数据确定为10%。一方面，在员工们从来没有见过或者听说过战略规划时就做出决定，并按照战略规划合作是困难的或不可能的。另一方面，即使规划得到了沟通，战略的清晰度通常也不高，因此没有人能够十分确定他们的行动是按照战略还是已经偏离了战略。

2. **确定报酬与激励** 将报酬与绩效结合在一起，对于成功执行战略十分重要。然而，激励计划通常与短期财务成果相结合，而不是与企业的战略规划相结合，甚至没有与企业运营计划中制定的战略激励相结合。短期计划的最大主导作用也比不上理性决策。Palladium Group（Norton，2007）指出70%的企业不能将它们的战略与中级管理层的激励机制相结合。

3. **焦点** 管理层通常将大把的时间花在外围问题上而不是集中于核心问题。时间通常被花在对一系列预算问题的争论上，而对企业的战略、财务计划与战略的联系或者隐含在这些联系中的设想并不关心。Palladium Group（Norton，2007）指出在许多企业中超过85%的管理者每个月讨论战略的时间少于1小时。

4. **资源** 除非战略性的提案有足够的资金和资源，否则失败是显而易见的。Palladium Group（Norton，2007）发现，低于40%的企业的战略规划与企业预算紧密联系在一起。

3.2 节复习题

1. 企业为什么需要一个成熟的战略规划？
2. 制定战略规划的基本任务是什么？
3. 制定战略规划和实际执行战略规划之间的差距在哪儿？

3.3 计划：我们如何达到那里

当运营管理者知道了是什么（也就是企业的目的和目标），那么下一步是提出如何做（也就是具体的运营和财务计划）。运营和财务计划回答了两个问题：采取什么样的战术和举措才能达到战略规划所确定的绩效目标？执行这些策略所期望的财务结果是什么？

3.3.1 运营计划

运营计划将企业的战略目的和目标转化成一系列成熟的战术和举措，对资源需求和对未来一段时间（通常为一年，但并不总是一年）期望的结果。实质上，运营计划就像用于保证企业的战略能够实现的项目计划一样。大部分运营计划都由一组战术和举措组成。运营计划的关键是要一体化。战略驱动战术，战术驱动结果。基本上来说，运营计划中定义的战术和举措需要直接与战略规划中的关键目的和目标相结合。如果某个战术与一个或更多的战略目标没有关联，那么管理层应该质疑这一战术及与其相关的举措是否真的有必要存在。3.8 节将讨论的 BPM 方法就是用来保证这些联系存在。

（Axson，2007）文献指出，运营计划既可以以战术为中心，也可以以预算为中心。在以战术

为中心的计划中，战术的制定要符合战略规划中的目的和目标。相反地，在以预算为中心的计划中，财务计划或预算应与目标财务价值一致。最佳实践企业使用以战术为中心的运营计划。这意味着他们通过定义可受的战术和举措，制定运营计划步骤以达到特定的目标。例如，如果一个商业目标是利润率增长10%（也就是说，收益和花费分别除以收益的比率之间的差异），那么企业首先应该决定它提高这一比率是通过增加收益、减少花费或者是两者都应用。如果将以规收为重点，那么这个问题就变为是进入新的市场或增加已有市场的销售额，提高现有产品的产量还是引入产品，或者两者同时应用。各用场景和举措必须权衡总体的风险、资源要求和财务能力。

3.3.2 财务计划和预算

在大多数企业中，资源趋向于匮乏。如果资源不匮乏，那么企业完全可以在抓住机会、解决问题或者击败对手方面投入大量的人力和财力。由于资源的缺乏，企业需要将人力和财力投入到它的战略和与之相关的战术上。企业的战略目标和关键度量应该在如何分配企业的有形和无形资产中从上到下起到驱动作用。很明显可持续运营需要支持，应该将关键的资源分配给最重要的战略规划。大多数企业用它们的预算和资金来分配资源。为了战略的成功，这两种方法都需要与企业的战略目标和战术仔细地匹配。

企业实现这种匹配的最好做法是基于运营计划制定财务计划，或者更直接一些，就是按特定的战术和举措安排分配资源。例如，如果其中的一项战术是开发新的销售渠道，那么预算的收益和费用需要分配到渠道，而不仅仅是将费用分配给特定的职能部门，例如市场部、研发部。没有这样典型的战术资源计划，就不能评价战术的成功与否，进而就不能评价战略的成功与否。这种联系能够帮助企业避免“随意”削减与战略相关的预算。将特定的预算限额项目与特定的战术和举措联系起来，以战术为基础的预算就很好建立和明确了。

财务计划和预算过程的逻辑结构通常从那些产生某些形式的收入或收益的战术开始。在销售产品或服务的企业中，产生利润的能力是基于直接生产产品的能力和提供服务的能力，或者被授权销售产品或提供服务的能力。在制定了预期收入之后，就能够确定相应的分发费用。通常需要来自多个部门或策略的输入信息。这意味着流程必须相互配合，明确并理解职能之间的依赖关系。除了这些合作输入信息，组织需要增加各种经常费用，以及需要的资本费用。一旦这些信息巩固，就可以显示按计划实施战略所需的费用、现金和资金需求。

3.3 节复习题

1. 运营计划的目的是什么？
2. 什么是以战术为中心的计划？什么是以预算为中心的计划？
3. 财务计划最重要的目的是什么？

3.4 监控：我们做得怎么样

在实施运营计划和财务计划的过程中，监控企业的绩效是必要的。监控绩效的综合框架应该包括以下两点：监控什么和怎样监控。因为方方面面都要兼顾是不可能的，所以组织需要集中监控特定的问题。在制定了关注的指标和措施之后，企业需要创建监控这些因素和进行有效反应的战略。

3.7 节和3.8 节将详细讨论在 BPM 系统中，如何确定评价的内容。我们暂时只需注意“评价什么”，通常由 CSF 确定，并且企业的目的与目标是在战略规划的制定过程中确立的。举例来说，如果一家乐器生产商的某个战略规划，是在未来的3年中每年都将现有生产线的总利润率提高5%，那么企业就需要对利润率进行全年的监控，用以观察能否达到每年5%的增长率。同样，

如果这家企业计划在未来的2年里，每个季度都引进一种新产品，那么企业需要在指定的时间期限内跟踪新产品的引进。

3.4.1 诊断控制系统

很多企业都在应用一种被称为诊断控制系统的工具监控企业的绩效，纠正目前绩效水平的偏差。甚至普遍存在于那些没有正式BPM过程或系统的企业。诊断控制系统是一种基于控制论的系统，这表明它包括输入，将输入转换为输出的过程，用于对比输出的标准或标杆，允许输出结果和标准之间差异信息进行沟通和参照的反馈通道。事实上，任何一种信息系统如果满足以下几点都可以用做诊断控制系统：（1）事先建立目标；（2）测量输出；（3）计算绝对或相对的绩效偏差；（4）将偏差信息反馈，用于调整输入或过程，使得绩效与现有目标和标准相一致。图3-2阐述了诊断控制系统的关键因素。平衡记分卡、绩效仪表盘、项目监控系统、人力资源系统和财务报告系统都是可以用于监控的系统的例子。

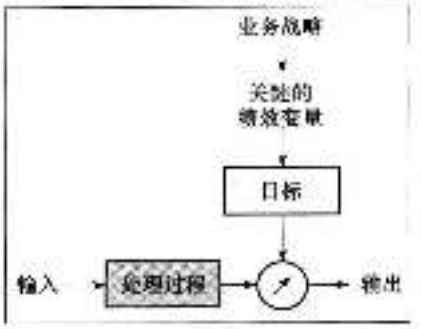


图3-2 诊断控制系统

来源：R. Simons, *Performance Measurement and Control Systems for Implementing Strategy*, Prentice Hall, Upper Saddle River, NJ, 2002, P. 207.

有效的诊断控制系统支持异常管理。并不是对内部过程和目标价值持续地监控，而是将实际结果与计划结果进行比较，管理者通常还会收到异常报告。管理者通常不会花费很多精力在与预期一致的评价上。但是，如果发现了巨大的偏差，那么管理者就需要投入时间和精力去调查出现偏差的原因并着手实施恰当的补救办法。

3.4.2 差异分析的困难

在很多企业中，当职能小组或者部门不能达到目标时，绝大多数的差异分析集中于消极差异上。很少集中于例如发现潜在机会一类的积极差异上，而且很少做差异模式下的假设分析。请考虑图3-3所描述的两条路径。在这张图中，从A到B的虚线表示某一特定时间的计划或目标结果。通过识别与计划之间存在的微小差异，我们可能希望实际结果与目标结果之间有轻微的偏差。当偏差超过预期的假设时，通常认为是运营发生了问题，需要进行纠正。这时候，管理者通常为了使计划重回正轨，通常命令员工不惜一切代价。如果没有达到预期收益，那么员工们就会受到责备进而更加努力的工作。如果费用超过计划，那么员工们就会被告知停止支出。

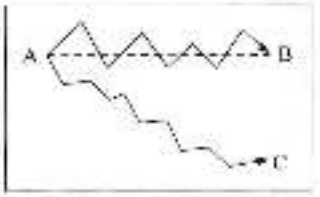


图3-3 运作的差异还是战略问题

然而，如果我们制定的战略假设出现错误——而不是执行出现错误，怎么办？如果企业需要将它的战略方向调整到C点而不是继续执行原来的计划，怎么办？就像应用案例3.1中所描述的那样，按照错误的前提假设行动其结果是灾难性的。做出此类决定的唯一办法就是对计划的绩效实行更密切的监控。不论企业应用哪一种诊断分析系统，都需要有对初步假设、因果关系和预定策略的被体有效性进行追踪。例如，试想企业要实行以推出新产品为主的成长战略。这类战略通常要依赖市场需求或零部件供应商的生产能力等方面的假设。在战略开展的同时，管理者不仅要监控与新产品有关的收益和费用，并且要关注市场需求，或者零部件的可用性，或者其他关键的假设与预期之间的偏差。

应用案例3.1 发现驱向型计划：咖啡之战

在过去的几年中，星巴克（Starbucks）、唐恩都乐（Dunkin' Donuts）（dunkindonuts.com）及麦当劳（McDonald）都卷入了占领精选咖啡市场的战争中。对于星巴克和唐恩都乐，这场战争的主要部分是围绕着增加店面数量展开的。但是麦当劳并没有这么做，因为它早就已经将店面开遍全世界了。

2000年以来，星巴克就以"非凡的速度"增开新店。它的店面数目从2000年不到4 000家迅速增长为2007年的15 000家左右。这一行动的假设前提是，人们对精选咖啡的需求还没有完全释放出来，如果不开更多的店满足人们的需求，那么它的竞争对手就会抢足先登。其中竞争者之一就是唐恩都乐。在2007年，唐恩都乐决定在数量和地域覆盖上拓展其特许经营权。在2007年之前，唐恩都乐仅在东北部就有5 000家特许经营店（Weier，2007）。它制定的新目标是在全球范围内开设15 000家特许经营店。唐恩都乐与星巴克不同的是，它没有自己的店面。相反，它依赖于个人申请特许经营权，通过申请审批后支付特许经营权使用费，特许经营权使用费从日常收入中扣除。

唐恩都乐为了实现特许经营的目标，引进了新的仪表盘应用程序（参看3.10节），用于帮助其查看哪里的生意不好做了，哪些交易之间太接近了，关闭一家特许经营交易的平均周期是多长，交易的平均规模有多大（Weier，2007）。假设唐恩都乐发现平均周期比预计的要长或者生意不好做了，那么应该采取什么样的应对措施？

按照唐恩都乐在战略中的说法，首要的措施当然是考虑增长周期或者确定为什么生意不好做。采取的最后一项行动才是对整个在全球范围内增设新店的战略和人们被抑制的需求这一基本假设进行质疑。事实上，这就是星巴克的做法。

即使面临店面销售额大幅下降的趋势，这也是评价开设至少一年的店面销售额增长速度的指标，星巴克仍然持续以飞快的速度开设新店（Wailgum，2008）。星巴克对这一问题的第一反应是如何解决销售额下降。在2007年，它宣布了一系列针对这一问题的战略举措。它推出新的混合咖啡，用新的设备替换已有的咖啡机，推出回报顾客工程及开设一个新网站。直到2008年1月，星巴克才意识到需要修改它的扩张战略，就是将新开的店与现有的店合并。相应地，星巴克缩减其扩张计划，减少其每年新开店数目、取消它开设40 000家店的长期目标，并着手关闭在美国不盈利的店面。

发现驱向型计划

当星巴克和唐恩都乐这样的大公司在着手企业级的扩张战略时，很大一部分成就取决于企业中的每个人是否都在努力。如果步入歧途，就会出现各种偏差，不惜任何代价为了坚持执行这一计划直接或间接向员工施加压力。特别是在竞争激烈、完全公开的环境中，企业都有这些倾向：

- **确认偏差** 导致只接受那些对已有的假设支持的信息，而抵制对其质疑的信息。
- **近期偏差** 由于对最初关键性假设的淡忘，造成通过实施经验来解释或理解上的困难。
- **赢家偏差** 在竞争中过分看重输赢，即使付出的代价超出了得到的利益。
- **社会或政治偏差** 过分坚持"公共"计划，而不接受无知或错误。

部分问题是星巴克和唐恩都乐所采用的那类常规计划过程并不能在研究和分析基础假设方面提供什么。为了改变常规的计划制定过程，McGrath and MacMillan（2009）年提出，公司

应该使用发现驱动型计划（Discovery-Driven Planning，DDP）。对于大多数成长型战略来说，结果是很难确定和预料的。它们同样也要依靠在计划执行过程中会发生改变的种种重要假设。随着战略的进行，关键问题是降低"假设－知识"的比例（即将假设都变成事实）。这就是发现驱动型计划的核心所在。发现驱动型计划提供了系统的方法，能够发现那些不被关注和没有引起争论但又存在问题的假设。之所以称之为发现驱动型是因为在计划的执行过程中，会涌现新的数据和发现新的可能性。

DDP包括一系列的步骤。其中一些与常见的计划制定过程相同（例如，创建成长型战略），另一些则十分不同。在这里的讨论中，DDP中存在3个与常见计划过程不同的步骤：

1. **逆向财务** 第一步是用一组财务单据模拟计划中所有种种假设如何相互影响，随着获取信息的增加，确定计划要得到更多的支持还是存在风险。

2. **支出规范** 第二步是安排产品、销售、服务所需要的全部活动，并将这些产品或服务提供给顾客。这些活动就代表了可列支的费用。

3. **假设清单** 业务开展需要的所有活动都依赖于关键性假设。在这一步中，列出了与第2步中支出项目相关的各项假设的书面清单。

假如你要开一家高级法式餐厅，你的目标是在经营的第一年就有所突破（获得200万销售额）。这里存在一个问题，"这是不是真实的销售数据？"更确切地说，"如果想要得到200万的销售额，应该采取什么样的行动？这些行动有意义吗？"

回答这些问题的一个方法就是考虑你的餐厅一年接待的顾客数量，以及他们在餐厅进餐时的平均消费金额是多少。按照每个人的平均消费，就可以猜测出平均每餐的规模（例如小菜、主菜等）及这些菜的平均成本，或者你可以查看整个餐厅的账单。比如在你生活区域内，其他高端法式餐厅每餐的平均消费为每人120～150美元。有了这些数据，你就知道每年需要接待13 333～16 667位顾客，或者每晚接待44～56位顾客。问题是：这些数据有意义吗？是不是太乐观了？如果是，你就需要调整目标。不论答案是什么，你仍然需要安排这些行动和为了实现这一目标的相关费用。

一旦推出了成长战略，DDP就帮助识别检查点和假设清单，使企业不仅能够估计现在的绩效，并且能够判断以前和现在计划所依靠的假设的持续有效性。如果星巴克使用了DDP，那么它也许能更早地发现其成长战略的缺陷。

来源：Compiled from R. McGrath and I. MacMillan, *Discovery-Driven Growth*, Cambridge, MA, Harvard University Press, 2009; T. Wailgum, "How IT Systems Can Help Starbucks Fix Itself," *CIO*, January 25, 2008, cio.com/article/176003/How_IT_Systems_Can_Help_Starbucks_Fix_Itself (accessed January 2010); M. Weier, "Dunkin' Donuts Uses Business Intelligence in War Against Starbucks," *Information Week*, April 16, 2007.

3.4 节复习题

1. 监控系统回答了哪些关键问题？
2. 构成诊断控制系统的关键因素是什么？
3. 什么是意外管理？
4. 从管理的观点看，差异分析的主要缺陷是什么？

3.5 行动和调整：我们需要做什么不同的吗

不论企业想发展它的业务还是仅仅想改善其运作，事实上，所有的战略都基于新的计划——设计新产品、进入新市场、获取新客户或业务，或者使业务流程合理化。大多数企业在实施它们

的新计划时过于乐观而不够客观，忽略了事实上大多数新项目或企业都是以失败告终（Slywotky and Weber，2007）。失败的概率有多大？很明显，这与项目类型有关。好莱坞电影失败的概率为60%，与收购类并失败的概率相同。IT项目的失败率为70%，对于新的食品，失败率为80%。而对于新的药品，失败率则更高，达到90%左右。总体来说，对于大多数新项目或者冒险的失败率在60%～90%之间。

一个项目会以多种不同方式失败，比如，考虑的选项或场景太少，不能成功地预测竞争者的行动，忽略经济或社会环境的变化，错误地预测需求，低估要取得成功所需要的投资等，这里只是列举了一些可能性。这就是为什么企业要持续监控结果，分析发生了什么，确定为什么发生，并适时调整它的行动的重要原因。

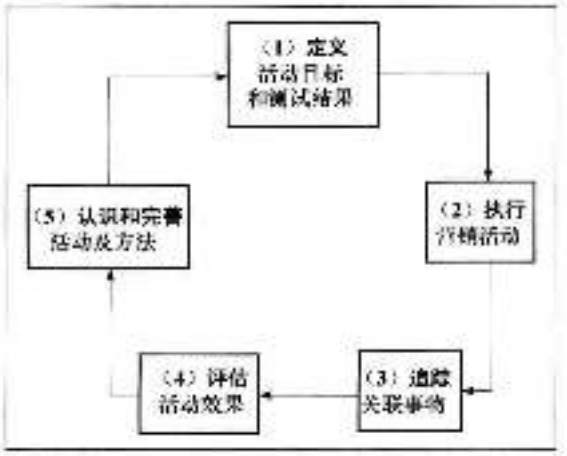


图3-4 Harrah公司闭环营销模型

来源：Watson, H. , and L. Volonino. "Harrah's High Payoff from Customer Information," *The Data Warehousing Institute Industry Study 2000—Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions*, Jan 2001. terry. uga. edu/~hwatson/Harrahs. doc(accessed January 2010).

同顾一下在开篇场景中，Harrah公司的闭环营销系统。图3-4描述了这一系统。就像图3-4所示，这个过程分为5个步骤：

1. 这一循环首先确定市场活动或测试步骤量化指标，方式是对比实验组与实验，对照组客户期望值或预期结果。

2. 接下来的活动或实验称为执行。这些活动用于提供及时准确的报价或信息。被选中的顾客及他们享受的待遇与他们在Harrah公司先前的经历有关。

3. 响应这个活动的每位顾客都要被追踪。不仅要评价响应率，而且对其他一些指标也要进行评价，比如激励产生的收益，以及这个激励有没有对顾客行为产生积极影响（例如，光顾频率的增加，光顾收益的增加，或者在各种赌博场所间穿梭）。

4. 通过这一活动产生的净值及其相对其他活动的盈利能力评价活动是否有效。

5. Harrah公司认识到激励对顾客行为的影响最明显，并导致了最佳盈利能力的提升。这些知识也被继续用于完善它的营销方法。

在过去的几年中，Harrah公司实际上进行了成千上万次这样的测试。尽管这5个步骤都很重要，但事实上Harrah公司为了得到最理想的结果而不断地分析和调整其战略，以期在竞争中取得优势。

像Harrah公司一样，很多企业花费大量的时间和金钱制订计划、收集数据和生成管理报告。然而，大部分企业在绩效管理实践方面缺乏竞争力。Saxon Group的研究结果表明（Axson，2007）：

> 大多数企业试图应用已经存在了半个多世纪的管理实践去管理日益不确定的、复杂的流程。详细的5年战略规划，静态年度预算、定期报告、缺乏灵活性的财务预测等，在管理变革、不确定及复杂的环境中，大部分是无效的管理工具。但是，很多企业还保留着这种管理方式。

Saxon Group咨询公司由曾在Hackett Group任职的David Axson领导，这是一家全球性的咨询公司，在最佳实践顾问、标杆管理、变革咨询服务方面尤为突出。Axson个人参与实践了300多起标杆管理案例。从2005年中期到2006年中期，有1 000多名来自北美、欧洲和亚洲的财务主管参与了Saxon Group领导的调查或工作会议，致力于研究当前商务管理艺术的发展状态。所有

主要的产业集团公司都参与其中。其中，25%的企业年利润少于500万，55%的企业年利润在500万～50亿之间，另外的20%年利润超过50亿。

以下为Saxon Group研究小组的调查总结（Axson，2007）：

- 只有20%的企业使用了综合的绩效管理系统，尽管5年前还不到10%。
- 少于30%的企业实行的计划清晰地反映了主要项目和举措的预期结果。相反，它们关注于出现错误的事情。每个项目的财务计划没有显示出每项举措预期的成本和收益，也没有确定与之相关的总投资。战术计划不能描述实施的主要举措。
- 报告给管理层的信息中，超过75%是滞后的或者是集中于内部数据；只有少于25%的数据具有预测价值或者是关注于市场的。
- 普通的员工在所谓的高价值分析和决策支持任务花费的时间少于20%。基础的工作，如收集和验证高价值工作所需数据占掉了普通员工的大部分时间。

对普通公司制订计划和报告行动最大的影响是，管理层很少有时间从战略角度审视结果，决定应该采取哪些不同的行动及如何改进计划。事实上企业的战略、战术和期望的结果之间存在的关联少之又少（Axson，2007）。

……当事情没有严格按照计划发生时——通常是这样的，导致很多企业危险的暴露。对策略和目标之间的因果关心缺乏清晰的理解，你就无法决定现在的行动能产生预计的结果。最佳的实践企业不一定有更好的预测或计划；但是，它们能够快速地发现变化或问题，找出根本原因，采取纠正措施。

3.5 节复习题

1. 为什么60%～80%的新项目或企业都以失败告终？
2. 描述Harrah公司闭环销售系统模型的基本步骤。
3. 根据Saxon研究小组的研究结果，普通公司的绩效管理实践是什么？
4. 为什么很少有企业有时间分析战略和战术结果并根据这些分析采取纠正措施？

3.6 绩效评价

BPM基本上是一种绩效评价系统。依据Simons（2002）的描述，绩效评价系统：

> 帮助管理者通过对比实际结果与战略目标和目的，跟踪商业战略的实现情况。绩效评价系统通常指确定商业目标，并定期反馈进展报告的系统性方法。

所有的指标都是相对而言的。未经处理过的数据几乎没有价值。如果告诉你，一个销售人员在一个月内完成了他应该完成销售额的50%，这几乎没有任何意义。现在，告诉你同一个销售人员去年的月完成率仅为38%。很明显，这个趋势是好的。如果你还被告知这家公司的全体销售人员的平均完成率为80%又如何？很明显，这名销售人员需要加快速度了。就如Simons的定义，在绩效管理中，关键的对比包括战略、目标和方向。

3.6.1 KPI和业务指标

“一般化”指标与“针对战略”的指标之间有很大差异。**关键绩效指标**（Key Performance Indicator，KPI）一词通常表示后一种指标。关键绩效指标表示要达到一个目标所需的战略方向和业绩衡量。Eckerson（2009）指出，KPI是多维的。简单的解释就是，KPI具有多种不同的特点，包括：

- **战略** KPI体现了战略目标。

- **目标** KPI针对特定目标进行业绩衡量。目标在战略、计划或进行预算时确定，可以采取不同的形式（如，完成任务、缩减目标、绝对目标等）。
- **范围** 目标有绩效范围（如，高于、等于或者低于目标）。
- **编码** 将范围在软件中编码，方便直观地显示绩效（如，绿色、黄色、红色等）。编码可以使用百分比或者更复杂的规则。
- **时间范围** 制定目标时，必须有时间范围，即明确它们在什么时间必须完成。时间范围通常被细分为更短的时间间隔，用来提供绩效的里程目标。
- **标准** 用基准线或标准来评价目标。前些年的成果通常作为标准，但是也可以使用任意的数据或外部的标准。

KPI可分为"结果型"和"驱动型"。结果型的KPI（有时称为滞后指标）用于评价过去活动的产出（如收益）。通常它们实质上就是财务指标，但也并不总是。驱动型的KPI（有时又称为先行指标或价值动因）用于评价对KPI结果有重要影响的活动（如销售机会）。

在某些情况下，驱动型的KPI又称为运营KPI，这是一种矛盾修辞法（Hatch，2008）。大多数企业都有多种多样的运营指标。就如它的名字显示的，这些运营指标用来评价企业的运营活动或绩效。以下列举的例子表现了这些运营指标覆盖的不同范围：

- **顾客绩效** 顾客满意度、解决问题的速度和准确度、顾客维系等指标。
- **服务绩效** 服务电话的解决率、服务更新率、服务水平协议、交付效率和回报率指标。
- **销售运营** 新的销售渠道账户、落实的销售会议、将潜在转变为机会、服务订单的平均完成时间等指标。
- **销售计划/预测** 价格与购买之间关系的准确性、采购订单的履行率、取得的数量、预测与计划的比例以及所有完成的合同。

运营指标是否是战略性的，取决于企业和它采用的评价。在许多情况下，这些指标代表了战略成果的关键驱动因素。例如，文献Hatch（2008）回顾了一段时间内中档酒经销商都集中到上游成为供应商或者集中到下游成为零售商。相应地，要重点关注4种运营指标：手头/实时存货能力、突出"开放"订单价值、新网络账户、促销成本和市场投资回报率。这一努力的最终结果是1年内收益增长12%。很明显，这些运营指标是关键驱动因素。然而，就像在下一节中说的那样，在很多情况下，企业仅仅是出于方便，很少考虑为什么收集这些数据。结果是浪费了大量时间、精力和资金。

3.6.2 现有绩效评价系统存在的问题

如果你对大量企业进行调查，那么你就会发现企业时时刻刻在抱怨它的绩效评价系统（相对绩效管理系统而言）。使用最广泛的系统与Kaplan和Norton的平衡记分卡（Balanced Scorecard，BSC）有一些不同。各种调查和标准研究显示50%～90%的企业已经一次或分多次实现了不同形式的BSC。例如，从1993年以来，每年Bain & company都对国际高管进行调查，确定哪种管理工具在全球范围内使用最广泛（Rigby and Bilodeau，2009）。参与2008年度调研的有1 400名管理人员。根据这项调查，53%的企业正在实验某种BSC。在大多数调查中，当请管理人员描述他们的BSC时，似乎他们对什么构成"平衡"有些不解。但是，BSC的创始人，Kaplan and Norton（1966）对此很清楚：

> BSC方法的核心是对组织战略评价系统的整体构想。这一方法包含4个方面的视角：顾客视角、组织内部视角、学习与成长视角以及基于这三个指标的财务分析视角。

然而，如同Saxon Group发现的一样，绝大多数评价指标实际上都是财务指标（65%），或者是滞后指标（80%），还有实际上都是内部指标而不是外部指标（75%）。这些企业真正引有

的是“记分卡”——一系列报告、报表和特定的显示方式使他们能够用各种指标将实际结果与预计结果进行对比。

日历驱动型财务报告是绩效评价系统中的主要组成部分。这并不奇怪。第一点，这些系统中的很大一部分属于财务部门的权限。第二点，大部分企业（Saxon 认为是67%）认为计划过程是每年需要完成的财务活动。第三点，大部分管理者几乎不关注财务或运营数据之外的数据。研究表明，管理者重视种种不同形式的数据（如财务、运营、市场、顾客），但是他们认为大多数财务或运营之外的数据并不可靠，并且不愿意将他们的工作花费在那些信息的质量上。

将财务数据作为绩效评价系统核心的缺点很明显，最常见的缺陷如下所列：

- 财务指标通常由组织结构提出报告（如研发的支出），而不是由产生财务费用的过程提出。
- 财务指标是滞后指标，表明已经发生了什么而不是为什么发生或将来可能发生什么。
- 财务指标（如管理费用）通常是与产生费用的基本流程不相关的分摊结果。
- 财务指标关注短期行为，几乎不能提供长期信息。

如今对绩效评价系统产生危害的并不仅仅是财务的缺乏远见。指标过量和指标倾向性也是目前危害系统的主要问题。

企业自豪地宣布它们有200个甚至更多的企业级指标的现象屡见不鲜。很难想象如何驾驶一辆有着200个仪表盘的汽车。然而，似乎我们觉得管理一家有着200个指标的企业难度并不大，即使我们知道人类只能同时关注少量问题，其他问题就简单地推到一边。企业很少放弃它们已经收集的指标，导致指标过量的现象愈发严重。如果出现新的数据或数据要求，企业往往简单地把它们添加到已有的列表中。如果今天有200个指标，那么到了明天就会有201个，后天就会增加为202个。即使计划变了，或者出现了机会和问题，而且迅速增加，也很少有人考虑使用的指标清单是否适用于当前的情况。

相对于许多指标被跟踪，很多指标缺乏连续的管理。Michael hammer（2003）指出，这就是倾斜原理。一方面，像每股收入、股本回报率、盈利能力、顾客满意度这类的指标需要监控。另一方面，这些指标只能以帕顿向性的方式取得。能够控制的只有每个员工或雇员的行动。不幸的是，任何个别的行动对于一个企业的战略或商业团体来说的影响几乎可以忽略。控制的关键是实行一种战略性的商业模式或方法，这种战略性的商业模式或方法从顶层开始，围绕企业战略和目标，一直延伸到底层，通过执行者的行为表现出来。

3.6.3 有效的绩效指标

很多书都提供了一些识别绩效指标好坏与否的秘诀。其中好的要素包括以下方面：

- 指标应该关注关键的因素。
- 指标应该包括过去、现在和将来。
- 指标应该平衡股东、员工、合作伙伴、供应商和其他利益相关者之间的需求。
- 指标应该上至高层下到基层。
- 指标需要有研究和实际的目的性而不是随意的。

就像这节中在KPI部分所指出的那样，虽然所有这些因素都很重要，但对一个有效的绩效评价系统，真正关键的是要有恰当的战略。指标应该能够从企业或商业团体的战略中以及从分析实现这一战略的关键商业流程中提炼出来。当然，说起来容易做起来难。如果这很简单的话，大多数企业肯定已经拥有了有效的战略评价系统，但是事实是他们没有。

应用案例3.2描述了Expedia.com的基于网络的KPI记分卡系统，讲述了定义结果和驱动型KPI的困难，以及将部门KPI与企业方向匹配的重要性。

应用案例3.2 Expedia.com的顾客满意记分卡

Expedia公司是很多世界领先旅游公司的母公司，它向美国和全世界的个人和团体提供旅游产品和服务，它拥有并经营着许多著名的多样化的产品，包括Expedia.com、Hotel.com、Hotwire.com、TripAdvisor、Egencia、Classic Vacations及其他本土的和国际的业务。公司的旅游业务包括航班、酒店住宿、汽车租赁、目的地服务、游轮以及折扣线、住宿、汽车租赁公司、目的地服务、游轮及其他旅游产品和服务公司联合起来实行的一揽子旅游方案。它也提供预定旅店房间，机票、租车和来自旅游公司的目的地服务。它在这些交易中充当代理的角色，帮助旅客向其他公司预定航班、酒店、租车和游轮。这些受欢迎的品牌和新颖的业务，使Expedia成为全球最大的在线旅游机构，美国第三大旅游公司，全球第四大旅游公司。它的使命是成为全世界最大、最盈利的旅游公司，帮助各个地方的每个人计划并购买旅游中的一切。

问题

顾客满意度是Expedia公司的使命、战略和成功的关键。因为Expedia.com是在线服务，所以顾客的购买经历对Expedia的利润至关重要。在线购买经历可以成就或毁灭在线交易。顾客在线购买经历是愉快的旅行经历的写照。因为顾客的在线购买经历很重要，所以应该追踪、监控顾客的所有评论，当出现问题时尽快解决。不幸的是，几年之前，Expedia并没有重视"顾客之声"。没有统一的评价满意度的指标，分析满意度的驱动因素或者确定顾客满意度对公司盈利能力或整个企业目标的影响。

解决方案

Expedia所出现的问题并不是因为缺少数据。Expedia负责顾客满意度的小组知道他们有足够的数据。20个不同拥有者一共拥有20个不同的数据库。最初，公司指派其中一名商业分析师，将这些不同的数据库进行合并与集成，从中找出影响顾客满意度的关键指标。这名商业分析师每月花2-3周的时间进行数据的合并和集成，实际上没有时间进行分析。最终，小组发现仅仅进行数据集成是不够的。需要在战略背景下对数据进行分析，员工们有权知道分析的结果。

为了解决这一问题，小组决定完善显示方式。从部门绩效的基本驱动因素以及这一绩效和Expedia整体绩效的联系进行详细分析。接下来，小组修改了这些驱动因素并使之与记分卡关联。这一过程包括3个步骤：

1. **确定如何评价满意度** 这要求小组确定20个数据库中哪些指标对描述顾客满意度适用。这成为记分卡和KPI的基础。

2. **设定正确的绩效目标** 这要求小组决定KPI目标既要有短期的回报还要有长期的回报。顾客对其在线经历满意并不意味着会对卖方提供的旅游服务满意。

3. **输入数据** 小组需要将数据持续不断地与顾客满意计划相结合。

图3-5提供了这一系统的技术概述。将各种不同的实时数据资源输入到一个主数据库（称其为决策支持工厂）。对顾客满意度小组来说，该系统包括顾客调查、CRM系统、交互式语音应答系统和其他顾客服务系统。决策支持工厂中的数据从业务数据库中加载到数据集市和多维立方体。用户可以通过不同的方式访问数据库以满足他们不同的商业需求。

获得收益

最后，顾客满意度小组实现了10~12个直接与Expedia公司整体目标相关的目标。这些目标依次与顾客满意度小组的200多个KPI关联。KPI拥有者可以建立、管理、使用他们自己

的记分卡，管理者和经理对战略执行的情况如何心知肚明。记分卡同时向顾客满意度小组提供向下钻取数据的功能，用于发现潜在的发展趋势和形式。在过去，所有的这些需要几个星期甚至几个月的时间完成。在使用了记分卡之后，顾客满意度小组可以立即看到KPI方面的表现如何，这些指标依次反映小组和公司的目标。

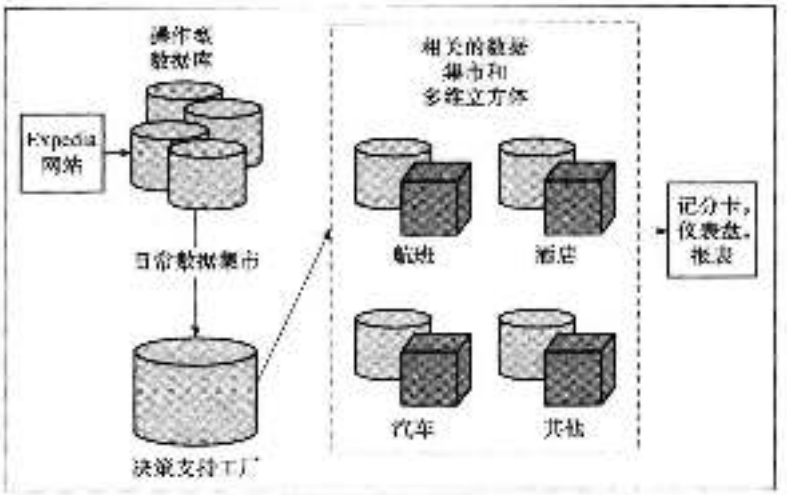


图3-5 Expedia的记分卡系统

作为额外的收益，系统中的数据不仅仅为顾客满意度小组提供支持，也支持企业中其他部门的工作。例如，基层管理人员可以进一步分析每个市场的飞机费用，以评价谈判合同的绩效，或者确定在单程运输中通过合并支出节省开支的可能性。旅行部门的经理可以利用商务智能发现哪个区域闲置或没有被订购最频繁，提出能够调整现象的战略，增加节约。

来源：Based on Microsoft, "Expedia: Scorecard Solution Helps Online Travel Company Measure the Road to Greatness," April 12, 2006, microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=49076 (accessed January 2010); R. Smith, "Expedia-5Team Blog: Technology," April 5, 2007, expedia-team5.blogspot.com (accessed January 2010).

3.6 节复习题

1. 什么是绩效评价系统？
2. 什么是KPI，它有什么特性？
3. KPI和运营指标有什么不同？
4. 仅仅依靠财务指标进行绩效评价的缺点是什么？
5. 什么是倾角原理？
6. "好"的绩效指标应该具有什么样的特征？

3.7 BPM方法

与简单的记分相比，还有很多绩效评价方法。有效的绩效考核评价系统应该做到以下几点：

- 上层的战略目标应与下层的举措相匹配
- 及时发现机会和存在的问题
- 决定优先权，并根据优先权分配资源
- 如果基本流程或战略改变，评价标准进行相应的调整

- 描述责任，明确实际绩效和责任的关系，奖励以及识别成就
- 如果系统中的数据显示有必要，就及时采取相应的措施改进流程和步骤
- 及时、可靠地进行计划和预测

整体或系统的绩效评价体系应该能够完成这些要求以及其他的一些要求。在过去的40年甚至更久的时间里，涌现出了很多不同的系统。其中，作业基准成本法（Activity-based Costing，ABC）或作业基准管理法是以财务为核心的。另一些，如全面质量管理则以流程为基础。在接下来的讨论中，我们注重介绍两种使用广泛、支持基本流程、强调BPM的方法：平衡记分卡（参见thepalladiumgroup.com）和六西格玛（参见Motorola.com/motorolauniversity.jsp）。

3.7.1 平衡记分卡

知名度最高、使用最广的绩效管理系统也许要属平衡记分卡法。1992年，Kaplan和Norton在他们发表在《Harvard Business Review》上的论文"The Balanced Scorecard：Measures That Drive Performance"中第一次提出了这一概念。几年后，1996年他们又出版了具有开创性的书——《The Balanced Scorecard：Translating Strategy into Action》。在这本书中，他们阐述了企业如何利用BSC，不仅仅提供财务指标和非财务指标，还能够通过沟通来实施他们的战略。经过过去几年的发展，BSC已经成为通用的名词（如同可乐和施乐），用于真实表示各种记分卡的应用和实施，无论是平衡的还是战略的。为了回应这一词的滥用现象，Kaplan和Norton，2000年又出了一本新书，《The Strategy-Focused Organization：How Balanced Scorecard Companies Thrive in the New Business Environment》。写这本书的目的是重新强调使用BSC方法要基于以战略为核心的环境。几年之后，在2004年，在《Strategy Maps：Converting Intangible Assets into Tangible Outcomes》一书中，阐述了将战略目标转化为操作层面战术和举措的具体流程。最后，他们在2008年最新出版的一本书，《The Execution Premium》，注重战略缺口——战略准则与只体运营实施的融合。

平衡记分卡的意义 从整体层面来看，平衡记分卡既是进行绩效评价的方法也是一种管理方法，这种管理思想能够帮助将企业的财务、顾客、内部流程及学习与成长的目标和任务转换为一系列的行动方案。作为一种管理思想，BSC的设计可以克服企业以财务为中心的系统局限。它通过将组织的愿景和战略转换为一系列的相关联的财务与非财务的目标、度量措施、目的和动机。非财务的目标分为3个部分：

- **顾客** 这一目标定义了组织如果想要实现自身的愿景应该以怎样的形象出现在顾客面前。
- **内部业务流程** 这一目标强调组织为了满足利益相关者和顾客的要求必须详细说明其流程。
- **学习和成长** 这一目标阐述了组织怎么做才能改变并提高其达到愿景的能力。

基本上，非财务的目标之间有这样的因果关系：通过"学习与成长"使"内部业务流程"改变，产生"顾客"成果以达到企业"财务"目标。这一因果关系的范例可以参见图3-6。

在BSC中，强调平衡这个词，因为一系列组合测量措施包括以下几种指标：

- 财务的和非财务的
- 处于领先的和落后的
- 外部的和内部的
- 数量的和质量的
- 短期的和长期的

调整战略和行动 作为一种战略管理方法，BSC使组织的行动能够与其总体目标相一致。BSC通过一系列相互关联的步骤完成这个目标。不同的书涉及的具体步骤不一样。在最新的译文中，Kaplan and Norton（2008）提出了包括6个步骤的流程：

	战略图：相关的项目	平衡记分卡：度量值和目标		战略方案：行动计划
财务	增加净收益	净收益增长率	提高25%	
顾客	提高客户保留	维持顾客的保留率	提高15%	修改授权许可和维修合同
业务流程	提高客户服务中心的绩效	讨论周转时间	提高30%	将客户服务中心的业务流程进行标准化处理
学习和成长	降低职工流动率	自发的员工流动率	提高25%	工资和奖金增加

图3-6 战略地图和平衡记分卡

1. **制定并阐述战略** 制定、阐明组织的使命、价值和愿景；通过战略分析，识别影响企业战略的内部和外部实力；明确组织的战略方向，详细说明组织如何实施战略。

2. **战略计划** 将战略的陈述转化为具体的目的、度量值、目标和举措，指南行动的预算，使组织与有效的战略实施保持一致。

3. **与组织保持一致** 确保企业的业务单元和辅助部门的战略和企业的总体战略保持一致，激发员工的积极性，实现企业的目标。

4. **实施计划** 确保战略的改变能够转化为实施过程、资源能力、实施计划、影响战略的预算和战略需要的改变。

5. **监督和学习** 通过正式运营审查会议，决定企业的短期财务和实施绩效是否和目标保持一致，通过战略审查会议来查看战略目标是否成功执行。

6. **测试和调整战略** 通过战略测试和调整会议确定战略是否有效、基本假设是否有效以及战略是否随时间变化而进行更改和调整。

表面上，这些步骤和图3-1中闭环的BPM系统很相像，这不奇怪，因为BSC方法是BPM方法中的一个。然而，BSC方法和其他方法的一个不同点在于，它应用了两个独一无二的新型工具——战略地图和平衡记分卡。

战略地图和平衡记分卡相互关联。战略图描述了通过一系列的因果关系来实现企业的增值过程，这些因果关系是BSC中的4个关键指标：财务、顾客、业务流程、学习和成长。平衡记分卡是对和各种目的相关的行动测量实现的和目标进行追踪。战略地图和BSC帮助企业对各个战略进行调整、传播和度量。

图3-6是一个虚构企业的战略地图和平衡记分卡的例子。其中还包含了帮助企业实现战略目标的方案组合。从图中可以看出，企业包含了贯穿4个BSC观点的7个目标。和其他战略地图一样，该图以财务目标开始（例如，增加净收入）。该目标受客户目标（例如，增加客户保留量）驱动，反过来，客户目标是企业内部目标的结果（例如，提高呼叫中心绩效）。沿该图直至地图底层，就找到了学习目标（例如，减少员工的流动）。

战略图上的每个目的都和一个度量值、目标和行动方案相关。如“增加客户保留量”的目标和“维持保留率”指标相关，该指标可能有一个每年增长15%的目标。其中方法之一就是可

以通过改变（简化）许可证和维持合同来实现该增长。

总之，图3-6中的战略地图是一个虚构的业务战略模型，当特定的名字（如人或者团队）被安排到各种方案时，该模型实现了低层次的行动与高层次的战略目标的一致性。将真实结果和战略目标进行对比，可以决定假设表现的战略是否值得怀疑，对与假设有关的各种行动是否需要调整。

图3-6是一个相对简单和明确的部分业务战略图，大多数的战略图是比较复杂的，而且覆盖一系列的目标。由于这些战略地图的复杂性，Kaplan和Norton最近提出了一个叫做"战略主题"的概念，"战略主题将一个战略分为几个独立的增值过程。"每一个战略主题代表一系列相关的战略目标。例如，图3-6中的战略目标可以表示为"客户管理"。如果图3-6的虚构企业试图通过捕获一个竞争对手来提升净收入，可能会有一个"兼并和获取"的主题。战略主题背后的思想是一个简化实施、执行、追踪和调整战略的过程。

3.7.2 六西格玛

六西格玛兴起于20世纪80年代中期，已经被世界上的许多公司所采用。最重要的是，它被作为一个绩效管理方法。然而，许多公司将它作为一个帮助企业核查实施流程、发现问题和找出解决办法的流程改进方法。近几年，有些公司，如Motorola已经意识到六西格玛在战略中的作用。在这些实例中，六西格玛提供了测量和监控与公司盈利相关的关键流程的手段，并用于提升企业的总体效益。由于关注于业务流程，六西格玛也提供了一种在识别或发现问题之后，处理绩效问题的直接方法。

六西格玛定义 许多六西格玛的思想在早期的质量措施里出现过，但是六西格玛的历史要追溯到20世纪70年代（参考en.wikipedia.org/wiki/Six_Sigma）。六西格玛是由Motorola的工程师Bill Smith创建的。实际上，六西格玛是Motorola的一个联邦注册商标，在20世纪70年代末和20世纪80年代早中期，Motorola迫于内外界环境的压力下实施了六西格玛。从外部来说，被一些提供质高价廉产品的竞争者打败。从内部来说，一个日本的公司接管了美国Motorola生产Quasar电视机的工厂，在正常操作程序下生产的电视机的不合格率是5%，Motorola管理者不得不承认他们的质量不行。为了应对这些压力，Motorola的CEO，Bob Galvin领导公司走六西格玛的质量路线。从此，世界上包括General Electric、Allied Signal、DuPont、Ford、Merrill Lynch、Caterpillar和Toshiba在内的数百个企业开始利用六西格玛，使最高增长达数十亿美元，并提高了最低收入。

表3-1 商业流程的种类

会计和测量
行政和战略管理
审计和改进
商业计划和执行
商业政策和程序
全球市场营销和销售
信息管理和分析
领导和盈利能力
学习和创新
维持和合作
合伙和联盟
生产和服务
购买和供应链管理
招聘和发展
研究和发展

在六西格玛理论中，一个商务活动被看做是各种业务流程的集合，一个业务流程是一系列活动的集合，活动将包括供应商、资产、资源（如资产、物料和员工）和信息等的一组输入转化为提供给其他的人或流程的一组输出（如产品或服务）。表3-1列出了一些商务处理过程对企业总体效益的影响。

σ（西格玛）是一个希腊字母，统计学家用它来评价一个过程的变化范围，在质量管理中，变化性和不合格数表示同样的意思。一般情况下，公司在商业活动中存在很大的可变性。在数量上，每100万个机会有6 200～6 700次缺陷是正常的（DPMO）。例如，一个保险公司每处理100万个索赔事件中，会有6 200～6 700个是不能令人满意的（如处理不当、表格错误）。这种可变性对应3～4个西格玛值。为了达到六西格玛，企业要将缺陷减少到少于3.4DPMO以下。因此，六西格

玛是一个用于减少商业流程中缺陷率并尽量将 DPMO 减小到 0 的一个绩效管理方法。

DMAIC 绩效模型 六西格玛依赖于一个简单的绩效改进模型 DMAIC。就像 BPM，DMAIC 是一个闭环的商务绩效改进模型，包括定义、度量、分析、改进和控制一个流程，步骤如下：

1. **定义** 定义改进活动的目标、目的和范围。在最高层次目标是企业的战略目标，在较低层次——部门或项目级别——是指每个操作过程的目标。

2. **度量** 对存在的系统进行度量，确定能够产生有效统计数据的度量数值，数据用于对前一步定义的目标进行监控。

3. **分析** 对系统进行分析，消除目前系统或流程绩效与目标的差距。

4. **改进** 初始方案通过找出更好、更便宜、更快的方法来减小差距，用项目管理和其他的计划工具来实现新方法。

5. **控制** 通过修改补偿和激励系统、政策、步骤、制定资源计划、预算、操作指导或者其他的管理系统使系统改进制度化。

对于新的商业过程，使用的模式称为定义、度量、分析、设计和校验（Define Measure Analyze，Design、Verify，DMADV）。传统意义上，DMAIC 和 DMADV 主要用于解决操作性问题。但是，将它们应用于企业的战略问题上，如企业利润，毫无问题。

精益六西格玛 近几年，人们开始关注将六西格玛方法和精益生产一起使用，精益生产或者简称精益（该方法概要可参考 en. wikipedia. org/wiki/Lean_manufacturing）。精益的早期概念可以追溯到 Henry Ford 的工作流批量生产。最近，精益生产的概念和 Toyota（Toyota 生产系统）生产过程相关联。精益生产一词是由 John Krafcik 于 1988 年发表在《Sloan Management Review》上名为"Triumph of the Lean Production System"的文章提出的（Krafcik，1988），也是基于他在麻省理工大学斯隆管理学院的硕士毕业论文。在麻省理工学院工作之前，Krafcik 是丰田和通用汽车合作项目中的质量工程师。

六西格玛和精益生产均用于质量管理，表 3-2 对两者进行了比较。

表 3-2 精益生产和六西格玛对比

功能	精益生产	六西格玛
目的	减少浪费	减少变化
关注点	集中于工作流	集中于发现问题
方法	许多小的提高	消除问题根源
绩效度量	减小工作流时间	统一输出
结果	减少浪费，增加效率	减少变化，统一输出

就像表 3-2 列出的，精益生产的核心在于减少浪费和非增值的活动，然而六西格玛的核心在于减少变化，使输出一致。Six Sigma Institute，从精益的观点来看，浪费（或认识浪费）来自于各种可变的因素（Six Sigma Institute，2009）：

- 生产多于需求
- 等待下一个处理过程的信息
- 不必要的材料运输
- 过量而非增值过程
- 存货量大于最低限度
- 员工多余的操作
- 不合格零件生产

精益生产可以应用于任何产品和工作流，不仅仅是生产活动中。目的是对工作流进行检查，消除浪费。下面是处理顾客要求或电话投诉过程中出现的一些不必要的浪费：

- 过量生产——向每个人发布所有信息
- 等待——等待信息的人
- 传输——将呼叫转给许多操作员
- 处理——过多地批准发布信息

- 存货清单——访客有待回答
- 举动——检索印刷说明书
- 过失——向顾客提供错误的信息

精益给六西格玛带来的是速度提升，精益生产通过消除非增值的步骤来实现加速（Poppendieck，2009）。一旦组成流程的都是有价值的步骤，六西格玛能够确保这些步骤的前后一致性。例如，上面所举的客服中心例子，一旦为取回打印手册制定了合理的步骤，那么下一步就要决定如何严格按照该步骤实施。

六西格玛的回报 六西格玛理论的专家和学者对该方法提出好评并用通用电气（General Electric，GE）和Honeywell例子证明了六西格玛方法的优势。GE的前CEO，Jack Welch，从1995年开始利用六西格玛方法，并公开表明"六西格玛帮助企业2000年的营业利润率从4年前的14.8%达到了18.9%"。最近，Caterpillar Inc（2009）指出利用六西格玛方法节约了30亿美元。其他学者指出，Home Depot的事例说明了利用六西格玛方法有可能导致失败（Richardson，2007）。广为报道的Home Depot采用六西格玛，源于它的前CEO，来自于GE的Robert Nardelli。从此，Home Depot业绩开始下降，市场地位被主要竞争对手Lowes占领。Nardell离开公司后指出六西格玛并没有像自己承诺的那样奏效。六西格玛方法的反对者同样认为，六西格玛在提高生产效率上能发挥很好的作用（Hindo，2007）。对于通过革新来提高业绩的企业，六西格玛不能起到很好的作用。Honeywell的代言人对六西格玛的争议提出了一个更加适度的观点，"六西格玛不是终极目的，它只是一个处理工具，我们不能把企业的业绩仅仅和使用的工具相关联"（Richardson，2007）。

六西格玛和其他的商业方案没有什么区别。你可以制订计划，并设出评价指标评估执行过程，如果没能按照自己希望的方式部署，那么可以对此做出调整。下面的措施能够显著提高六西格玛的成功（Wurtzel，2008）：

- **六西格玛和企业战略相集成** 六西格玛在减少过程的变化上起到很大的作用。如今，越来越多的企业开始实施六西格玛方法，并将它作为企业战略的一部分。
- **六西格玛帮助企业实现目标** 企业取得好的业绩需要依靠六西格玛方法解决面临的主要商业挑战或风险。识别企业面临的挑战意味着，所有的企业领导者都明白为什么要将六西格玛作为企业制定战略的准则。
- **关键管理人员的参与** 一个企业的所有业务管理者必须帮助企业设计六西格玛的部署，如果管理者认为六西格玛只是占用了企业的资源，而没有增加企业的能力并帮助他们成功实现目标，或者认为六西格玛减少了企业的预算分配而没有带来财务上的回报，那么他们就不会支持六西格玛的实施。
- **基于潜在价值项目的选择** 一些成功实施六西格玛的企业通过评估该项目能够给利益相关者带来多少价值，严谨地选择所要实施的项目，这是一个通过比较创造价值和付出成本来权衡决策的过程。
- **大量的项目和资源** 一些企业在实施新项目时，对大量的人员进行培训，但是在项目论证上投入较少。另外一些企业加大当前企业级项目的开发，在6个月时间里培训大量的"黑带"并启动几十个项目。这里，黑带是指经过六西格玛培训，并投入100%的时间用于执行六西格玛计划的员工。每种方法都是可行的，但是对每个企业来讲，都存在一个六西格玛投入上的临界点。
- **积极管理进行中的项目** 假设大多数的企业想在6个月或1年内达到可评价的重大结果，那么就要尽可能地将更多的项目投入到精益六西格玛中。最好是能够选择少量的有潜力

的项目，而不是大量的不重要的项目。将正确的资源放置到正确的项目中，才能在短期内获得最大的学习和结果。

强调小组领导技巧 六西格玛的应用需要加入一些技术性的技巧，如处理和分析数据的能力。但是，优秀的领导技巧更为重要，这表明企业要考虑如何选择合适人选作为黑带的角色。刚开始，将企业很多有前途的人员作为黑带人选是困难的，但是这样能够很快取得成效并快速改变企业的组织状况。

严格追踪结果 六西格玛的实施应该是"量入为出"，并且结果能够被客观证实。许多企业没有完全可靠的方法来判断项目的结果和影响，或者低估了实施中的困难。一个项目应当都是计划好的，企业必须考虑潜在的影响财务结果的各种评价办法和关键绩效指标。项目的运行周期及项目的价值都应当作为基本的评价指标，并为这些指标制定一个可接受的变动范围。

为了提高实施六西格玛方案成功的可能性，一些企业，如 Motorola 和 Duke University Hospital 将六西格玛方案和企业的 BSC 方案一起实施。这样，它们的质量方案就和企业的战略目标相关联。同时，（Gupta，2006）制定了一个叫做六西格玛业务记分卡的混合方法，该方法将六西格玛的提升处理过程和 BSC 的财务指标相互结合。技术前沿 3.1 介绍了这种结合的优点和结构。

技术前沿 3.1 BSC 遭遇六西格玛

2006 年 Praveen Gupta 在一本名为《Six Sigma Business Scorecard》的书中提到了平衡记分卡和六西格玛方法的区别，见表 3-3 所示。简而言之，BSC 的重点在于优化战略结果，而六西格玛在于优化流程。

由于存在着这些不同点，所以许多企业都分开实施 BSC 和六西格玛方案。然而，自波士顿 Aberdeen 集团的已退休董总裁 Stan Elhuaro 指出，BSC 和六西格玛是相互补充的。（Leahy，2005）文献指出，如果两者不相互结合，那么它们中任何一个的优势都不能发挥出来。BSC 方法帮助企业迅速精确地认识到关键绩效中的不足，并为企业的发展提供机会。但是，BSC 不能帮助企业改进绩效问题。相比之下，六西格玛项目处于挣扎处境，因为项目团队"将整个组织主要专注于找出绩效的缺陷或者将关注点放在提升企业的边际收益方面"（Leahy，2005）。这两种方法之所以是相互补充的，是因为 BSC 为提高绩效方案提供了战略内容，六西格玛可以找出绩效不足之处的基本原因并减小目标和现实之间的差距。

不久前，一个针对采用 BSC 或六西格玛项目的企业调查（Docherty，2005）表明，采用这些项目的企业中几乎有一半在前 3 年无法实现收支平衡，但是那些运作得好的企业在财务上获得了很大的财务利益，那些获得最大纯收益的是能够将 BSC 和六西格玛两种方法集成起来的企业。通过以下过程可以实现两者的集成：

- **将企业的战略转化为可计量的目标** 这可以通过制定战略地图和利用相关度量值的记分卡来实现。
- **通过组织关系将各个目标串在一起** 利用六西格玛分析企业内部各种因果关系，将企业级的目标分解为较低层次的操作型目标。
- **制定基于顾客需求的目标** 通过将 BSC 和六西格玛方法结合起来，确保操作型目标能够直接影响顾客的期望。
- **利用六西格玛方法实施战略项目** 利用六西格玛方法，驱动产品和流程质量的提升。
- **始终用一种形式来实现商业目标** 以流程的观点看待组织活动。六西格玛用户控制过程中的变化，BSC 包含了流程评价指标。

成功将两者合并起来的企业者总，他们不明白为什么一些企业只采用其中的一种办法。同时也表明这需要花费 1 年的时间来组织员工培训并克服存在的文化和组织障碍。

来源：Compiled from p. Gupta, *Six Sigma Business Scorecard*, 2nd ed., McGraw-Hill Professional, New York, 2006; P. Docherty, "From Six Sigma to Strategy Execution," 2005, www.solutionsglobal.com/secure/FromSixSigmaToStrategy_AACBC.pdf (accessed January 2010); and T. Leahy, "The One-Two Performance Punch," *Business Finance*, February 2005, businessfinancemag.com/magazine/archives/article.html?articleID=14364 (accessed January 2010).

表3-3 平衡记分卡和六西格玛之间的比较

平衡记分卡	六西格玛
战略管理系统	绩效评价系统
和关键商业活动相关	提供绩效的即时信息，确定驱动因素的绩效
用于制定平衡的一套指标	用于确定度量值对利润影响的一套指标
确定影响质量和价值的评价办法	确定健康和长期能力的领导责任
实现明确质量/战略、沟通、计划、制定目标、策略一致性方案和提高反应速度等关键管理过程	包括所有的商业过程——管理和运作
无须明确地定义领导角色，平衡顾客和企业内部操作	平衡管理者和员工角色关系；平衡重要流程的成本和收益关系
强调每个度量目标	强调每个度量及不相关目标的强劲的增长率
核心是增长	核心是利润最大化
充满战略内容	充满对利益的执行情况
组成度量值的管理系统	基于流程管理的管理系统

资料来源：P. Gupta, Six Sigma Business Scorecard, 2nd ed., McGraw-Hill Professional, New York, 2006.

3.7 节复习题

1. 有效的绩效管理系统的特征有哪些？
2. BSC 的 4 个观点是什么？
3. BSC 中的“平衡”指的是什么？
4. BSC 是如何和企业的战略和实施方案保持一致的？
5. 什么是战略地图？
6. 什么是战略主题？
7. 什么是六西格玛？
8. DMAIC 模型有哪些基本过程？
9. 比较精益生产和六西格玛。
10. 六西格玛成功实施的方式有哪些？
11. 比较 BSC 和六西格玛。
12. BSC 和六西格玛是如何结合在一起的？

3.8 BPM 技术和应用

本章开始介绍了 BPM 的概念，BPM 包含的企业运营的流程、方法、度量和技术，对企业的绩效进行度量、监控和管理。3.3～3.8 节研究流程、度量和方法。本节介绍剩下的因素——技术和应用。

3.8.1 BPM 架构

术语系统架构包括系统逻辑设计和物理设计。逻辑设计包括系统的组成元素和各元素之间的交互关系。物理设计是指系统逻辑设计的实现方式以及通过一些专业技术对系统进行部署，如网页浏览器、应用服务器、通信协议、数据库等。从物理设计来看，任何特殊的绩效管理方案和实施都比较复杂，而逻辑设计比较简单。从逻辑上看，一个绩效管理系统由以下 3 个部分或层次组成（如图 3-7 所示）：

- **绩效管理应用** 该层通过变换用户交互和源数据，形成企业的预算、计划、预测、报表、分析等信

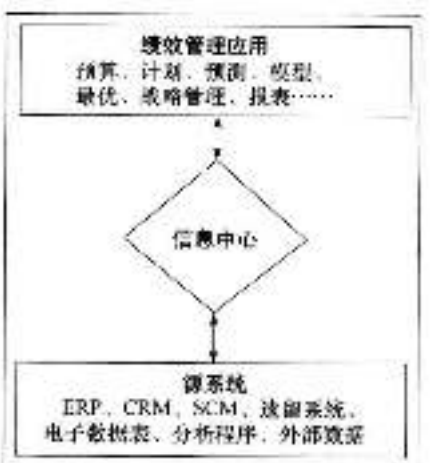


图 3-7 绩效管理逻辑系统架构

息，支持企业BPM。这种特殊应用使得不同组织之间实施一个又一个BPM，这主要取决于他们的需求和战略目标。任何BPM方案应当是足够灵活和可扩展的，从而使组织找到适合自己的路线，包括选择何种软件以及如何运行该软件。特别要说的是，实际上，有些BPM应用程序经常被使用。

- **信息中心** 大多BPM系统需要从各种不同的源系统（如ERP或CRM系统）中提取数据和信息，而且获取数据和信息的方式多种多样。然而，一个设计良好的BPM系统通常将这些数据进行集中映射和存储，一般存储在数据仓库或数据集市中。
- **源系统** 该层包含了向BPM信息库中提供信息的所有数据源。对大多数企业来说，该层包含了财务和企业其他系统的运营数据。一个完整的解决方案还可以任意访问企业的外部信息，如行业趋势、竞争对手情报，从而对企业的绩效进行更深入地分析。BPM系统很少直接访问这些源数据，一般要对数据进行提取、转换和加载，企业利用集成系统或网络服务，将这些数据转换和连接到信息库中。

BPM应用 BPM是一个包含了各种应用的闭环系统，涵盖了从战略规划到运营计划，从预算到监控、到调整再到实施的内容。尽管BPM包含的流程范围很广，但Gartner Group公司的行业分析家将BPM的主要流程分为以下几个方面（Chandler et al.，2009）：

1. 战略管理 战略管理应用为企业的战略制定、建模和监控提供了一套方案，从而提高企业的绩效、促进企业决策的制定和合作。这些方案通常与企业的战略地图或方法（如平衡记分卡）相关。战略管理包含了以下能力：

- 在场景模型中利用“基本情况”或“主动型”的方法对高层次的商业计划进行制定和评估。
- 在目标管理中，应用项目管理工具使于负责的管理人员实施战略中的任务。
- 记分卡和战略地图记录企业的战略、目标和各项任务，评估绩效，并提供有效的、面向企业沟通和合作环境。
- 仪表盘（或驾驶舱）集中表现了各种显示度量指标和关键绩效指标，通过对这些指标进行观察，从而利用BI工具对这些指标进行深入分析。

BPM套件至少能够提供仪表盘功能，以用户容易理解的方式体现出企业的绩效现状。一些较复杂组织将BPM的记分卡软件与其他BPM软件相关联，从而实施企业战略地图。因此，战略管理成为BPM中一个越来越重要的功能。

2. 预算、计划和预测 这些功能可帮助企业进行预算、计划和预测。它包括了短期的财务预算、长期计划和高层战略规划。这些功能以工作流的形式进行预算/计划的制订、提交和核实，而且还具有动态预测和制订方案的功能。它们也应该支持连接操作规划和财务预算的企业规划模型。另外，它们也能够与特殊领域的应用，如供应链规划，实现数据共享。

3. 财务合并 该功能使组织将不同会计准则和联邦法规下的财务数据进行统一、合并、简化和聚集。这些应用都是BPM的基础，因为它们需要审计，并与其他BPM应用分享的企业级财务信息。

4. 收益模型和最优化 该功能包含了作业成本法（Activity Based Costing，ABC）。该方法对组成总成本的每个作业成本进行高层次的决定和分配。作业管理帮助用户模拟出不同成本对收益的影响和资源分配战略有些应用程序除了具备传统的ABC所具有的功能外，还从包装成本模型、捆绑销售、定价和销售渠道战略来分配收入。

5. 财务、法定和管理报告 BPM应用程序需要专业的报表工具将财务状况按标准的格式输出。这些输出要符合公认会计准则（Generally Accepted Accounting Principle，GAAP）中的规则，如美国的公认会计准则或国际财务报表标准。同时，应用程序中还包含了可视化的技术，这些可视化技术可应用于从预算到目标等不同方面的分析，如双曲线树。

3.8.2 商业BPM套件

BPM软件厂商提供的套件至少可以实现3个核心的功能（如：预算、计划和预测；收益模型和最优化；记分卡、财务合并以及法定财务报表）。根据Gartner（Chandler et al.，2009）的估计，2007年BPM软件的许可证费用和维护费几乎达到了18亿美元，比2006年增加了19%。相比之下，国际数据公司（International Data Corporation，IDC）表明，2007年BPM软件市场交易额大约为20亿美元，预计2012年将达到32亿美元。（Vesselle and McDonough，2008）如此来看，每年的增长率超过了10%。

这种增长的主要原因是更强大的分析系统代替了电子表格。除了部分行业部门以外，BPM几乎和所有的组织相关，因为所有的组织都向财务总监和财务小组提供业务分析（如，盈利分析和财务计划编数）和管理信息（如财务管理报表、预算和法定报告）的信息，从而将这些管理信息传递给领导小组，这也是BPM的主要功能。

在过去的3～4年间，BPM领域的主要变化是BPM软件厂商的合并。过去的几年中，BPM软件市场被一些单一业务厂商（如Hyperion、Cognos和SAS）所占有。这是在Oracle兼并Hyperion，IBM兼并Cognos，以及SAP兼并Business Object之前。现如今，该市场领域被一些大的厂商所占有，如Oracle Hyperion、IBM Cognos、SAP Business Object，以及Infor和SAS，这些厂商占领了BPM市场70%的份额。

如同对它所关注的许多软件市场一样，Gartner为CPM套件供应商创立了一个魔力象限（Chandler et al.，2009）。象限依据各供应商的执行能力和视角的全面性进行定位，从这两个角度进行组合便得到了4种类型的供应商（见表3-4）。根据Gartner的这种方法，Oracle Hyperion、SAP Business Object及IBM Cognos都处于领导地位。由此可知，这些大供应商引领着BPM市场。

表3-4 Gartner的魔力象限

	视角	
执行	有限	强大
强大	挑战者	领导者
有限	观望者	深耕者

能被Gartner魔力象限所认同的BPM套件至少包含了3个BPM基本应用，也就是说许多供应商提供的软件只包含少量的功能。表3-5列出了在象限中处于领导地位的供应商所提供的BPM套件功能。

表3-5 SAP、Oracle和IBM绩效管理软件的功能

BPM应用	SAP Business Object 企业绩效管理	Oracle Hyperion 绩效管理	IBM Cognos BI和财务绩效管理
战略管理	战略管理	战略财务、绩效记分卡、计划	BI记分卡、BI分析、计划
预算、计划和预测	商业计划和合并	计划	计划
财务合并	财务合并、公司间协作	财务管理	控制
收益模型和最优化	收益和成本管理	收益和成本管理	
财务、法定和管理报表	业务对象、可扩展商业报告语言发布	绩效记分卡	BI报表、BI记分卡、BI仪表盘
其他绩效管理应用	绩效管理费用、供应链绩效管理	主要财产计划、工作人员计划、集成经营计划	
数据管理应用	财务信息管理	财务数据质量管理、数据关系管理	决策

来源：Compiled from sap.com/solutions/sapbusinessobjects/large/enterprise-performance-management/index.epx（accessed January 2010）；oracle.com/appserver/businessintelligence/hyperion-financial-performance-management/hyperion-financial-performance-management.html（accessed January 2010）；ibm.com/software/data/cognos（accessed January 2010）。

3.8.3 BPM 市场与 BI 平台市场对比

除了 BPM 市场以外，Gartner 同样密切关注 BI 平台市场。根据 Gartner 的观点，BI 平台不是一个简单的 BI 观望者，而是包含了各种综合的功能。按照 Gartner，BI 包含了以下功能（McKay，2009）：

- BI 基础架构
- 元数据管理
- BI 应用开发
- 工作流和协同管理
- 报表
- 仪表盘
- 查询
- Microsoft Office 集成
- OLAP
- 高级视图
- 预测模型和数据挖掘
- 记分卡

和 BPM 市场相比，BI 平台市场的规模更大。从 2009 年绩效管理市场规模的各种分析数据可知，BPM 交易额为 20 亿美元到 35 亿美元之间，每年的增长率至少为 25%。相比之下，BI 软件市场 2007 年超过了 50 亿美元，每年的增长率超过 10%。

然而，所有处于领导地位的 BPM 软件供应商也提供 BI 平台，BI 平台市场比 BPM 市场的差异性更大。2009 年，Gartner 的魔力象限中的 BI 平台“领导者”不仅仅包含了 BPM 软件的领导者（IBM、Oracle 和 SAP）还包含了 Information Builders、Microsoft、SAS 和 MicroStrategy。这一象限中所有供应商的 BI 平台都有功能强的产品，这些产品在使用和分析能力上稍有差异。在这些领军产品中，MicroStrategy 和 Teradata 大学（teradatastudentnetwork.com）合作，使大学生在教育和研究中应用他们提供的产品，并因此出名。

3.8 节复习题

1. 什么是逻辑系统架构？
2. BPM 架构的 3 个关键组成部分是什么？
3. 描述 BPM 应用的主要种类。
4. 在过去的 3 年到 4 年里，BPM 市场发生了什么变化？
5. Gartner 的魔力象限有哪几个基本类别？哪些厂商是 BPM 市场的领导者？
6. 什么是 BI 平台？在 Gartner 的 BI 魔力象限中，哪些厂商是市场领导者？

3.9 绩效仪表盘和记分卡

记分卡和仪表盘不是全部也是绝大多数软件中应包括的组件，如绩效管理系统、绩效评价系统、BPM 套件和 BI 平台。仪表盘和记分卡两者都将重要信息集中可视化显示在一个独立的界面上，因此，通过简单的浏览就可以理解这些信息。图 3-8 是一个典型的仪表盘示例。该仪表盘上列出了一个虚拟软件公司的关键绩效指标，该公司为软件开发人员提供了专业化的图表和可视化的组件。该公司通过网页和在网站发布横幅广告的形式来增加网站主页上的访问量。从仪表盘上可以看出，通过“The Code House”网站发布的横幅广告去访问该公司所占的比例最大，而且“The Code House”网站的点击率最大（也就是说，该软件公司的横幅广告

在"The Code House"网站上每出现100次，就会有稍微多于2个访问者点击该广告）。大体上，仪表盘的网站横幅广告上显示总共有超过2.05亿次点击，有220万访问者进入公司主页，其中有120万访问者进入产品介绍界面，并且最终有100万访问者下载了该公司产品。该统计数据表明了，截止到此时"通过访问网页广告的用户"和"下载用户"呈现上升的趋势，超过了企业的目标值（如，在阴影区以上），每点击一次的成本为80美分。这个有特色的仪表盘能够使最终用户看到不同的横幅广告统计，和按时间段或产品进行的度量（图3-8中右上方的下降）。

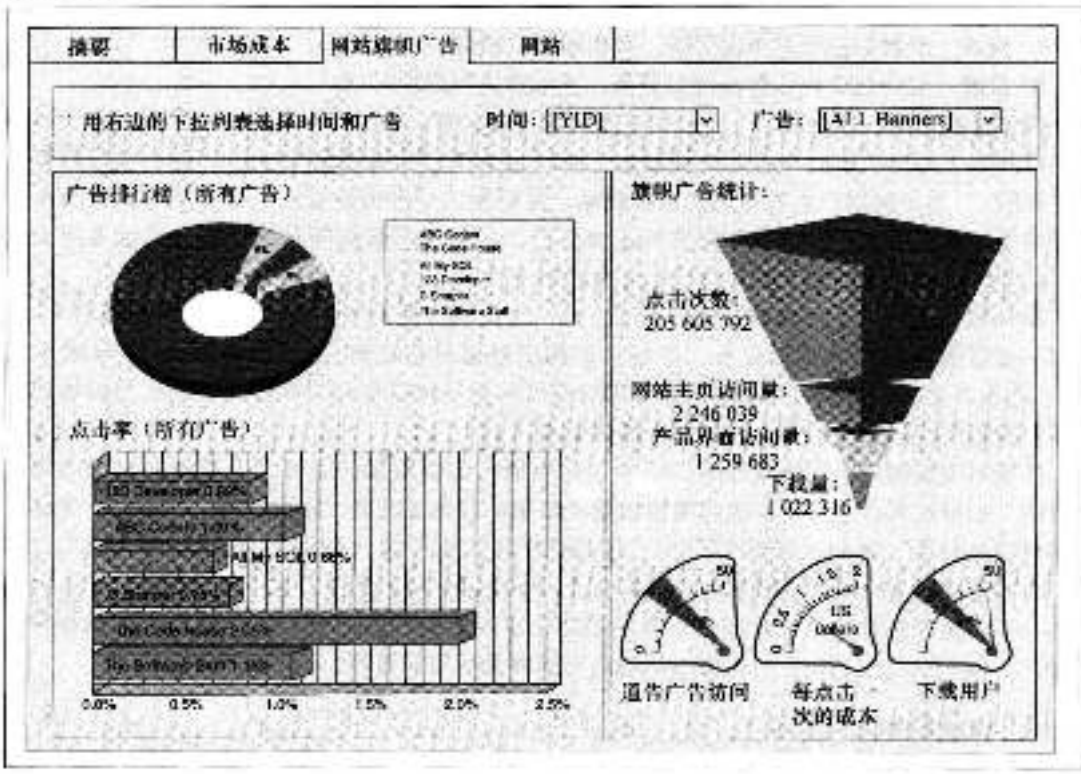


图3-8 仪表盘示例

来源：Dundas 数据可视化公司，dundas.com/Gallery/Flash/Dashboards/Index.aspx(accessed January 2009).

3.9.1 仪表盘和记分卡

在商业刊物中，仪表盘和记分卡是可以互相替代的，尽管如此，可以从表3-4中看出，各个BPM厂商均提供独立的仪表盘和记分卡应用程序。虽然仪表盘和记分卡有许多相同之处，但两者也存在差异。一方面，经理、管理者和员工用记分卡制订战略计划，从而成功实施战略方向和目标。最典型的应用就是平衡记分卡。另一方面，仪表盘应用于实施和作业层次上。管理者、经理和作业人员用作业仪表盘工具管理每周、每日甚至是每小时的细节性的作业绩效。例如，作业仪表盘可用于监控产品质量。同时，管理人员和员工用战术仪表盘来监控战术实施。例如，战略仪表盘可用于市场活动或营销渠道战略的制定。

3.9.2 仪表盘设计

仪表盘已经不是一个新的概念，它的根源至少可追溯到20世纪80年代的高级管理人员信息系

统。现如今，仪表盘的使用很普遍。例如，几年前弗雷斯特研究公司（Forrester Research）估算，占世界上40%的2 000个大型企业中使用该技术，仪表盘开发网站（dashboardspy.com/about）提供了更多关于广泛使用它们的证据（Ante and McGregor，2006）。该网站对各种规模的企业，各行业、公益单位和政府机构等拥有的成千上万个BI仪表盘、记分卡和BI接口进行了描述和截图。本章末的案例详细介绍了该网站上最新的一个仪表盘组件——纽约城市绩效报告系统。

根据BI领域，特别是仪表盘领域的专家Eckerson（2006）的观点，仪表盘最典型的特征是它的3层信息结构：

1. 监管　利用图表和摘要资料进行核心绩效监管。
2. 分析　对数据进行多角度分析，找出问题的根源。
3. 管理　对详细的作业数据进行分析，找出解决问题的行动。

因为这些层次，仪表盘的一个界面中要包含许多信息，根据Few（2005），"仪表盘设计最根本的目的是在一个简单的界面上能够清楚、明了地展示出所需要的信息，而且这些信息容易被用户理解。"为了使用户更容易理解这些数据，需要将这些数据放置在一个背景中进行分析。可以通过将一些基础数据或目标数据进行比较分析，判断这些数据质量的优劣，预示发展趋势的好坏，或者用专业化的分析工具进行比较和评估背景。

BPM软件中包含了一些典型的对比项，如：历史值、预测值、目标值、标准值或者平均值，对同一度量值进行实例应用分析，并与其他的度量值进行比较分析（例如，收入与成本的比较）。在图3-8中，各种关键绩效指标分别与该指标的目标值进行对比分析，收入与市场成本对比，各阶段的销售数据与另一阶段的销售数据进行对比。

尽管对度量值进行了比较分析，但专业化地指出数据质量的好坏以及数据的发展趋势是很重要的。如果没有这些评价，决定数据质量或结果的好坏要花费大量的时间。特别地，无论是专业的可视化对象（例如，交通信号灯），还是可视化的属性值（例如，黄色标志），都可以认为是可以评价的内容。再次回到图3-8的仪表盘，颜色编码用于指定关键绩效指标的好差，绿色箭头表示不同阶段销售额是上升还是下降的趋势，以及这种趋势的好坏。尽管该例中没有使用其他颜色，例如红色或橘色，但也可以体现出不同度量对象的属性。

3.9.3　仪表盘展示的内容

尽管绩效仪表盘和平衡绩效记分卡不同，但是两者也有相同之处。首先，它们均适用于大型的BPM或绩效评价系统中。这说明它们是BI的一个架构或者可以作为大系统的绩效管理架构。其次，所有的仪表盘和记分卡均有以下功能：

- 使用可视化组件（如图表、柱状条、折线、仪表、计量器、停止信号灯）突出数据或需要处理的异常，使人一目了然。
- 对用户来说简明易懂，这就意味着即使不需要大量培训也能够容易使用。
- 将不同系统的数据整合为一个能够反映商业信息、独立的、概括性的图形界面。
- 能够实现对数据进行挖掘或钻取，发现潜在的数据资源或报表，从而为用户提供更多的、潜在的比较和评价内容的细节。
- 它们呈现了动态的、真实世界的、及时的数据更新，使终端用户时刻能够看到商业数据的最近变化情况。
- 它们几乎不需要用户编码就可以实施、部署和维护。

3.9.4　数据可视化

数据可视化定义为"对数据探索、数据理解和数据交流的可视化描述"（Few，2008）。这与

信息图表、信息可视化以及统计图表相关。最近，应用于 BPM 和 BI 中的数据可视化包含了图表、图像以及一些可应用于记分卡和仪表盘中的其他可视化元素（例如，停止信号灯和测量仪器）。Seth Grimes（2009）指出，数据可视化技术和工具的应用呈现上升的趋势，这使得 BPM 和 BI 系统的用户能更好地"交流联系，增加历史背景，发掘隐藏的业务关系，并且能够通过讲述有说服性的故事来澄清和号召行动。"

在 BPM 和 BI 应用中，可视化面临的主要难题是对复杂的大型数据集进行直观地、多维度、多种方式地分析。这些应用中的大部分图表、图像和其他可视化工具通常包含了 2 个维度，有时是 3 个，以及一些小型的数据集。这些系统的数据存储在一个数据仓库中。这些数据仓库至少包含了：一些维度（例如，产品、地址、组织结构、时间）、度量值以及成千上万个数据单元。为了解决这些问题，许多研究人员开发了一些新的数据可视化技术。

其中的一些新技术对图表和图像进行扩展，例如，Gapminder 网站（gapminder.org）上的气泡图，提供了对世界健康和人口数据进行多维分析的功能。图 3-9 描述了该网站上展示的数据种类（这是临摹图，不是网站上的真实图）。表面上，这些特殊数据表现了世界上不同国家的生活期望和人口出生率之间的关系。每个气泡代表一个国家，气泡的尺寸代表国家的人口规模，每个颜色表示这些国家所在的洲。图 3-9 下方的菜单可供用户选择分析的年份，点击按钮可以显示不同年份时的变化。

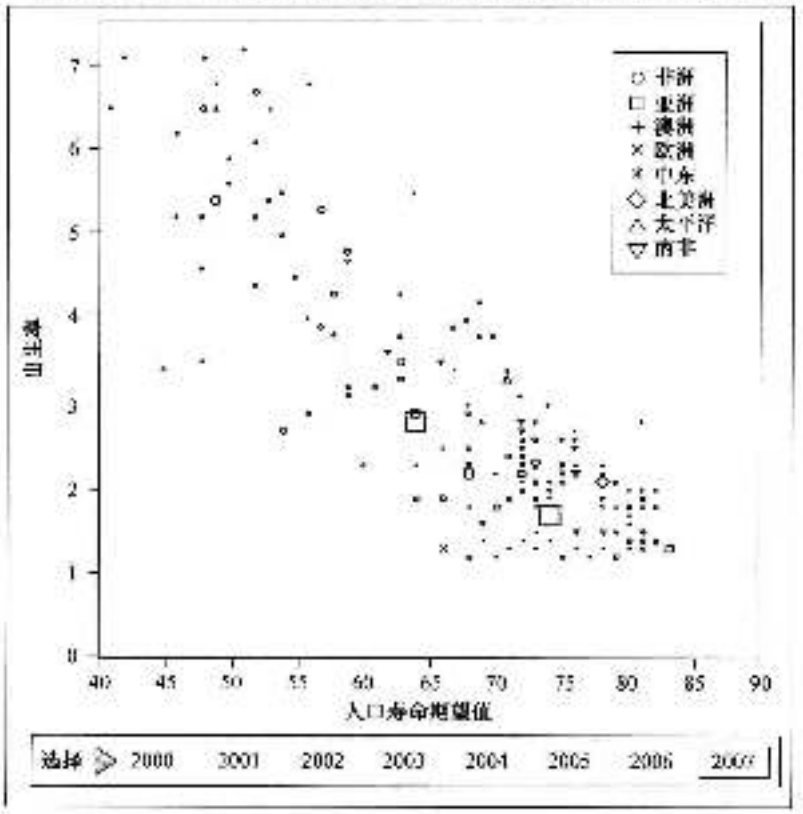


图 3-9 人口数据的气泡图

其他一些技术包含了较新的可视化格式，如，美国马里兰州大学的 Ben Shneiderman 提出的树图。树图按嵌套的矩形格式将数据分层（见图 3-10）。因此，这非常适合于展示数据仓库中包含多个维度的数据。每层的一个维度可以用一个矩形框表示，该框和一个更小的矩形框相互关联，这个小的矩形框是父数据的一个子数据（子分支）。每个矩形框和一个特定的度量值（通常

是求和）相对应。例如，一个树图可供一个软件公司分析产品说明书中的缺陷。在顶部，矩形框展示了不同种类产品的问题数量。在每个矩形框中，小的平铺的矩形框代表了这些种类中单独的一个产品出现的所有问题。另外，用矩形框的颜色表示正在使用的产品说明书在市场上的时间长度。总之，树图的方式能帮助用户发现其他技术所不能发现的类型，而且，由于树图对空间的有效使用，使其一次能够展现成千上万个条目。

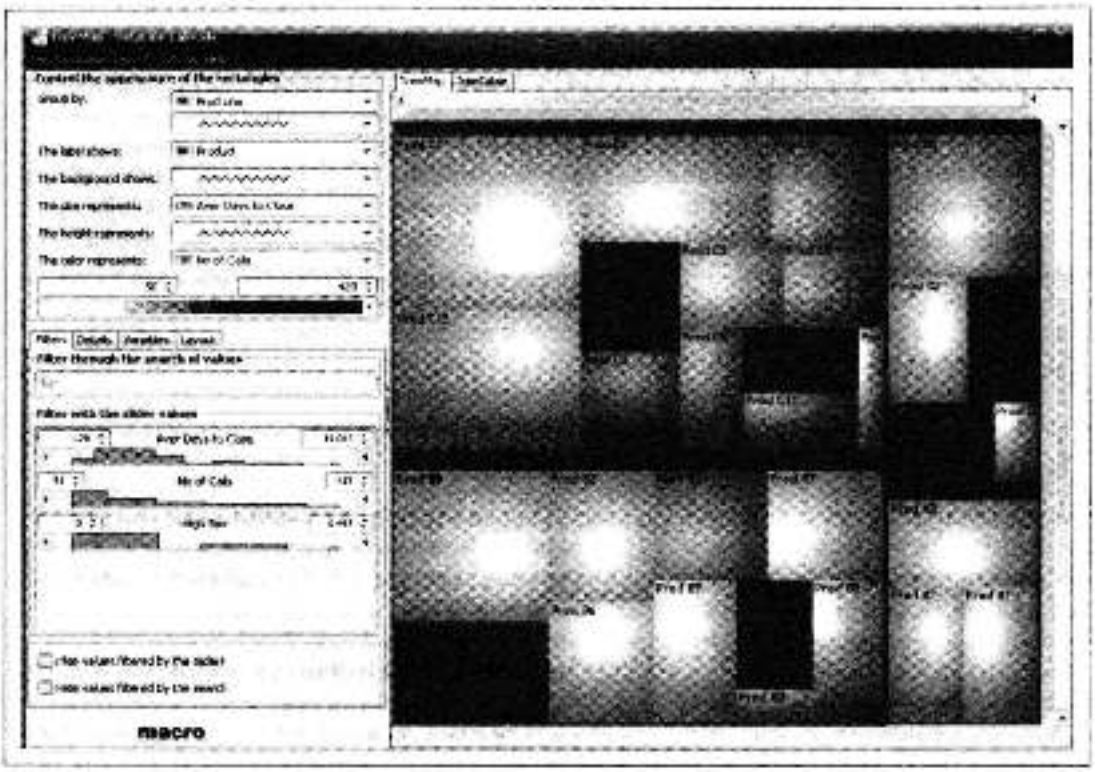


图3-10 呼叫中心的数据树图

其他的数据可视化技术可参考网页链接：webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization，以及 smashingmagazine.com/2007/08/02/-data-visualization-modern-approaches。Grimes（2009）给出了一些重要的擅长于 BPM 和 BI 应用的数据可视化厂商（例如，tableausoftware.com）的清单及描述。

3.9 节复习题

1. 记分卡和仪表盘主要区别有哪些？
2. 作业仪表盘和战术仪表盘有什么区别？
3. 仪表盘每层包含的信息是什么？
4. 为仪表盘的度量值选择展示工具的标准有哪些？
5. 一个设计优秀的仪表盘有哪些特征？
6. 什么是数据可视化？BPM 和 BI 中展示数据遇到的主要问题有哪些？

本章重点

- BPM 是指对企业的绩效进行度量、监控和管理的过程、方法、度量和技术。
- BPM 是 BI 的产物，包含了许多 BI 中的知识、应用和技术。

- BI 成为了描述访问、分析、报告企业数据相关技术的术语。
- BI 的实践和软件成为全面的 BPM 解决方案的主要部分。
- BI 和 BPM 的主要区别在于 BPM 通常是战略驱动的。
- BPM 是一个从决策到执行并使企业的绩效达到最优的闭环过程。
- BPM 的关键步骤是战略、计划、监管、实施和调整。
- 战略回答了“我们未来要走向哪里”的问题。
- 几十年的研究使战略和执行之间的差距加大。
- 战略和执行之间的差距存在于沟通、结盟、焦点和资源中。
- 作业计划和战术计划主要强调“我们如何能通向未来”。
- 作业计划中的策略和方案必须要服从于战略计划的方向和目标。
- 一个组织的战略目标和关键指标应当和企业的自上而下的有形资产和无形资产分配方式相匹配。
- 监管强调了“我们正在如何做的”。
- 平衡记分卡、绩效仪表盘、项目监管系统、人力资源系统和财务报表系统都归属于诊断控制系统。
- 多数监管强调负差异，忽视了最基本的假设和战略。
- 传统的计划承受着各种偏见，包括证实、相近因素、成功的示例及社会或政治偏见。
- 发现—驱动计划提供了系统性地发现问题假设的方法，否则这些问题就会潜伏在计划和监管过程中。
- 新项目和管理的失败率在 60% ~ 80% 之间。
- 一般企业中企业管理人员制订计划和报告，几乎没有时间回顾战略的观点，决定什么应该按照不同方法的去做，并相应地修改计划。
- 绩效评价系统通过将战略目标和实际进行对比，来帮助管理者跟踪战略的实施情况。
- 财务数据作为绩效评价系统的核心工作所遇到的困难已经很明确。
- “运营一个工厂”度量标准和“按战略执行”度量标准有区别。
- 绩效评价源于企业或业务战略，以及实现这些战略的关键业务过程。
- 最流行、最常用的绩效管理软件是平衡记分卡。
- BSC 的核心是按照组织战略方案对业绩进行评价比较。
- BSC 作为一个评价绩效的方法，克服了只关注财政上的系统限制。
- 按日历驱动的财务报告是绩效评价系统的一个重要组件。
- BSC 作为一个战略性的管理方法，使组织能够将实施方法和战略保持一致。
- 在 BSC 中，企业战略地图展现了企业战略目标和各个目标之间的关联。
- 多数企业利用六西格玛作为流程改进方法，使企业对运营过程进行详细检查、指出存在问题并进行补救。
- 六西格玛绩效管理方法的主要目的是减少企业运营过程中的缺陷，使每百万次采样数的缺陷率达到零。
- 六西格玛利用 DMAIC 的流程，形成一个闭环的商务模型，包含了对流程的定义、测量、分析、改进、控制。
- 近几年，人们开始关注六西格玛方法和精益战略之间的关联。
- 精益生产是指减少非增值活动和浪费，然而六西格玛的核心是减少不一致，提高流程的一致性。
- 通过将平衡记分卡和六西格玛一起使用，有助于企业绩效管理。
- 一个标准的 BPM 架构由多个层次组成，包含了 BPM 程序、信息中心、不同来源系统的数据。
- BPM 的主要功能包括：战略管理，预算、计划和预测管理，财务数据合并，利润分析和最优化，以及法定的财务管理报告。
- 在过去的 3 ~ 4 年间，BPM 市场的最大变化是 BPM 厂商的合并。
- 记分卡和仪表盘几乎是绩效管理系统、绩效评价系统和 BPM 套件中都包含的组件。
- 尽管记分卡和仪表盘都提供了可视化视图，但是两者有明显区别。
- 仪表盘与众不同的地方在于它的 3 层结构：监管、分析和管理。
- 仪表盘设计的根本问题在一个简单的图像上没有歧义地、易于理解地展示所有必要的信息。
- 新的数据可视化技术能够帮助用户分析 BPM 和 BI 应用中大型的、复杂的多维数据。

关键术语

Balanced Scorecard（BSC，平衡记分卡）
data visualization 数据可视化
learning 学习
strategic goal 战略目标
strategic objective 战略目的
Business Performance Management（BPM，业务绩效管理）
diagnostic control system 诊断控制系统
optimization 最优化
strategic vision 战略愿景
strategy map 战略地图
Critical Success Factors（CSF，关键成功因素）
DMAIC 一个闭环业务改进模型，指定义、测量、分析、改进、控制5个阶段
Key Performance Indicator（KPI，关键绩效指标）
performance measurement systems 绩效评价系统
strategic theme 战略主题
dashboards 仪表盘
Lean Manufacturing 精益生产
scorecards 记分卡
Six Sigma 六西格玛
system architecture 系统架构

讨论题

1. SAP 的战略企业管理，Cognos 的企业绩效管理以及 Hyperion 的业务绩效管理，它们的基本观点相同吗？并说明依据。
2. BPM 包括了5个基本过程：战略、计划、管理、实施和监控。选择其中的一个过程，阐述可以支持这个过程的软件工具和应用的种类。图3-1给出了类示。可以查阅 Bain 和 Company 的管理工具（bain.com/management_tools/home.asp）。
3. 选择一个上市公司，用该公司2008年的年度报告，为2009年制定3个财务战略目标。为每个目标制定一个目的和指标。该目标要和2008年的财务状况保持前后一致。
4. Netflix 的在线视频下载战略在多个文献中被应用。Netflix 战略的基本目标是什么？该战略的主要假设条件有哪些？这些假设条件合理吗？
5. 近几年，超越预算圆桌会议（Beyond Budgeting Round Table，BBRT）（bbrt.org）已经向传统的预算实践提出了挑战。网上的许多文献中提到了超越预算圆桌会议的应用，在超越预算圆桌会议的观点中，对于如今的预算实践来说哪些是错误的？超越预算圆桌会议提出了哪些替代方法？
6. 描述 BSC 是如何应用于诊断控制系统的。
7. 绩效管理和绩效评价的区别。
8. 欧洲质量管理基金会（EFQM）的"卓越模型"提供了可选的绩效评价方法和管理架构。首先，EFQM 代表什么意思？其次，借助网上资料，说明该框架的主要原则。将该框架与平衡记分卡、六西格玛进行比较。
9. 为你感兴趣的战略目标制定一个评价办法（可以使用问题3中制定的一个目标）。选择评价办法，完成本章3.2.1节中的评价模型。
10. 用评价记分卡中的4个观点，为一个假定公司制定一个战略，阐述战略中的一系列策略目标。制定一个战略地图，描述这些目标之间的关系。
11. 比较 DMAIC 模型和闭环 BPM 系统。
12. 选择表3-4Gartner 魔力象限中的两个公司（除了 SAP、Oracle 和 IBM），说明他们的 BPM 套件中包含的组件，比较他们的绩效管理中组件的应用和功能。

练习

Teradata 大学和其他的动手练习题

1. 进入 teradatastudentnetwork.com，选择"Articles（文献）"，在文献列表中找出一个题为"Business/Corporate Performance Management：Changing Vendor Landscape and New Market Targets"的文章，在阅读该文章的基础上，回答下列问题：
   a. 该文章的基本观点是什么？

b. 该文章中的"take away"的前提是什么?
c. 本文中，关键路径法中哪个功能或角色最详细?
d. 关键路径法包含了哪些功能?
e. 本文中的关键路径法与Gartner的关键路径法包含的功能相比，两者之间有什么相同和不同之处?
f. 什么是CRC，它和企业绩效之间有什么联系?
g. 在过去几年中，关键路径法市场定位的关键点有哪些?
h. 选择该文章中提到的2个公司（除了SAP、Oracle和IBM），指出每个公司的关键路径法战略是什么?作者对这些战略有什么看法?

2. 进入teradatastudentnetwork.com，选择"Case Studies（案例学习）"，在案例列表中找出一个题为"Real-Time Dashboards at Western Digital"的案例，在阅读该文章的基础上，回答下列问题：
   a. 什么是VIS?
   b. BPM架构和VIS架构之间有哪些相同点和不同点?
   c. 闭环BPM和OODA决策环之间有哪些相同点和不同点?
   d. 系统中的仪表盘种类有哪些?它们是操作型的还是战略型的，或者说它们是真正的仪表盘吗?解释说明。
   e. Western Digital的VIS和仪表盘有哪些优势?
   f. 对于一个将要实施VIS和仪表盘的公司，你能提出哪些建议?
3. 进入Stephen Few的博客"the Perceptual Edge"（perceptualedge.com），选择"Examples（举例）"，可以看到各种仪表盘示例，阅读其中的几个示例。然后进入dundas.com，选择"Gallery"，点击"Digital Dashboard（数据仪表盘）"，将看到各种仪表盘示例，并看其中的几个示例。
   a. 这些数据分别包含了那些信息和度量值?根据这些能采取什么行动?
   b. 利用Few阐述的几个概念，指出这些示例中的优缺点。
4. 利用一个仪表盘模型展示出一个上市公司的财务状况，该模型可用文本文档或者Excel。用2个上市公司2008年的数据说明你指定的仪表盘的功能。

小组作业和角色扮演

1. 几乎所有的BPM/CPM供应商都在网上提供了自己的案例学习材料，小组选择其中的2个供应商（可从Gartner或AMR列表中得到这些供应商的名称）网站。从每个网站中选出2个案例，总结出每个案例中客户面临的问题、解决方案以及带给用户的好处。
2. 进入到仪表盘网站（enterprise-dashboard.com/sitemap），该网站提供了各种业务管理仪表盘。小组选择一个行业（如保健、银行、航空），举出几个该行业应用仪表盘的例子，指出每个仪表盘的度量值及用哪些方式展示的信息。用所知道的仪表盘模型，为这些信息制定一个仪表盘模型。

网络练习

1. 经济学人智囊团完成并发表在S. Tools的调查报告"Closing the Strategy-to-Performance Gap"，CFO Magazine，2005年2月22日（cfo.com/article.cfm/3686974?f=related），研究探索了战略和实施方法之间的关系。根据该调查，绩效管理、战略和实施方法中哪个更重要?为什么实施效果不好，采用什么方式来提高绩效?
2. 进入到纽约城市管理报告网（nyc.gov/html/ops/html/home/home），按报告题目将这些报告进行归类（参考本章应用案例）。参考该网站回答下列问题：
   a. 该城市使用了哪些绩效指标?哪些用于社区服务?哪些用于教育事业?
   b. 总体来看，这些指标中哪些正在改进，哪些稳定不变?哪些正在下降?
   c. 在教育事业，这些指标中哪些正在改进，哪些稳定不变?哪些正在下降?
   d. 选择"Citywide Themes（城市主题）"（网页的左上角），其中的一个主题是"Social Services（社会服务）"，选择该主题。然后，选择"View the performance report for Social Services（社会服务绩效报告）"。这里总共包含了多少绩效指标?哪些下降率大于10%?哪些指标在下降?纽约市是如何应付这些绩效指标问题的?
3. 一个著名的BSC例子就是西南航空公司创建的用于管理自己业务的BSC。Anthes的一个早期文献指出了该系统的战略图（computerworld.com/action/article.do?command=viewArticleBasic&articleId=78512）。

检索该文献，利用战略图描述西南航空公司的战略。西南航空公司是用哪些方法使绩效和战略目标相一致的？根据你对当今经济形势和航空行业的了解，你认为西南航空公司的战略适用于当今的经济形势吗？

4. 数据仓库协会（The Data Warehouse Institute，TDWI）每年对那些在开发、部署和维护 BI 和数据仓库领域表现突出的公司进行鉴定和排名。到网上查看 2008 年的排行榜（tdwi. org/research/display. aspx?id =9000）。TDWI 鉴定的厂商包含了哪些种类？哪些厂商是优胜者？他们是如何获得优胜者的称号的？
5. 许多网站提供了仪表盘和记分卡的例子和指导手册，为这些网站提供的每个功能制定一个 Excel 原型。
6. 最近，Oracle 的一个白皮书"Business Intelligence and Enterprise Performance Management：Trends for Mid-size Companies"（oracle. com/appserver/business-intelligence/hyperion-financial-performance-management/docs/bi-epm-trends-for-emerging-businesses. pdf），对中型和大型公司的 BI 和绩效管理应用进行了比较分析。首先，该书中什么样的公司是中型公司？其次，这些公司的应用包含了哪些种类？这两种类型公司有什么相同点和不同点？他们给这些厂商提出了哪些结论和意见？
7. 进入 webdesignerdepot. com/2009/06/50-great-examples-of-data-visualization。选择两种数据可视化技术，并指出这种技术是如何在 BI 和 BPM 系统中发掘和描述数据的。要包含这些技术的优缺点。

本章结尾应用案例

跟踪城市绩效管理

大量的系统产生了大规模的度量值，目前用户遇到的问题是如何对数据进行切片，特别是当用户拥有丰富的信息技术经验时，使用户更方便使用数据而不是被数据困扰用户。这是纽约城市绩效报告（Citywide Performance Reporting，CPR）在线系统所面临的一个主要问题，一个交互式的仪表盘为政府机构和市民提供了"以友好的形式访问每个城市机构的关键绩效指标，并且每月自动更新，自动地评估专业领域的趋势"。

CPR 的发展趋势

CPR 是 NYCStat 的一部分（nyc. gov/html/ops/nycstat/html/home/home. shtml），NYCStat 是纽约城市提供的一站式服务点，提供了城市服务的基础数据、报告和统计数据。NYCStat 提供了和各种绩效相关信息的接口，包括全市专业机构的信息，纽约"311"市民服务热线数据，根据选择的绩效数据和生活质量指标进行交互的地图特征。

CPR 建立于 2008 年 2 月，CPR 最初是 2005 年市长运作办公室与信息技术和通信技术部（Department of Information Technology and Communication，DoITT）相互合作开发的 2005 中期项目，该项目包含了以下 3 个组件（NYCStat，2009）：

- **绩效管理应用** 后端计算机系统为专业行政机构设置了输入数据的单一访问点。
- **分析工具/仪表盘** 前端系统提供了标准的报表格式，包含了向下展开信息、绩效概况和趋势图。

数据定义 查看并定义 44 个主要行政机构的 CPR 系统中包含的主要度量值和指标。

这些组件开发完成于 2007 年 7 月，那时，系统对 44 个行政机构和主要的办公系统进行开放并进行回顾。2007 年夏季末，市长 Bloomberg 指出，该系统将尽快通过纽约城市网站对公众开放。随后的工作将致力于使公众更容易、更方便地使用该系统的仪表盘。系统最终在 2008 年 4 月 14 日对外开放，并作为 2008 年度市长管理报告显示的一部分。

精炼度量值

城市绩效报告是市长管理报告（Mayor's Management Report，MMR）的一个分支，是城市行政机构每年对 1 000 个指标进行评估的结果，这些指标包括了从学校考试分值到处理问题所花费的时间等各个指标。CPR 不仅是 MMR 的在线版本，而且还要求行政机构随时更新他们的服务，并指出怎样最好地评价这些服务的产出。起初，这些指标成千上万，而且许多没有被使用。最终，这些指标数减少到 525 个，其中主要的指标用于评价影响城市居民的最终服务输出情况。这就是为什么该系统向公众开放。

为了精简这些指标，还要考虑其他一些关键因素。首先，数据按月或按年获得，因此要决定如何对数据进行切片或切块，以及对不同时期的数据进行横向比较。其次，决定怎样评价和呈现发展趋势及绩效和目标之间的关系，指明指标所期望的方向及绩效好坏的标准。最后，由于指标和行政机构的种类很多，公众对每个行政机构不是太了解，因此要提供简明的表现形式和导航说明。开发人员根据政府对在纽约市工作和居住的市民提供的服务形式，将这些指标分为 8 个"主题"类型。包括：全市管理、社区服务、经济

发展和商务事件、基础设施、教育、法律事件、公众安全和社会服务。

CPR 的影响

从市长办公室的角度来看，CPR 通过负有责任、透明度和可接近性 3 个角度来提高绩效管理效果，并通过以下功能实现绩效的提升（NYCStat，2009）：

- 对重要的"产出"绩效指标进行评价，这些指标反映了城市管理政策对居民生活的影响。
- 将现在的数据和以往的数据进行比较，使行政机构对每年的绩效提升负责。
- 利用图表和不同的颜色标志来表示行政机构的绩效，很明确地看出绩效的趋势是积极的还是消极的。
- 提供钻取的功能，使用户可以看到 5 年内的发展趋势状况。
- 将反映不同种类行政机构管理的重要指标加入到城市管理主题中，得到城市绩效管理的情况。
- 对数据按月、季度或年进行更新，确保最近的数据是可用的。
- 提供各种形式的数据下载，以便进行深入分析。
- 对每个指标提供详细的说明，包括指标的含义、汇报的频率和其他有用的细节。

2009 年 3 月，CPR 系统被哈佛肯尼迪政府学院"国家管理和改革 Ash 机构"评定为 2009 年度最优秀的 50 个管理革新方法之一（New York Nonprofit Press，2009）。

课堂学习

Sarlin 指出"看到政府在各个方面的管理工作像经营一个公司一样的良好实践很有趣。"这可以从纽约城市市长运作办公室的案例中得到证实，因为 Bloomberg 市长来自商业界，更确切地说是实时的财务信息界（bloomberg.com）。然而，有些政府机构表明在绩效表盘和商务几乎没有联系。如，美国的首席信息官 Vivek Kundra 最近指出，美国的联邦 IT 仪表盘提供了行政管理和预算局（Office of Management and Budget，OMB）的数据报告，包括 7 000 项联邦 IT 投资的基本信息及其中的将近 800 个主要的投资项的详细数据。Mr. Kundra 将他全部的事业都投入到了政府部门。商业活动可以从这些仪表盘方案中获得许多有价值的东西。

这些政府仪表盘方案具有以下优势（Boytendijk，2008）：

- **透明度有很大差异** 这些方案提供了大量的公众可访问的绩效管理数据，许多企业可以对行政机构提供的信息进行深入和广泛的学习。
- **合作的重要性** 许多 BI 项目遇到的困难是不同部门要求做不同的仪表盘或者记分卡，每个部门实现了短期的投资收益率。然而，这样导致总投资收益率达不到最优。CPR 仪表盘和联邦 IT 仪表盘示例表明，开发一个跨多个领域的组织层面的方案是可行的。
- **不断的改进** CPR 主要基于趋势分析，而不是目标驱动。这表明绩效指标是不断地改进，而不是以静止目标为导向。这为作业型仪表盘和战术型仪表盘提供了一个很好的例子。

本章结尾应用案例的问题

1. CPR 仪表盘的主要组成有哪些？
2. 在 CPR 仪表盘中制定和实施过程中包含了多少行政机构部门？
3. 制定和实施 CPR 的主要步骤有哪些？
4. CPR 仪表盘的"主题"扮演者什么角色？
5. CPR 仪表盘主要的功能有哪些？
6. 商务活动可以从 CPR 这样的政府方案中学到什么？

来源：Compiled from NYCStat, "CPR Fact Sheet," Mayor's Office of Operations, February 2009, nyc.gov/html/ops/cpr/downloads/pdf/cpr_fact_sheet.pdf (accessed January 2010); R. Sarlin, "Mayor Unveils Web Database Tracking Performance," New York Sun, February 15, 2008, nysun.com/new-york/mayor-unveils-webdatabase-tracking-performance/71347? print=511986642 1 (accessed January 2010); F. Boytendijk, "The Mother of All Accountability Tools," February 20, 2008, blogs.oracle.com/frankbaytendijk/2008/02/the_mother_of_all_accountabil.htm (accessed January 2010); New York Nonprofit Press, "Eight NYC Programs Among 50 Selected for National Honors," March 31, 2009, nynp.biz/index.php/breaking-news/620-eight-nyc-programs-among-50-selected-fornational-honors- (accessed January 2010); J. Hiner, "U. S. Federal IT Dashboard is a Great Example of How to Promote IT," July 1, 2009, blogs.zdnet.com/BTL/?p=20157 (accessed January 2010).

参考文献

Antic, S., and J. McGregor. (2006, February 13). "Giving the Boss the Big Picture." *BusinessWeek* **businessweek.com/magazine/content/06_07/b3971083.htm** (accessed January 2010).

Axson, D. (2007). *Best Practices in Planning and Performance Management: From Data to Decisions.* New York: Wiley.

Buytendijk, F. (2008, February 20). "The Mother of All Accountability Tools." **blogs.oracle.com/frankbuytendijk/2008/02/the_mother_of_all_accountabili.html** (accessed January 2010).

Calumo Group. (2009, January 30). "Planning on Microsoft BI Platform." **calumo.com/newsblog** (accessed January 2010).

Caterpillar Inc. (2009, September). "Caterpillar Inc. Announces 2Q 2009 Results." *Caterpillar Press Release* **cat.com/cda/components/fullArticleNoNav?m=37523&x=7&id=1654623** (accessed January 2010).

Chandler, N., N. Rayner, J. Van Decker, and J. Holincheck. (2009, April 30). "Magic Quadrant for Corporate Performance Management Suites." *Gartner RAS Core Research Note G00165786.* **mediaproducts.gartner.com/reprints/oracle/article51/article51.html** (accessed January 2010).

Colbert, J. (2009, June). "Captain Jack and the BPM Market: Performance Management in Turbulent Times." *BPM Magazine.* **bmpmag.net/mag/captain_jack_bpm** (accessed January 2010).

Conference Board. (2008, December 2). "Weakening Global Economy and Growing Financial Pressures are Increasing CEO Concerns." Press Release. **conference-board.org/utilities/pressDetail.cfm?press_id=3529** (accessed January 2010).

Docherty, P. (2005). "From Six Sigma to Strategy Execution." **i-solutionsglobal.com/secure/FromSixSigmaToStrateg_AACSC.pdf** (accessed January 2010).

Eckerson, W. (2009, January). "Performance Management Strategies: How to Create and Deploy Effective Metrics." *TDWI Best Practices Report.* **tdwi.org/research/display.aspx?ID=9390** (accessed January 2010).

Eckerson, W. (2006). *Performance Dashboards.* Hoboken, NJ: Wiley.

Few, S. (2005, Winter). "Dashboard Design: Beyond Meters, Gauges, and Traffic Lights." *Business Intelligence Journal,* Vol. 10, No. 1.

Few, S. (2008). "Data Visualization and Analysis—BI Blind Spots." *Visual Perceptual Edge.* **perceptualedge.com/blog/?p=367** (accessed January 2010).

Green, S. (2009, March 13). "Harrah's Reports Loss, Says LV Properties Hit Hard." *Las Vegas Sun.* **lasvegassun.com/news/2009/mar/13/harrahs-reports-loss-says-lv-properties-hit-hard** (accessed January 2010).

Grimes, S. (2009, May 2). "Seeing Connections: Visualizations Makes Sense of Data." *Intelligent Enterprise.* **i.cmpnet.com/intelligententerprise/next-era-business-intelligence/Intelligent_Enterprise_Next_Era_BI_Visualization.pdf** (accessed January 2010).

Gupta, P. (2006). *Six Sigma Business Scorecard,* 2nd ed. New York: McGraw-Hill Professional.

Hammer, M. (2003). *Agenda: What Every Business Must Do to Dominate the Decade.* Pittsburgh, PA: Three Rivers Press.

Hatch, D. (2008, January). "Operational BI: Getting 'Real Time' about Performance." *Intelligent Enterprise.* **intelligententerprise.com/showArticle.jhtml?articleID=205920233** (accessed January 2010).

Henschen, D. (2008, September). "Special Report: Business Intelligence Gets Smart." **intelligententerprise.com/showArticle.jhtml?articleID=210500374** (accessed January 2010).

Hindo, B. (2007, June 11). "At 3M: A Struggle between Efficiency and Creativity." *BusinessWeek.* **businessweek.com/magazine/content/07_24/b4038406.htm?chan=top+news_top+news+index_best+of+bw** (accessed January 2010).

Hiner, J. (2009, July 1). "U.S. Federal IT Dashboard is a Great Example of How to Promote IT." **blogs.zdnet.com/BTL/?p=20157** (accessed January 2010).

Kaplan, R., and D. Norton. (2008). *The Execution Premium.* Boston, MA: Harvard Business School Press.

Kaplan, R., and D. Norton. (2004). *Strategy Maps: Converting Intangible Assets into Tangible Outcomes.* Boston, MA: Harvard Business School Press.

Kaplan, R., and D. Norton. (2000). *The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment.* Boston, MA: Harvard Business School Press.

Kaplan, R., and D. Norton. (1996). *The Balanced Scorecard: Translating Strategy into Action.* Boston, MA: Harvard University Press.

Kaplan, R., and D. Norton. (1992, January–February). "The Balanced Scorecard—Measures That Drive Performance." *Harvard Business Review,* pp. 71–79.

Knowledge@W.P. Carey. (2009, March 2). "High-Rolling Casinos Hit a Losing Streak." **knowledge.wpcarey.asu.edu/article.cfm?articleid=1752#** (accessed January 2010).

Krafcik, J. (1988, Fall). "Triumph of the lean production system." *Sloan Management Review,* Vol. 30, No. 1.

Leahy, T. (2005, February). "The One-Two Performance Punch." *Business Finance.* **businessfinancemag.com/magazine/archives/article.html?articleID=1436** (accessed January 2010).

McGrath, R., and I. MacMillan. (2009). *Discovery-Driven Growth.* Boston, MA: Harvard Business School Press.

McKay, L. (2009, February). "Megavendors Look Smart in Gartner Magic Quadrant for Business Intelligence." **destinationcrm.com/Articles/CRM-News/Daily-News/Megavendors-Look-Smart-in-Gartner-Magic-Quadrant-for-Business-Intelligence-52600.aspx** (accessed January 2010).

Microsoft. (2006, April 12). "Expedia: Scorecard Solution Helps Online Travel Company Measure the Road to Greatness." **microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=49076** (accessed January 2010).

New York Nonprofit Press. (2009, March 31). "Eight NYC Programs Among 50 Selected for National Honors." **nynp.biz/index.php/breaking-news/620-eight-nyc-programs-among-50-selected-for-national-honors-** (accessed January 2010).

Norton, D. (2007). "Strategy Execution—A Competency that Creates Competitive Advantage." The Palladium Group. **thepalladiumgroup.com/KnowledgeObjectRepository/Norton_StrategyExec creates competitive adv WP.pdf** (accessed January 2010).

Novell. (2009, April). "Executive Dashboards Elements of Success." Novell White Paper. **novell.com/rc/docrepository/public/37/basedocument.2009-03-23.4871823014/**

Executive Dashboards_Elements_of_Success_White_Paper_en.pdf (accessed January 2010).

NYCStat. (2009, February). "CPR Fact Sheet." Mayor's Office of Operations. **nyc.gov/html/ops/cpr/downloads/pdf/cpr_fact_sheet.pdf** (accessed January 2010).

Poppendieck, LLC. (2009). "Why the Lean in Six Sigma." **poppendieck.com/lean-six-sigma.htm** (accessed January 2010).

Richardson, K. (2007, January 4). "The Six Sigma Factor for Home Depot." *Wall Street Journal Online.* **mkimcintosh.com/articles/WSJjan4.pdf** (accessed January 2010).

Rigby, D., and B. Bhockan. (2009). "Management Tools and Trends 2009." Bain & Co. **bain.com/management_tools/Management_Tools_and_Trends_2009_Global_Results.pdf** (accessed January 2010).

Sarlin, B. (2008, February 15). "Mayor Unveils Web Database Tracking Performance." *New York Sun.* **nysun.com/new-york/mayor-unveils-web-database-tracking-performance/71347/?print=5119866421** (accessed January 2010).

Shill, W., and R. Thomas. (2005, October). "Exploring the Mindset of the High Performer." *Outlook Journal.* **accenture.com/Global/Research_and_Insights/Outlook/By_Issue/Y2005/ExploringPerformer.htm** (accessed January 2010).

Simons, R. (2000). *Performance Measurement and Control Systems for Implementing Strategy.* Upper Saddle River, NJ: Prentice Hall.

Six Sigma Institute. (2009). "Lean Enterprise." **sixsigmainstitute.com/lean/index_lean.shtml** (accessed August 2009).

Slywotzky, A., and K. Weber. (2007). *The Upside: The 7 Strategies for Turning Big Threats into Growth Breakthroughs.* New York: Crown Publishing.

Smith, R. (2007, April 5). "Expedia-5 Team Ring: Technology." **expedia-team5.blogspot.com** (accessed January 2010).

Stanley, T. (2006, February 1). "High Stakes Analytics." *InformationWeek.* **informationweek.com/shared/printableArticle.jhtml?articleID=177103414** (accessed January 2010).

Swabey, P. (2007, January 18). "Nothing Left to Chance." *Information Age.* **information-age.com/channels/information-management/features/272256/nothing-left-to-chance.thtml** (accessed January 2010).

Tucker, S., and R. Dimon. (2009, April 17). "Design to Align: The Key Component in BPM Success." *BPM Magazine.* **bpmmag.net/mag/design-to-align-key-component-in-bpm-success-0417** (accessed January 2010).

Vesselle, D., and D. McDonough. (2008, November). "Worldwide Business Analytics Software 2008–2012 Forecast and 2007 Vendor Shares." **DC Doc # 214904. sas.com/offices/europe/denmark/pdf/idc_ba_1108.pdf** (accessed January 2010).

Wailgum, T. (2008, January 25). "How IT Systems Can Help Starbucks Fix Itself." *CIO.* **cio.com/article/176003/How_IT_Systems_Can_Help_Starbucks_Fix_Itself** (accessed January 2010).

Watson, H., and L. Volonino. (2001, January). "Harrah's High Payoff from Customer Information." *The Data Warehousing Institute Industry Study 2000—Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions.* **terry.uga.edu/~hwatson/Harrahs.doc** (accessed January 2010).

Weier, M. (2007, April 16). "Dunkin' Donuts Uses Business Intelligence in War Against Starbucks." *InformationWeek.*

Wextel, M. (2008, June 13). "Reasons for Six Sigma Deployment Failures." *BPM Institute.* **bpminstitute.org/articles/article/article/reasons-for-six-sigma-deployment-failures.html** (accessed January 2010).

商务智能中的数据挖掘

学习目标

- 定义数据挖掘是商务智能的实现技术；
- 理解业务分析和数据挖掘的目标和好处；
- 认识数据挖掘的广泛应用；
- 学习数据挖掘标准化过程；
- 理解数据挖掘中数据预处理的基本步骤；
- 学习数据挖掘的各种方法和算法；
- 认识现有数据挖掘软件工具；
- 理解关于数据挖掘的一些谎言和误解。

通常来说，数据挖掘是从组织机构收集、组织和存储的数据中开发商务智能的一种方法。大量的数据挖掘技术被组织机构用来更好地理解他们的顾客和自己的运营，解决复杂的组织问题。这一章将研究作为商务智能技术的数据挖掘，学习实施数据挖掘项目的标准化过程，理解和建立数据挖掘技术使用的专门知识，发展对现有软件工具的认识程度，探索数据挖掘存在的缺陷和神话。

开篇场景：数据挖掘来到好莱坞

预测某一电影的票房收入（也就是财务上的成功）是一个有趣且具有挑战性的问题。有些领域专家认为，电影业是一个“依靠直觉和猜测”的领域。由于很难预测产品需求，所以在好莱坞做电影业务是一种冒险行为。与此观点对应的是，Jack Valenti（长期任美国电影协会主席和CEO）曾指出，“没有人能告诉你一部电影的市场表现将如何……直到电影在漆黑的影院开幕，在观众和荧幕之间擦出火花。”娱乐行业的贸易杂志中，大量例子、陈述和经验都完全支持这种说法。

就像其他很多研究者试图阐释这一极具挑战性的实际问题一样，Ramesh Sharda 和 Dursun Delen 在制片商未开始的阶段（在电影仅仅还是一个概念创意时）就应用数据挖掘进行电影的票房收入表现预测。在他们广为宣传的预测模型中，他们将预测（回归）问题转换为一个分类问题。换句话说，他们基于票房收入将电影划分为从“失败”到“拳头产品”的9大类，而不是对票房收入进行预测点估计，从而将问题转换成为一个多项式分类问题。表4-1阐明了票房收入的等级范围划分定义。

表4-1 基于收入的电影分类

分类号	1	2	3	4	5	6	7	8	9
范围（百万美元为单位）	<1（失败）	>1 <10	>10 <20	>20 <40	>40 <65	>65 <100	>100 <150	>150 <200	>200（一鸣惊人）

数据

数据来源于与各种电影相关的数据库（例如，ShowBiz、IMDb、IMSDb、AllMovie），然后将它们合并成为一个单独的数据集。最近开发的模型数据集，包括了从1998年到2006年间发布的2 632部电影。表4-2总结了各独立变量及其规格定义。要详细了解这些独立变量，读者可以参阅文献 Sharda and Delen（2007）。

表4-2 独立变量概要

独立变量	可能值的数目	可能值
MPAA（美国电影协会）等级	5	G、PG、PG-13、R、NR
竞争	3	高、中、低
明星价值	3	高、中、低
流派	10	科幻、历史剧、现代剧、政治相关片、惊悚片、恐怖片、喜剧、卡通、动作、纪录片
特技	3	高、中、低
续篇	1	是、否
屏幕数目	1	正整数

解决方案

应用各种数据挖掘方法，包括神经网络、决策树、支持向量机（Support Vector Machine，SVM），以及3种类型的组合算法，Shanks和Delen开发了预测模型。而2006年的数据则被用作测试数据来评估和比较模型的预测精确度。图4-1展示了用SPSS的PASW Modeler（即以前的Clementine数据挖掘工具）描述的预测问题的过程图。过程图的左上部分展示了模型开发过程，右下角则展示了模型的评估（测试或评分）过程。关于PASW Modeler工具及其使用细节可参考本书的Web网址。

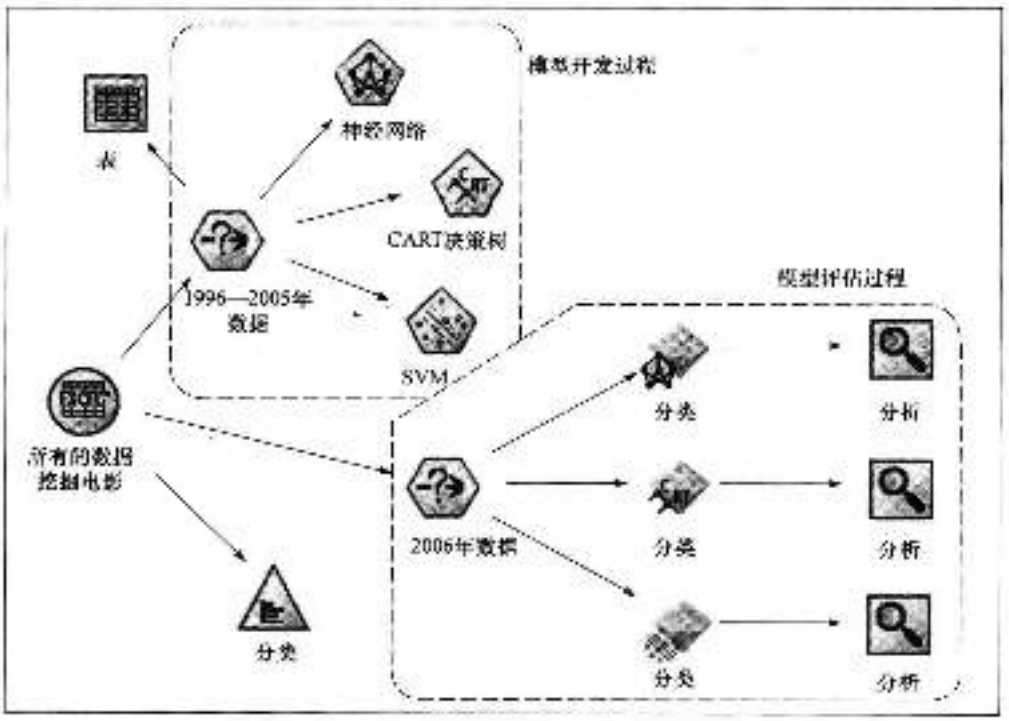


图4-1 票房收入预测系统流程截图
来源：SPSS. Used with permission.

结果

表4-3给出了所有3种数据挖掘方法和3种不同类型组合方法的预测结果。第1个性能指标称为bingo，表示等级正确分类的比率。表中还给出了在一个范畴内（1-Away）的正确等级分类比率。结果表明，在独立预测模型中，SVM表现最好，其次是ANN，最差的是CART决策树算法。总体上，组合模型表现优于独立预测模型。其中，融合算法表现最好。

表 4-3 独立和组合模型的预测结果表格

预测模型						
	独立模型			组合模型		
性能指标	SVM	ANN	CART	随机森林	增强树	整合（平均）
数目（Bingo）	192	182	140	189	187	194
数目（1-Away）	104	120	126	121	104	120
精确度（%Bingo）	55.40%	52.60%	40.46%	54.62%	54.05%	56.07%
精确度（%1-Away）	85.55%	87.28%	76.88%	89.60%	84.10%	90.25%
标准差	0.93	0.87	1.05	0.76	0.84	0.63

从表中我们可以看到，组合模型预测结果的标准差显著小于独立模型预测结果的标准差，这一点对于决策者可能更为重要。

结论

研究者认为，该预测结果优于本领域现有文献中的其他结果。对于票房收入的预测结果具有非常高的精确度。而且，这些模型可以被用于进一步分析和优化决策变量，从而使得财务收入最大化。特别地，可以使用已经训练过的预测模型改变建模参数，以更好地理解不同参数对于最终结果的影响。通过这样一个敏感度分析的过程，特定娱乐企业的决策者可以很精确地了解某位演员（或某个发布日期、某种技术效果等）对于公司财务成功的贡献程度。这使得该系统成为一种宝贵的决策辅助手段。

开篇场景的问题

1. 为什么好莱坞的决策者需要用到数据挖掘？
2. 好莱坞管理者面临的首要挑战是什么？你能否想出面临类似问题的其他行业？
3. 研究者是否使用了所有数据来建立预测模型？谈谈你的看法。
4. 研究者为什么选择将一个回归问题转变为一个分类问题？谈谈你的看法。
5. 你认为该如何应用这些预测模型？你能否为这些模型想出一个很好的生成系统？
6. 你认为决策者是否容易适应这样的信息系统？
7. 如何进一步改善实例中的预测模型？

我们从开篇场景中能够学到什么

娱乐业的决策者们面临着很多有趣而富有挑战性的问题。对于娱乐业市场上的很多公司来说，对巨额财富进行正确的管理决策对其成功（或仅仅是生存）而言至关重要。对于这样一个"数据丰富，知识贫乏"的业务环境，数据挖掘是其实现更好管理的重要选择。开篇故事中的研究清楚地说明了数据挖掘能够预测和解释电影财务表现的能力，尽管大多数观点认为电影是一种艺术形式因而是不能被预测的。在本章中，读者可以学习数据挖掘在各行业中的广泛应用。您将学习数据挖掘如何应用数据解决一系列复杂的产业问题，从而提升商业竞争优势。

来源：R. Sharda and D. Delen, "Predicting Box-Office success of Motion Pictures with Neural Networks", *Expert Systems with Applications*, vol. 30, 2006, pp. 243-254; D. Delen, R. Sharda, and P. Kumar, "Movie Forecast Guru: A Web-based DSS for Hollywood Managers," *Decision Support Systems*, Vol. 43, No. 4, 2007, pp. 1151-1170.

4.1 数据挖掘概念和定义

1999年1月，在《Computerworld》杂志的一次会议中，Arno Penzias 博士（诺贝尔奖得主，贝尔实验室前首席科学家）提出，在不久的将来，基于组织数据库的数据挖掘将成为一种企业

关键应用。对于《Computerworld》杂志提出的一个由来已久的问题："什么将成为企业的杀手级应用？"Penzias 博士回答："数据挖掘。"然后，他补充说："数据挖掘将比现在重要得多，企业将不会扔掉客户的任何信息，因为客户信息非常宝贵。如果企业不这样做，那么就做不成生意。"类似地，Thomas Davenport 于2006年发表在《Harvard Business Review》的一篇论文中认为，决策分析学是企业的最新战略武器。他给出的例子包括 Amazon.com、Capital One、Marriott International 和其他企业。这些企业应用分析学来更好地理解客户，优化其延伸供应链，从而在为客户提供最佳服务的同时，最大限度地提高其投资回报。决策分析有多成功，在很大程度上依赖于企业对其客户、供应商、业务流程和延伸供应链的理解程度。

这种认识理解的很大成分来自于对企业采集的海量数据所进行的分析。近来，数据存储处理成本的急速降低，带来了电子数据存储量的爆炸式增长。大型数据库的出现使得存储数据分析成为可能。数据挖掘这个词最初用于描述从数据中发现未知模式的过程。之后，一些软件厂商为了利用数据挖掘的噱头促销其产品，对该定义边界进行了扩展，将大多数形式的数据分析都包括到该定义中。在本章中，我们采用数据挖掘的原始定义。

虽然数据挖掘是一个相对比较新的名词，但其背后的思想却由来已久。数据挖掘采用的很多技术根源是传统的统计分析和20世纪80年代早期的人工智能技术。那么，为什么业界现在突然开始关注数据挖掘呢？主要原因如下：

- 客户需求的不断变化和市场需求的日趋饱和带来了全球范围内更激烈的竞争。
- 对于海量数据中隐藏的未开发价值的普遍重视。
- 数据库记录的合并和集成形成了关于客户、供应商和交易的单一视图。
- 数据库和其他数据资料库都合并到单一位置——数据仓库中。
- 数据存储和数据处理技术的指数增长。
- 数据存储和数据处理相关软硬件成本的显著降低。
- 商业行为中的分散化运动（将信息资源转换为非物质形态）。

因特网产生的数据数量和复杂度都在急速增长。全世界正在产生和积累大量的基因组数据。航天和核物理科学定期生成海量数据。医学和药剂学研究者也在不断地产生和存储数据，数据挖掘程序应用这些数据来识别精确诊断和治疗疾病的更好方法，还可以利用这些数据发现新的更好的药。

在商业方面，数据挖掘可能更普遍地应用于金融、零售和卫生保健部门。数据挖掘被用于检测和减少诈骗活动，特别是在保险索赔和信用卡使用中（Chan et al.，1999）；识别客户购买模式；开发有助于获利的客户；从历史数据中发现交易规则；以及应用购物篮分析帮助增加利润。数据挖掘正被广泛应用于更好地定位客户，而随着电子商务的发展，这一定会变得越来越迫切。请看应用案例4.1，了解1-800-Flowers是如何应用业务分析和数据挖掘取得商业成功的。

应用案例4.1 业务分析和数据挖掘帮助1-800-Flowers获取商业成功

1-800-Flowers是礼物馈赠业最为著名和成功的品牌之一。30多年以来，这家设在纽约的公司为世界各地的客户供应适合各种场合的鲜花、植物、礼品篮、精致食品、糖果、毛绒动物玩具。1-800-Flowers由Jim McCann于1976年创立。14年前，在开放了自己的Web网站之后，它迅速成为直接订购电子商务的领袖。

问题

尽管已经非常成功，但1-800-Flowers和其他涉足电子商务的企业一样，需要进行实时决策以增加收益、降低成本，并保留其最好的客户，使这些客户更多地成为回头客。由于该公司的业务已经由一家花店成长为一家拥有3 000万客户的在线礼品零售商，所以它需要做到最好来保持其竞争优势。

解决方案

1-800-Flowers坚信能同客户关系的价值，因而非常希望能够通过分析其拥有的每一条客户数据来更好地理解其客户需求。1-800-Flowers决定应用SAS数据挖掘工具来深入挖掘其数据资产，以发现新的客户行为模式，并利用这种知识促成商业交易。

结果

根据McCann，在业务分析和SAS数据挖掘工具的帮助下，不管整体经济环境如何，1-800-Flowers都能够实现业务增长。当其他零售商在为了生存而苦苦挣扎时，1-800-Flowers在过去5年中却几乎实现了收入翻倍增长。

业务分析带来的好处如下：

- **营销活动更有效率** 通过邮递直销，1-800-Flowers大幅度地减少了其用于客户分类的时间。客户知识管理副主管Aaron Cano说，"过去需要花2～3周时间，而现在只需要2～3天的时间。这就给了我们时间来进行更多的业务分析，从而确保我们发出的商业信息是恰当的。"
- **更少的邮件和更高的回复率** 公司已经能够在显著减少其营销邮件的情况下，保证更高的回复率。同时，对于电视和广播广告更有选择权。
- **更好的客户体验** 当一位回头客登录1-800-Flowers.com时，网站会很快显示该客户可能感兴趣的商品。"如果一位客户通常为其妻子购买郁金香，我们就会为其列出我们最新和最好的郁金香产品，"Cano说。
- **增加重复销售** 公司最好的客户更频繁地重复购买，因为1-800-Flowers了解客户是谁，并且知道他需要什么。公司营造了一种轻松方便的购物体验，并且在接触客户时就完成销售。

通过使用业务分析和数据挖掘，1-800-Flowers减少了其运营费用，使其最好客户的保持率增加到超过80%，吸引了2 000万新客户，并且将总体的重复交易比率从不到40%提高到超过50%（所有品牌的重复交易每增加10个百分点，就意味着增加4 000万美元的额外收入）。

来源：Based on "SAS Helps 1-800-Flowers.com Grow Deep Roots with Customers," sas.com/success/1800flowers.html（accessed on May 23, 2009）; "Data Mining at 1-800-Flowers," kdnuggets.com/news/2009/n10/3i.html（accessed on May 2006, 2009）.

4.1.1 定义、特征和好处

简单地说，**数据挖掘**这个名词是指从海量数据中发现或"挖掘"知识。人们很容易发现数据挖掘实际上用词不当。打个比方来说，从泥土和岩石中挖掘金子称为掘"金"，而不是"泥土"挖掘或"岩石"挖掘。因此，数据挖掘可能应该被称为"知识挖掘"或"知识发现"。尽管这个名称和其实际含义并不匹配，但大家还是选择了数据挖掘这个术语。还有很多其他名词也和数据挖掘有关，包括知识提取、模式分析、数据考古、信息采集、模式搜索和数据捕捞。

严格来说，数据挖掘是一个应用统计学、数学和人工智能技术从大数据集中提取和识别有用信息以及随之而产生的知识的过程。这些模式的表现形式可以是业务规则、类同关系、关联关

系、趋势或预测模型（Nemati and Barko，2001）。大多数文献将数据挖掘定义为“从结构化数据库中识别出合理的、新颖的、可能有用的、并且最终可理解的模式的一个非简单过程。”其中，数据以分类变量、顺序变量和连续变量结构化的记录形式组织（Fayyad et al.，1996）。该定义中的关键术语含义如下：

- 过程表明数据挖掘包括很多迭代步骤。
- 非平凡说明其中涉及一些实验搜索或推导，就像对预定数值进行计算一样明确。
- 合理的含义是，有足够程度的把握认为所发现的模式同样适用于新数据。
- 新颖是指对于所分析的系统，模式是用户此前未知的。
- 可能有用是指所发现的模式应该能够为用户或任务带来一些好处。
- 最终可理解意味着模式应当具有商业意义，不是立刻，但是至少在模式经过后置处理后，用户会由此说：“嗯，很有道理！为什么我没想到呢？”

数据挖掘并非一门新学科，而是一个应用很多学科的新定义。数据挖掘紧密定位于多学科的交叉，包括统计、人工智能、机器学习、管理科学、信息系统和数据库（见图4-2）。数据挖掘应用所有这些学科的进展，在从大型数据库中提取有用信息和知识方面取得进步。这是一个在很短时间内就吸引了诸多关注的新兴领域。

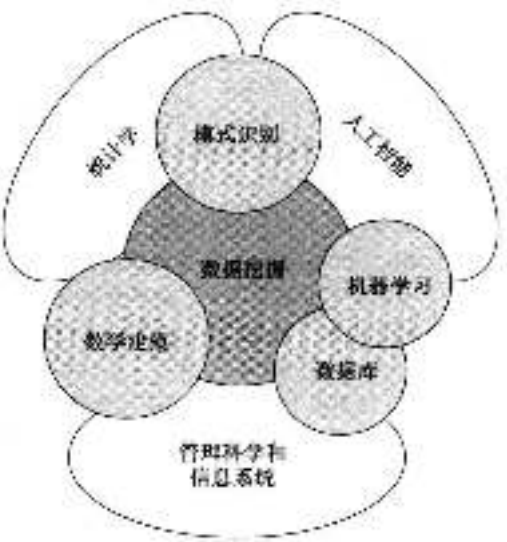


图4-2 数据挖掘——多学科交融的学科

以下是数据挖掘的主要特征和目标：

- 数据往往被深埋在非常大型的数据库中。这些大型数据库往往包含数年的数据。在很多情况下，数据被清洗然后合并到一个数据仓库中。
- 数据挖掘环境通常是一个客户机/服务器架构或一个基于Web的信息系统架构。
- 使用尖端新型工具，包括先进的可视化工具，来帮助移动埋藏于公司文件或公众档案记录中的信息宝藏。这需要对数据进行修改和同步以得到正确的结果。尖端的数据挖掘技术还在探索利用软性数据（即非结构化数据，存储在诸如Lotus Notes数据库、Internet文本文件，或者企业内部网络这样的位置）。
- 进行数据挖掘的人经常是具备很少或根本不具备编程技能的终端用户。利用数据钻探和其他强大的查询工具，用户可以提出特定的问题并迅速得到答案。
- 从数据挖掘中真正获益的过程常常伴随着发现某个意外结果，终端用户在整个过程中的创造性思考，以及对于发现结果的创造性解释。
- 数据挖掘工具很容易和电子制表软件等其他软件开发工具结合，因而使挖掘数据的分析和部署更加容易和快速。
- 由于数据量和搜索工作量都非常大，所以数据挖掘有时需要使用并行处理。

有效运用数据挖掘工具和技术的企业能够获得和保持战略竞争优势。通过将数据转换为一种战略武器，数据挖掘为组织提供了一个对于开拓新商机不可或缺的优化决策环境。Nemati and Barko（2001）对数据挖掘带来的战略利益进行了更详细的讨论。

技术前沿4.1 数据挖掘中的数据

数据是指从经验、观测或实验中得到的一系列事实的汇总。数据可能由作为一组变量测量值的数字，

文字、图像等构成。数据常被认为是从中获取信息和知识的抽象的最低层次。

在抽象的最高层次，可以将数据分为分类数据和数值数据。其中，分类数据又可以再细分为名目数据和序数数据，而数值数据则可以再细分为区间数据和比率数据。图4-3给出了数据挖掘中一个简单的数据分类体系。

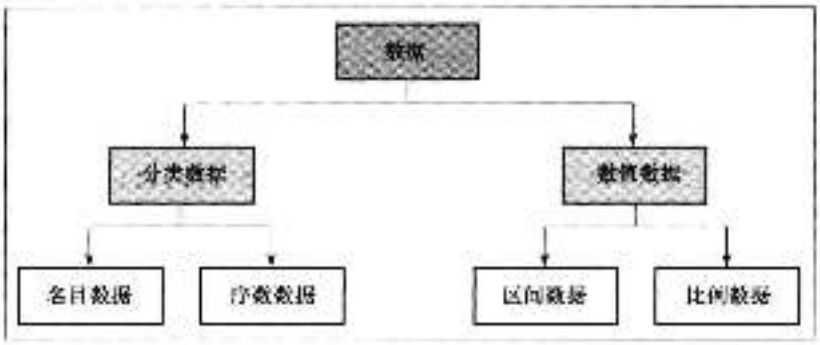


图4-3 数据挖掘中一个简单的数据分类体系

- **分类数据**表示用于将一个变量分为多个特定分组的多类标签。例如，种族、性别、年龄组和教育程度都是类别变量。虽然后两个变量也可以看成是数值变量，例如可以用具体的数值表示年龄，用所完成的最高级别数值表示教育程度，但更有用的往往是将这些变量分类成为相对少量的几个有序组。分类数据也可称为离散数据，即它所表达的是有限数量的非连续值。即使这些分类（离散）变量的值是数值型的，这些数值也仅仅是一些符号，并不能计算其分数值。
- **名目数据**包含为对象分配的简单代码标签，是不可测量的。例如，变量婚姻状况一般可分为（1）单身；（2）已婚；（3）离异。名目数据可能只有2种可能值（如：是/否、真/假、好/坏），也可能有3个或更多个可能值（如：棕色/绿色/黄色、白人/黑人/拉丁美洲人/亚洲人、单身/已婚/离异）。
- **序数数据**包含赋予对象或事件的代码标签，用于表达对象或事件的等级顺序。例如，变量信用评分可以分成（1）低；（2）中等；（3）高。类似的有序关系也可以在诸如"年龄组"（如：儿童、青年、中年、老年）和"教育程度"（如：中学、大学、研究生）等这样的变量中看到。有些数据挖掘算法（例如多元序数逻辑回归）将这种附加等级顺序信息纳入考虑，从而建立更好的分类模型。
- **数值数据**表达特定变量的数量值。年龄、子女数、家庭总收入（以美元计算）、旅行距离（用英里计算）和温度（用华氏温度计算）等都是数值变量。数量值变量可以是整型（只能是整数）或实数（也可以包含分数）。数值数据也可称为连续数据，即这种变量在特定范围内具有连续值，允许存在过渡值。与代表有限可数数据的离散变量不同，连续变量代表的值是可扩展的，可以包含无限个可分割的值。
- **区间数据**是基于区间尺度衡量的变量。区间尺度测量中一个熟为人知的例子是摄氏温度计量，其度量单位是标准大气压下水的熔点和沸点之差的1/100。换句话说，不存在绝对零度。
- **比率数据**包含的测度变量常见于物理学和工程学中。例如，体积、长度、时间、平面角、能量和电荷等都是物理量度中的一些比例尺度。比率类型得名于其测度方式，即其测量单位是一个连续量的总数量和单位数量的比例估计值。通俗地说，比率测度的显著特征是有一个非任意的零值。例如，开尔文温度有一个非任意的零点标志绝对零度，等于摄氏零下273.15度。该零点是非任意的，因为在此温度下构成物质的粒子具有的动能为零。
- **其他数据类型**包括日期/时间、非结构化文本、图像和音频。在应用数据挖掘算法处理这些数据类型以前，需要先将其转换为某种形式的类别数据或数值数据。数据还可以分为静态和动态（时间或时间序列）。

某些数据挖掘方法对所处理的数据类型是挑剔的。不匹配的数据类型将导致错误的模型或者更常见地，导致模型开发过程的终止。例如，一些数据挖掘算法要求所有的输入变量和输出变

量都由数值类型变量表达（例如神经网络、支持向量机、逻辑回归）。使用1—N伪变量，名称变量或序数变量被转换为数值表达。（例如，具有3个单值的分类变量可以被转换为3个具有二进制值1或0的伪变量）。由于该过程会造成变量数目增加，所以应当谨慎使用，特别是对于有大量单值的分类变量来说。

类似地，也有一些数据挖掘算法，例如ID3（一种经典的决策树算法）和粗糙集（一种相对新的规则归纳算法），要求所有变量都由分类变量表达。这些算法的早期版本还要求用户在运行数据处理前，先对数值变量离散化，将其表达为分类变量。好消息是，在广泛使用的一些软件工具中，对这些算法的大多数实现都能够同时接受数值变量和名称变量，而数据处理前的必要转换则在工具内部完成。

应用案例4.2 警察局应用数据挖掘打击犯罪

资源萎缩、线索不够和陈年旧案都加剧了打击犯罪的复杂性。在英国的一家警察局，侦查人员发现这些挑战限制了案件处理量。大量缺乏明确线索的案例，例如入室行窃和车辆盗窃，都缺乏明确的证据，因而在发现新证据前常被搁置一旁。因此，警察局面临的挑战在于确定一种能够轻松快速发现未决犯罪案件的模式和趋势的方法。

警察局的每一个电子案宗文件都记录了小偷的外貌和他们的犯罪手法。虽然，很多缺乏证据的案件过去被搁置了，但警察局现在可以对其重新审查，并比以往更迅速地进行处理。应用PASW Modeler，数据模拟器可以使用两个Kohonen网络对类似的外貌描述和犯罪手法进行聚类，然后联合聚类，检查类似的外貌描述和类似的犯罪手法是否相符。若两者非常匹配，且已知作案者的一个或多个罪行，则未决案件也有可能是同一人所为。

分析小组进一步调查了这些聚类，应用统计方法来查证这些相似点的重要程度。如果聚类结果指出待查作案者可能已经找到，则重新调查其他罪案；如果作案者仍然未知，但聚类结果显示很多案件的作案者是同一人，那么可以结合这些线索来变更案件的优先次序。同时，该警局还对多次累犯的行为进行了分析，以识别出符合其行为模式的案件。该警局希望PASW Modeler能够重新分析旧案，将已知累犯和这些旧案建立联系。

全球警察部门都在应用21世纪的创新科技——数据挖掘技术，提高打击犯罪的技术水平，阻止犯罪活动。在各大数据挖掘解决方案提供商（例如SPSS、SAS、StatSoft和Salford Systems）和咨询公司网站，都可以看到数据挖掘的成功应用故事。

来源："Police Department Fights Crime with SPSS Inc. Technology," spss.com/success/pdf/WMPCS-1208.pdf(accessed on May 25, 2009).

4.1.2 数据挖掘的工作原理

数据挖掘利用现有相关数据建立模型，以识别数据集呈现的属性模式。模型通过数学表示形式（简单的线性关系或复杂的高度非线性关系）来识别对象（例如客户）属性中的模式。其中，一些模式是解释性的（解释属性之间的相互关系），而另一些模式则是预测性的（预测某些属性的未来取值）。一般而言，数据挖掘旨在识别4种主要类型的模式：

1. 关联模式发现普遍共同发生的事物分组。例如，购物篮分析发现顾客常常同时购买啤酒和尿布。

2. 预测模式基于过去已发生的一切对某些事件的未来性质做出判断。例如，预测橄榄球超

级杯赛的获胜者或预报某一天的绝对温度。

3. 聚类基于已知特征识别事物的自然分组。例如，基于客户的人口统计特征和过往购买行为将其划分到不同的分组。

4. 顺序关系模式发现时序事件。例如，对于已经拥有支票账户的现有银行客户，可以预测其将在一年内开立储蓄存款户头，并在随后开立投资户头。

几个世纪以来，人们一直用手工方法从数据中提取模式。但现代数据量的增长产生了对自动化方法的需求。由于数据集的规模和复杂度都在增长，所以直接手工数据分析越来越多地融入采用复杂方法论、方法和算法的间接自动化数据处理工具。目前，通常将这种对于大数据集的自动化和半自动化处理方法的演变称为数据挖掘。

一般而言，数据挖掘任务主要可以分为3大类：预测、关联和聚类。根据其从历史数据中提取模式的方法，又可以将数据挖掘学习算法分为有监督的和无监督的。有监督的学习算法中，训练数据既包括描述属性（即独立变量或决策变量），也包括类属性（即输出变量或结果变量）。相反，无监督的学习算法中，训练数据仅包括描述属性。图4-4描述了一个简单的数据挖掘任务分类体系，以及每种数据挖掘任务的学习方式和基本算法。

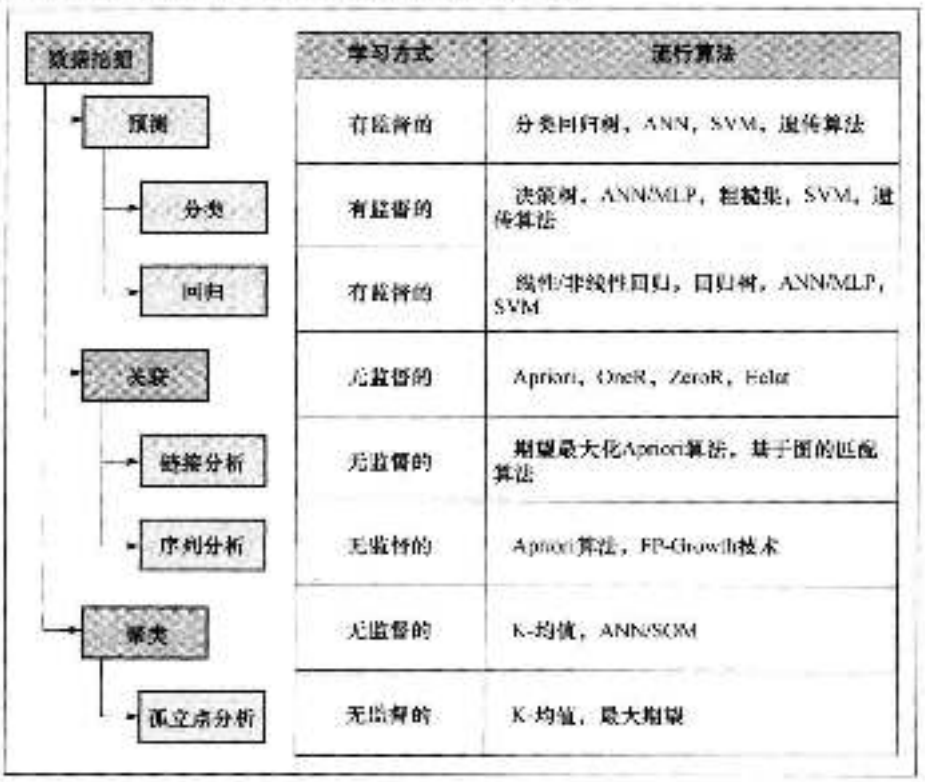


图4-4 数据挖掘任务分类体系

预测 Prediction通常指预告未来的行为。它不同于考虑经验、观点和其他相关信息而进行预言的简单猜测。常与Prediction联系的一个术语是forecasting。虽然很多人认为这两个术语是同义的，但两者之间还是存在细微却很重要的区别。Prediction在很大程度上是基于经验和意见的，而forecasting则基于数据和模型。也就是说，为可靠起见，应当将guessing、predicting和forecasting这些相关术语分别单列。在数据挖掘术语中，prediction和forecasting可以作为同义语使用，通常使用prediction表达预测行为。按照所预测的属性，预测又可具体分为分类预测（例如明日

天气，所预测的是事物类别，可以是“雨天”或“晴天”）和回归分析预测（例如明日气温，所预测的是一个实数，例如“65℉”）。

分类 分类或者称有监督的归纳，可能是最常见的数据挖掘任务。分类的目标在于通过分析数据库中存储的历史数据，自动生成未来行为的预测模型。所得到的模型是对训练数据集记录的概括，能够帮助识别预定义的类。人们希望模型能够用于预测其他未分类记录的类别。同时更重要的是，希望模型能够对未来实际事件的类别进行预测。

常用的分类工具包括神经网络、决策树（起源于机器学习）、逻辑回归和判别分析（起源于传统统计学）。还有一些新兴的分类工具，例如粗糙集、支持向量机和遗传算法。基于统计学的分类技术（例如逻辑回归和判别分析），由于其数据假设（例如独立性和正规性）不切实际而受到批评，所以这限制了其在分类数据挖掘项目中的应用。

神经网络是一种流行的机器学习算法（4.5小节对其进行详细介绍），其开发了一种类似人类大脑生物神经网络的数学结构，能够学习以结构化数据集表达的过往经验。当变量的数目非常多，且变量之间的关系非常复杂而不清晰时，神经网络算法特别有效。神经网络算法也有其劣势。例如，通常很难为神经网络得出的预测结果找到合适的理由。同时，神经网络要求更多的数据训练。不幸的是，当数据量增长时，训练所耗费的时间将呈指数级增长。因此，神经网络不能基于很大的数据集进行训练。这些因素限制了神经网络在数据充足环境中的应用。

决策树算法根据输入变量值将数据分为有限数量的类别。决策树本质上是一种条件语句的层次结构，因此其效率显著高于神经网络。决策树非常适合于分类数据和区间型数据。所以，对于决策树算法中遇到的连续变量，需要先对其进行离散化，也就是将连续变量值转换为取值范围或类别。

规则归纳也是一种相关的分类工具。和决策树算法不同的是，规则归纳算法中的条件语句是直接从训练数据中归纳得出的，而非一种层次结构。还有一些较新的其他技术，如SVM、粗糙集和遗传算法也逐渐加入分类算法的阵营，代表先进的智能系统。

聚类 聚类算法将事物集合（例如对象和结构化数据集中的事件）分割成多个部分（或自然分组），每一部分的内部成员之间共有某些相似特征。和分类不同，聚类中的类别标签是未知的。选定算法根据数据集中的事物特征识别其共同点，然后建立聚类。由于聚类是由启发式算法确定的，而且对于同一数据集，不同的聚类算法可能给出不同的聚类结果，所以在实际应用聚类结果以前，有必要由专家对其进行解释或可能的修正。所识别的合理的聚类结果可以用于对新数据进行分类和解释。

不足为奇的是，聚类算法中也包括优化。聚类的目标在于创造分组，使所生成分组的内部成员之间相似度最高，而不同组成员之间的相似度最低。最常用的聚类技术包括来源于统计学的k-均值算法和来源于机器学习的自组织映射算法，后者是Kohonen在1982年开发的一种独特的神经网络结构。

企业常常应用其数据挖掘系统展开聚类分析，有效地进行市场细分。聚类分析是一种识别物品类别的方法，同一聚类中的物品之间比其他类中的物品具有更多的共同点。这可以用于客户分类，并引导适合的产品销售，在合适的时间、以合适的格式和价格将其卖给特定的细分客户群体。聚类分析还应用于识别事件或对象的自然分组，以帮助识别和描述这些分组的共同特征。应用案例4.3描述了聚类分析是如何与其他数据挖掘技术结合来识别意外事故原因的。

关联 关联，或数据挖掘中的关联规则学习，是旨在从大型数据库中发现变量间有趣关系的一种流行的成熟技术。得益于条码扫描仪等自动信息采集技术，使用关联规则算法从超市的销售点（POS）系统的大规模交易记录中发现产品规律，在零售业中已经是很平常的知识发现任务。在零售业中的关联规则挖掘通常称为购物篮分析。

从关联规则挖掘中衍生了两种广为使用的工具：**链接分析和序列挖掘**。链接分析能够自动发现众多研究对象之间的联系，如Web网页之间的联系和学术著作作者群体之间的引用关系。

序列挖掘则通过分析事件的发生顺序来识别事件之间的时间关系。关联规则挖掘中使用的算法包括流行的 Apriori（识别频繁项目集）、FP-Growth、OneR、ZeroR 和 Eclat。

可视化和时间序列预测 可视化和时间序列预测是与数据挖掘相关的两种技术。将可视化技术和其他数据挖掘技术相结合，可以获得对挖掘结果更清晰的理解。时间序列预测利用一段时间中获取存储的同一变量的系列取值建立模型，并推测同一现象的未来取值。

假设驱动型数据挖掘和发现驱动型数据挖掘 数据挖掘可以是假设驱动的，也可以是发现驱动的。假设驱动型数据挖掘从用户提出的某一命题开始，然后对该命题真实性进行求证。例如，一位营销经理可能从提出如下命题开始数据挖掘："DVD 播放机的销售和电视机的销售有关联吗？"

发现驱动型数据挖掘 发现隐含在数据集中的模式、关联和其他关系。这种数据挖掘可以发现组织预先未知、甚至未曾想过的事实。

4.1 节复习题

1. 给出数据挖掘的定义。为什么数据挖掘有很多不同的名称和定义？
2. 近来有哪些因素促进了数据挖掘的广泛应用？
3. 数据挖掘是一门新学科吗？请给出解释。
4. 请给出主要的数据挖掘方式和算法。
5. 请说明主要数据挖掘方式之间的关键区别。

应用案例4.3 机动车事故与司机分心

驾驶者集中注意力对于公路交通安全至关重要。美国国家公路交通安全管理局于1996年发布的一项研究结果表明，大约25%～30%的车祸伤害是由于驾驶者注意力不集中造成的。1999年，由美国国家统计分析中心开发的致祸分析报告系统（Fatality Analysis Reporting System，FARS）显示，司机分心造成了11%的致命车祸（死亡4 462人）。

一项旨在从交通事故中提取司机分心因素模式的研究已经展开。该研究利用数据挖掘技术从FARS提供的车祸数据中提取出各种分心因素之间的关联关系。使用了3种数据挖掘技术（Kohonen神经网络、决策树和多层感应神经网络）来发现与高事故率相关并可能对其做出解释的不同种分心因素组合。Kohonen神经网络识别自然聚类，并发现数据集中输入变量的模式。决策树探讨了连续事件中每一事故的作用并对其进行分类，同时也提示了驾驶员分心和身体/精神状况之间的关系。最后，对多层感应神经网络模型进行训练和测试，以找到交通事故中驾驶员注意力不集中和其他驾驶员相关因素之间的关系。使用SPSS中的Clementine对FARS数据库中获取的数据进行3种模型的挖掘。

预测分析模型识别出由于注意力不集中导致车祸发生的事故中的1 255名驾驶员。车尾追撞、正面相撞、翻滚相撞等，都是对车祸发生和其严重程度有重要影响的因素。

来源：Based on W. S. Tseng, H. Nguyen, J. Lichowitz, and W. Agresti, "Distractions and Motor Vehicle Accidents: Data Mining Application on Fatality Analysis Reporting System (FARS) Data Files," *Industrial Management & Data Systems*, Vol. 105, No. 9, January 2005, pp. 1188 – 1205; and J. Lichowitz, "New Trends in Intelligent Systems," Presentation made at University of Granada, doctor-si.ugr.es/seminario2006/presentaciones/jay.ppt (accessed May 2009).

4.2 数据挖掘应用

数据挖掘已经成为一种颇受欢迎的复杂商业问题解决工具。在很多领域中，数据挖掘应用

被证明非常成功有效，以下列出了其中一些代表实例。很多数据挖掘商业应用的目标在于解决当前迫切问题，或者寻求新兴商业机会以形成可持续竞争优势。

- **客户关系管理** 客户关系管理（Customer Relationship Management，CRM）是传统营销的新兴扩展。CRM的目标在于通过对客户需求的深入理解，和客户建立一对一关系。由于客户关系是在各种交易过程（例如产品咨询、销售、服务请求和保修电话）中长期形成的，所以积累的数据量非常巨大。结合人口统计和社会经济属性，这些包含丰富信息的数据可以用于（1）识别最可能购买或响应新产品和服务的客户（即客户概况描述）；（2）理解客户流失的根本原因以改善客户维系（即流失分析）；（3）发现产品和服务的时变关联以最大限度地提高销售额和获取客户价值；（4）识别利润最高的客户及其需求偏好，从而加强客户关系，提高销售额。
- **银行业** 数据挖掘在银行业的作用如下：（1）通过精确预测最有可能的欠款者，实现借贷流程自动化；（2）检测信用卡和网上银行交易欺诈；（3）通过向客户提供其最有可能购买的产品和服务，识别客户价值最大化的途径；（4）通过精确预测银行实体（例如ATM机、银行网点）的现金流，优化现金回报。
- **零售业和物流** 在零售业，数据挖掘可用于（1）精确预测特定零售网点的销售额以确定正确的库存水平；（2）通过购物篮分析，识别不同产品之间的销售关系，从而改善商店布局、优化商品促销；（3）基于季节和环境条件，预测不同类型产品的消费水平，以优化物流并最大限度地提高销售额；（4）通过分析RFID等传感数据，发现产品（特别是易腐、易受污染，保存期有限的产品）流通过程中的有趣模式。
- **制造生产** 数据挖掘在制造业可用于（1）利用传感数据预测机械故障，实现基于条件的维修保养；（2）识别生产系统异常和共性，以优化生产能力；（3）发现改进产品质量的新模式。
- **证券交易** 证券经纪人应用数据挖掘（1）预测特定债券价格何时变动，变动多少；（2）预测股票波动的范围和方向；（3）评估特定问题或事件对于市场整体的影响；（4）识别并防止证券交易中的欺诈行为。
- **保险** 保险业应用数据挖掘技术（1）预测财产索赔额和医疗保险费用，以更好地进行保险业务规划；（2）基于客户和索赔数据分析，优化费率计划；（3）预测哪些客户更有可能购买新推出的特色保险；（4）识别并防止发生不合理的理赔付款和欺诈。
- **计算机软硬件** 数据挖掘可用于（1）及早预测磁盘故障；（2）识别并过滤不受欢迎的Web内容和电子邮件信息；（3）检测并阻止计算机网络安全隐患；（4）识别可能不安全的软件产品。
- **政府和国防部门** 在军事领域，数据挖掘也有很多应用。（1）预测军事人员和设备的流动成本；（2）预测对手行动以制定更成功的军事交往战略；（3）预测资源消耗以支持更好的规划和预算；（4）识别军事行动中的独特经验、策略和教训，以实现组织中更好的知识共享。
- **旅游业（航空、酒店/度假村、租车公司）** 数据挖掘在旅游业有很多成功应用。（1）预测各种不同服务（飞机的座位类型、酒店/度假村的房间类型、租车公司的车辆类型）的销售额，从而为随时间变化的交易提供最佳的价格服务，使收益最大化（即通常所说的收益管理）；（2）进行不同地点的需求预测，以更好地配置有限的组织资源；（3）识别利润最高的客户并为其提供个性化服务，以保持回头客；（4）识别员工流失的根本原因，并采取有针对的行动以保持有价值的员工。
- **卫生保健** 数据挖掘有很多卫生保健应用，其可用于（1）识别没有医疗保险的人群以

及造成这种不良情况的因素；（2）识别不同疗法之间的成本收益关系，以制定更有效的策略；（3）预测不同服务网点的需求水平和需求时间，以优化组织资源配置；（4）理解客户和员工流失的深层原因。

- **医学** 医学中的数据挖掘应用可以视为传统医学研究（本质上主要是临床和生物）的宝贵补充。数据挖掘分析可以（1）识别新模式以改进癌症患者的存活率；（2）预测器官移植患者的成功率，以更好地制定器官捐赠匹配策略；（3）识别人类染色体（被称为基因组）中不同基因的功能；（4）发现疾病和症状以及成功治疗手段之间的关系，以帮助医疗人员及时可靠地进行正确决策。
- **娱乐业** 在娱乐业，数据挖掘被成功地应用于（1）通过电视观众数据分析，确定黄金时段播放什么电视节目，以及何时插播广告以使收益最大化；（2）在电影制作以前，预测其财务表现，以帮助投资决策，优化收入回报；（3）预测不同时间和地点的需求，以更好地制定娱乐事件时间表和优化资源配置；（4）制定最佳定价策略，使收益最大化。
- **国土安全和执法** 在国土安全和执法领域也有很多数据挖掘应用。数据挖掘经常被用于（1）识别恐怖分子行为模式（最近的一个例子是应用数据挖掘进行恐怖分子活动资金追踪，参见应用案例4.4）；（2）发现犯罪模式（例如地点、事件、犯罪行为和其他相关属性），帮助及时破案；（3）通过分析特殊用途的传感数据，预测并消除对国家关键基础设施可能的生化攻击；（4）识别并阻止对关键信息基础设施的恶意攻击（常被称为信息战）。
- **运动** 在美国，数据挖掘被用于改善NBA的球队表现。NBA开发了一个基于PC的数据挖掘应用——Advanced Scout，给教练组人员使用，帮助发现篮球赛数据中的有趣模式。将这些模式和录像带关联，能够更好地解释这些模式数据。详见Bhandari et al.（1997）文献。

应用案例4.4 恐怖组织筹资活动挖掘

2001年发生在世贸中心的"9·11"恐怖袭击事件，强调了公开来源情报的重要性。美国爱国者法案和国土安全部门的成立预示了信息技术和数据挖掘技术在检测洗钱和其他形式的恐怖分子筹资活动中的潜在应用。一直以来，执法部门只关注通过银行和其他金融服务机构正常交易中的洗钱活动。

现在，利用国际贸易价格作为恐怖活动筹资工具，已经成为执法部门的关注焦点。利用国际贸易，洗钱者可以从一个国家悄悄转出钱，而不被政府注意。这主要是通过高估进口货价格，而低估出口货价格实现。例如，国内的进口商可以和国外出口商合伙，对进口货过高估价，从而将钱从祖国转移，进行海关欺诈、偷漏所得税和洗钱犯罪。其中的国外出口商可能是恐怖组织成员。

数据挖掘技术主要对美国商务部和其他商业相关实体的进出口交易数据进行分析。超出上限的进口价格和低于下限的出口价格数据都被追踪，而关注点主要在于公司之间不正常的转让价格造成的应纳税收入转移和偷税漏税。这种价格差异可能和偷漏所得税、洗钱，或恐怖分子筹资等相关。当然，贸易数据库错误也可能导致价格差异。

数据挖掘将提高数据评估的效率。反过来，这也有助于和恐怖分子的斗争。信息技术和数据挖掘技术在金融交易中的应用将造就更有用的情报资料。

来源：Based on J. S. Zdanowic, "Detecting Money Laundering and Terrorist Financing via Data Mining," *Communications of the ACM*, Vol. 47, No. 5, May 2004, p. 53; and R. J. Bolton, "Statistical Fraud Detection: A Review," *Statistical Science*, Vol. 17, No. 3, January 2002, p. 235.

4.2 节复习题

1. 数据挖掘的主要应用领域有哪些？
2. 说出至少5个数据挖掘具体应用，并列出这些应用的5点共同特征。
3. 你认为当前最重要的数据挖掘应用领域是什么？为什么？
4. 你能说出本节中未提及的其他数据挖掘应用领域吗？请解释。

4.3 数据挖掘流程

数据挖掘项目的系统实施通常要求遵循一般的流程。基于这方面的最佳实践，数据挖掘研究人员和从业人员提出了工作流或简单的逐步方法等流程，以最大限度地提高数据挖掘项目实施的成功率。其中一些流程已经被标准化，以下列出了其中应用最广泛的几种流程。

数据挖掘跨行业标准流程（Cross-Industry Standard Process for Data Mining CRISP-DM）可以认为是应用最为广泛的一种标准化数据挖掘流程。该标准于20世纪90年代中期由一个欧洲企业联盟提出，目的是建立一个数据挖掘的非专用标准方法（CRISP-DM，2009）。图4-5描述了该流程，包括6个步骤，以业务和数据挖掘项目（即应用领域）需求的深入理解开始，以能够满足特定业务需求的解决方案部署结束。尽管这些步骤是顺序进行的，但通常仍然存在诸多回溯。由于数据挖掘由经验和实验驱动，所以依赖于问题的实际状况和分析者的知识经验，导致整个数据挖掘流程可能非常迂回（即可能存在步骤之间的多次来回反复）和耗时。因为后续步骤是建立在前面步骤的结果基础上，所以应当特别注意前面的步骤，避免整个研究从一开始就进入错误路径。

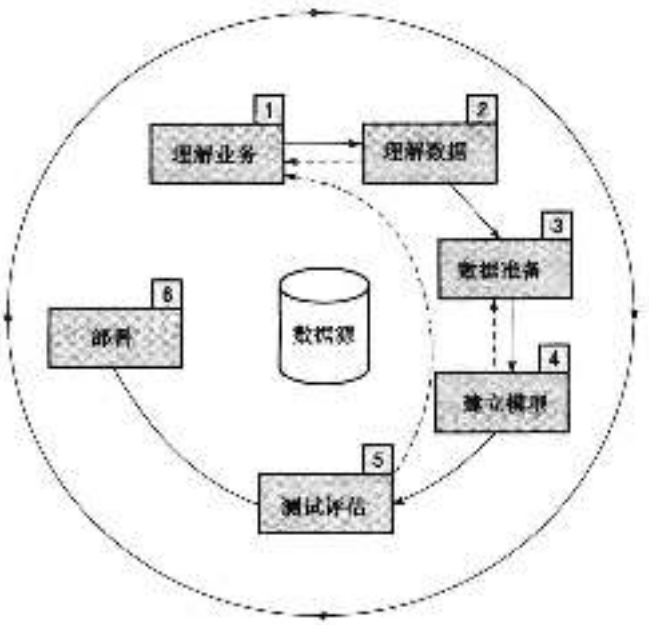


图4-5 CRISP-DM 6阶段数据挖掘流程
来源：转自 CRIP-DM. org.

4.3.1 步骤1：理解业务

理解研究目的是任何数据挖掘研究的关键要素。要回答这个问题，首先要透彻理解对新知识的管理需求，并清晰定义将要进行研究的商业目标。类似的问题，如“最近我们流失到竞争对手的客户具备哪些共同特征？”或者“我们客户的典型概况是怎样的？每位客户能为我们创造多少价值？”都需要解决。接下来，应当制定发现这些知识的项目计划，分别指定负责收集数据、分析数据和报告结果的负责人。在这样的早期阶段，至少应当在较高层次建立具有大概数目的研究预算。

4.3.2 步骤2：理解数据

数据挖掘研究主要是处理明确的特定业务任务，而不同的业务任务需要不同的数据集合。在理解业务以后，数据挖掘过程的主要活动是从很多可用数据库中识别相关数据。在数据辨识和选择阶段，需要考虑很多关键要点。首要的是，分析者应当清晰简洁地描述数据挖掘任务，以

识别出最为相关的数据。例如，一个零售业的数据挖掘项目可能想要识别购买应季服装的女性客户消费行为，依据是人口统计资料、信用卡交易记录和社会经济属性。此外，分析者应当深入理解各个数据源（例如，相关数据存储在哪里？以何种形式存储？数据收集过程是人工的还是自动的？谁收集了这些数据？数据多长时间更新一次？）和各个变量。（例如，最相关的变量有哪些？是否存在同义或同音不同义的变量？各个变量彼此之间是否独立？变量之间是否存在重叠或冲突信息？）

为了更好地理解数据，分析者经常利用各种统计技术和图形技术，包括对每个变量的简单统计摘要（例如，对于数值变量，可以计算其平均值、最大值、最小值、中间数、标准差；对于分类变量，可以计算其模式和频率表）、相关分析、散点图、直方图和方块图。应当仔细辨识这些数据源和相关变量。这样，数据挖掘算法更容易快速发现有用的知识模式。

数据挖掘的数据源可能是多样的。通常，商业应用的数据源包括人口统计数据（如收入、教育程度、家庭人口和年龄）、社会学数据（如爱好、俱乐部会籍和娱乐）、交易数据（销售记录、信用卡消费、签发支票）等。

数据可以是定性的或定量的。定量数据使用数值进行度量，分为离散型（如整数）和连续型（如实数）。定性数据，或称分类数据，包括名目数据和序数数据。名目数据具有有限个无序值（例如性别有2个可能值：男或女）。序数数据具有有限个有序值。例如，客户信用等级是序数数据，其值可以是优秀、一般和不好。

定量数据很容易由某种概率分布表达。概率分布描述了数据的分散形态。例如，正态分布数据是对称的，通常指的是一个钟形曲线。定性数据可以用数字编码，然后由频率分布描述。一旦已经按照数据挖掘业务需求完成了相关数据源选择，就应当继续进行数据预处理了。

4.3.3 步骤3：数据准备

数据准备也常被称为数据预处理，其目的是处理前一阶段识别的数据，为后面的数据挖掘算法分析做好准备。和CRISP-DM中的其他步骤相比，数据预处理最为耗时费力。普遍认为，这一步大约耗费整个数据挖掘项目80%的时间。造成这一步骤耗费如此巨大的原因是，现实世界中的数据一般不完整（缺失属性值、缺少所感兴趣的某些属性或者仅包含汇总数据），有噪声（包含错误或离群值）、不一致（代码或名称存在差异）。如图4-6所示，要将现实世界的原始数据转换为可挖掘的数据集，主要需要完成4个步骤。

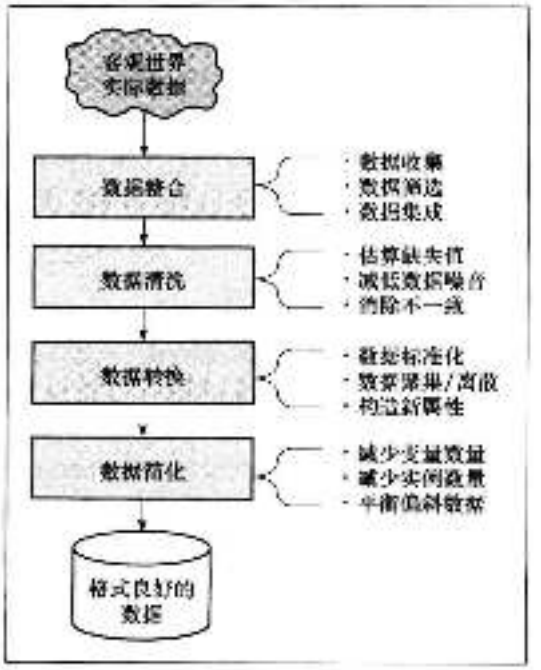


图4-6 数据预处理步骤

在数据预处理的第一阶段，首先从前面的步骤（即CRISP-DM过程中的数据理解过程）所识别出的数据源中收集相关数据。然后基于对数据的深入理解过滤掉不必要的部分，筛选出必要的记录和变量；最后仍然是在深入理解数据的基础上，恰当处理同音不同义和同义不同名的情况，将多个数据源的数据记录进行集成。

数据预处理的第二阶段是数据清洗，这一步骤识别并处理数据集的值。在有些情况下，数据集中的缺失值是不正常的，需要对其进行估算，

域上其最可能的值，或者简单忽略该值；在其他一些情况下，数据集中某些部分的缺失值是正常的。例如，处于收入最高层的人经常不填写家庭收入一栏。在这一阶段，分析者还要识别并消除数据中的噪声值（极端值）。此外，应当利用领域知识或专家意见对数据不一致（变量的离群值）进行处理。

在数据预处理的第三阶段进行数据转换，以便更好地完成数据处理。例如，在很多情况下，所有变量都被标准化到一个特定的最大值/最小值区间中，以减轻可能存在的个别大数值变量（例如家庭收入）对其他小数值变量（例如赡养人数或服务年限，可能更为重要）的支配偏向。在这一阶段还要进行数据离散或（和）数据聚集。有些时候，要将数值变量转换为分类变量（例如，低、中、高）；还有些时候，利用概念分层将名称变量的独特值域缩减到较小的集合中，以使数据集更适合计算机处理。例如，对于一个指示位置的变量，人们可能不愿意用50个不同州的名称，而希望使用几个大的区域作为变量值。还有一些其他情况，人们可能会根据已有变量生成新的变量，以使从数据集的变量集合中得到的信息量更大。例如，在器官移植数据库中，人们可能选择使用一个单独的变量来表示血型匹配（1：匹配；0：不匹配），而不是使用多项值将捐助者和接受者的血型值分开表示。这种简化操作一方面降低了数据关系的复杂度，另一方面增大了信息量。

数据预处理的最后一阶段是数据简化。尽管数据挖掘者往往希望拥有更大的数据集，但过多的数据也会成为问题。简单来说，可以将数据挖掘项目中通常使用的数据可视化为一个二维平面文件，包括变量（列数）和实例/记录（行数）。在有些情况下，例如图像处理和含有复杂效应矩阵数据的基因组项目，可能拥有相当多的变量，分析者必须将变量数量缩减到可控大小。由于变量被视为从不同角度描述现象的不同维度，所以在数据挖掘中，这一过程通常被称为维度缩减。虽然并不存在完成此任务的单一良方，但还是可以利用先前发表的文献发现、咨询领域专家、运行恰当的统计测试（例如主成分分析或独立成分分析），帮助完成维度缩减。当然，最好是综合应用这些技术，成功地将数据维度缩减到更可控、更相关的数据子集。

对另一维度（即实例数量）而言，有些数据集可能包括数以百万甚至十亿计的记录。即使计算能力在以指数级增长，但处理如此巨大数量的记录仍然是不切实际或不可行的。在这种情况下，人们可能需要对数据集进行采样以供分析。数据采样的基本假设是数据样本将包含完整数据集中的所有相关模式。对于同质数据集，可能可以做这种假设，但真实世界中的数据很少是同质的。分析者在筛选数据子集时应当非常慎重，使其能反映完整数据集的本质，而非局限于特定的子类数据。数据通常按某些变量进行排序，所以从上到下选取的数据段可能导致数据子集对索引变量的特定值存在偏见。因此，永远应当随机选取样本集合记录。对于偏斜数据，直接随机取样可能不够，需要分层取样，即按一定比例从不同的数据子类中选取样本数据。关于偏斜数据，很好的一个做法是对表达较少的数据进行超取样，或对表达较多的子类进行欠采样，以平衡高度偏斜数据。已有研究表明，平衡数据集生成的预测模型往往优于不平衡的数据集（Wilson and Sharda，1994）。

表4-4总结了数据预处理的实质，描述了数据预处理每一阶段中的问题、任务和常用算法。

表4-4 数据预处理各阶段任务和处理方法一览

主要任务	子任务	常用方法
数据整合	访问和收集数据	SQL查询、软件代理、Web服务
	筛选和过滤数据	领域专家、SQL查询、统计测试
	综合和统一数据	SQL查询、领域专家、本体驱动的数据映射

（续）

主要任务	子任务	常用方法
数据清洗	处理数据缺失值	用合适值（平均数、中间值、最小值/最大值、模式等）填充缺失值；用常量，例如“MIS”对缺失值重新编码；剔除含缺失值的记录；不处理
	识别并降低数据噪声	使用统计技术（例如平均值或标准差）或聚类分析技术识别数据极端值；删除或采用分级、回归或简单平均等方法消除发现的强噪值
	发现并消除错误数据	识别错误数据值（非极端值），例如奇怪的值、不一致的分类标签和奇怪的分布；在识别错误数据后，借助领域专家修正错误，或剔除含错误值的记录
数据转换	数据标准化	采用各种标准化或计算技术将各数值型变量的取值范围缩小到一个标准范围
	离散或聚集数据	必要时采用基于元图或级别的分级技术，将数值型变量转换成离散形式；通过恰当应用层次概念减少分类变量的取值数量
	构造新属性	广泛应用数学函数（简单的如加法和乘法，复杂的如欧氏变换的多变量组合），由已知变量推导出更有启发性的新变量
数据简化	减少属性数量	主成分分析、独立成分分析、卡方检验、相关分析和决策树归纳
	减少记录数量	随机取样，分层取样，基于专家知识的有目的取样
	平衡偏倚数据	对代表较少的实例超取样，或者对代表较多的实例少取样

4.3.4 步骤4：建模

这一步骤将从各种建模技术中进行选择，并将其应用到已经准备好的数据集，以解决具体商业需求。建模过程还包括对各种模型的评估和比较分析。由于并不存在公认“最好的”数据挖掘方法或算法，所以应当利用各种可行模型，并采用清晰定义的实验和评估策略识别出对于给定目标“最好的”方法。即使是对单个方法或算法，也需要进行很多参数调整才能获得最优的结果。有些方法可能对数据格式有特殊要求，因此从这里返回数据准备阶段也是经常有必要的。

依据不同的商业需求，数据挖掘任务可以分成3种类型：预测（分类或回归）、关联或聚类。在执行各种数据挖掘任务时，都可能用到多种数据挖掘方法和算法。本章前面讲述了其中一些数据挖掘方法，还有一些广为应用的算法，包括分类决策树、k－均值聚类和用于挖掘关联规则的Apriori算法，都将在本章后续部分讲到。

4.3.5 步骤5：测试和评估

第5步是对所建立模型的精确性和一般性进行评估。在这一阶段，确定所选用的模型是否能满足业务目标。如果认为模型能满足目标，那么还要评估其能在多大程度上满足目标，也就是要判断是否有必要建立和评估更多的模型。如果时间和预算允许，还可以考虑在真实世界背景中测试模型。尽管根据预期所建立的模型结果和最初的业务目标应当是相关的，但也会有一些其他发现虽然不一定和最初目标相关，却也可能揭示一些额外信息，或者对未来的研究方向有所启发。

测试和评估阶段的任务很关键也很有挑战性。除非能够识别分辨由所发现的知识模式所带来的商业价值，否则数据挖掘工作就没有增加价值。确定知识模式的商业价值有些类似于玩拼图。所提取出的知识模式就像拼图中的板块，需要按照具体业务目的将其组装在一起。辨识能否成功，有赖于数据分析员、业务分析员和决策者（如业务经理）之间的相互作用。数据分析员可能对数据挖掘的目标及其对于业务的意义缺乏整体理解，而业务分析员和决策者则可能缺乏解释复杂数学方案的

技术知识，因此二者之间的交互是非常必要的。要很好地解释知识模式，常常需要应用各种图表和可视化技术（例如数据透视表、结果交叉制表、饼状图、直方图、方块图、散点图）。

4.3.6 步骤6：部署

建立模型并评估并不说明数据挖掘项目已经结束。即使模型的目的只是简单了解数据，也需要对研究过程所获取的知识进行组织，并以终端用户能够理解和受益的方式表达这些知识。根据需求不同，部署过程可能很简单，例如生成报告；也可能很复杂，例如可重复的跨企业数据挖掘实施过程。在很多情况下，部署步骤是由客户而非数据分析员进行。然而，即使分析员并不实现部署工作，但客户也需要预先理解其需要执行的操作，才能真正利用所生成的模型。

部署阶段还可能包括对所部署模型的维护工作。由于业务总在变化，业务活动数据也一直在改变。随着时间的推移，基于过往旧数据的模型及其内置模式可能变得过时，不相干或者引发误解。因此，如果要把数据挖掘结果作为日常业务环境中的一部分，那么对模型的监控和维护就非常重要。认真预备维护策略有助于避免不必要地长期错误使用数据挖掘结果。为了监控数据挖掘结果的部署，项目需要制定详细的监控过程计划，对于复杂的数据挖掘模型，这可能是一项艰巨的任务。

应用案例4.5 癌症研究中的数据挖掘

据美国癌症协会的数据，2009年新增癌症病例大约达到150万例。癌症是美国乃至全世界的第2大常见死因，仅次于心血管疾病。2010年，预计美国将有562 340人死于癌症，平均每天超过1 500人，几乎占到死亡人数的四分之一。

癌症是一组疾病，其一般特征表现为异常细胞不可控制的生长和扩散。如果癌症的生长扩散不能得到有效控制，那么就会导致死亡。尽管确切病因不详，但一般认为癌症是由外部因素（例如，吸烟、器官感染、化学物质、辐射）和内部因素（例如，遗传突变、激素、免疫疾病、由代谢导致的突变）共同导致的。这些影响因素可能同时或顺序作用导致癌症引发或恶化。当前，癌症的治疗方法有手术、放射线治疗、化疗、激素疗法、生物疗法、目标疗法。癌症的存活率随不同的种类和诊断期而有很大差别。

癌症总体的5年相对存活率已经从1975年到1977年的50%，上升到1996年到2004年的66%。存活率的上升反映了癌症早期诊断的发展和癌症治疗手段的进步。癌症防治还需要进一步完善加强。

传统的癌症研究在性质上属于临床和生物领域，但近些年来也常常采用数据驱动的分析研究作为补充。在已经成功应用数据和分析驱动研究的医学领域，新的研究方向也已经产生，用以促进临床和生物研究的进步。利用各种类型的数据，包括分子、临床、文献数据、临床实验数据，同时应用合适的数据挖掘工具和技术，研究者已经能够识别出新模式，从而为战胜癌症成为无癌社会奠定了基础。

Delen在2009年的一项研究中，采用三种常用数据挖掘技术（决策树、人工神经网络ANN、支持向量机），并结合逻辑回归方法，建立了前列腺癌存活率预测模型。数据集包含大约120 000条记录和77个变量，同时，应用k折交叉确认方法进行模型的建立、评估和比较。结果表明，该领域中预测最为精确的是支持向量模型（测试集精确度为92.85%），其次是人工神经网络和决策树。此外，应用敏感度分析评估方法的研究结果还揭示了前列腺癌预后因素的相关新模式。

2006年，Delen的一项相关研究针对一个包含超过200 000例病例的大规模数据集，应用两种数据挖掘算法（人工神经网络和决策树），结合逻辑回归方法建立了乳腺癌存活状况的预测模型。研究采用一种10折交叉确认方法进行预测模型的无偏估计测量，进行模型性能比较，结果认为决策树（C5算法）预测最为精确，测试组样本精确度达到93.6%，是文献报道中预测精度值最高的；其次是人工神经网络，精确度为91.2%；最后是逻辑回归，精确度为89.2%。对预测模型的进一步分析表明预后因素相当重要，可以作为进一步临床和生物研究的基础。

这些实例（以及医学文献中的很多其他实例）说明，先进的数据挖掘技术能够用于建立具有高度预测和解释能力的模型。虽然数据挖掘方法能够挖掘出深度隐藏在大型复杂医疗数据库中的模式和关系，但如果没有医学专家的合作和反馈，这些结果是没有多大用处的。通过数据挖掘方法发现的模式应当由具备数年相关问题领域经验的医学专家进行评估，以确定其是否合乎逻辑，是否可行，以及是否新颖而可作为新的研究方向依据。简而言之，数据挖掘不是要取代医学专家和研究人员，而是要让数据驱动的新研究方向与他们的重要工作相辅相成，最终拯救更多的人类生命。

来源：D. Delen, "Analysis of Cancer Data: A Data Mining Approach," *Expert Systems*, Vol. 26, No. 1, 2009, pp. 100-112; I. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space," *Expert Systems with Applications*, 2009, in press; D. Delen, G. Walker, and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," *Artificial Intelligence in Medicine*, Vol. 34, No. 2, 2005, pp. 113-127.

4.3.7 其他标准化数据挖掘过程和方法

一项成功应用的数据挖掘研究必须遵循标准方法，而不是将一组自动化软件工具技术无序组合。除了CRISP-DM以外，由美国SAS软件研究所于2009年开发的SEMMA也是一种广为人知的方法。缩略字SEMMA代表"抽样（Sample）、探索（Explore）、修正（Modify）、建模（Model）和分析（Assess）"。

SEMMA从数据的统计代表取样开始，可以很容易地应用统计探测和可视化技术，选取转换最重要的预测变量，建立变量预测结果模型，并加强模型的精确度。图4-7给出了SEMMA的图形表示。

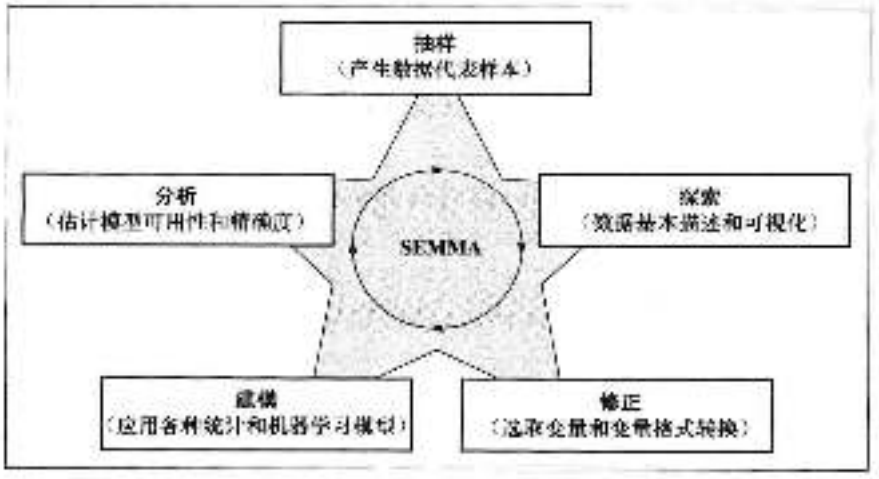


图4-7 SEMMA数据挖掘过程

通过对SEMMA过程中的每一阶段进行结果评估，建模人员能够针对前面结果所提出的新问题，决定如何建立模型，并接着返回调研阶段进一步优化数据。也就是说，SEMMA和CRISP-DM同样是由高度迭代的实验循环驱动的。CRISP-DM和SEMMA之间的主要区别在于，CRISP-DM采用的方法更为全面，包括对数据挖掘项目相关业务和数据的理解。而SEMMA则隐含假设数据挖掘项目的目标及其恰当数据源已经被识别和理解。

业界也通常使用术语数据库中的知识发现（Knowledge Discovery in Databases，KDD）指代数据挖掘。Fayyad等在1996年将数据库中的知识发现定义为应用数据挖掘方法从数据中发现有用信息和模式的过程。与数据挖掘形成对照的是，KDD应用算法从KDD过程获得数据中的识别模式。KDD是一个涵盖数据挖掘的综合过程。KDD过程的输入包括组织数据。企业数据仓库可以提供一个单一的挖掘数据源，因而能够促进KDD的实施效率。Dunham在2003年将KDD过程概括为如下步骤：数据筛选、数据预处理、数据转换、数据挖掘和解释/评估。Kdnuggets.com网站于2007年8月就问题"你主要使用什么数据挖掘方法？"进行了一项调查，图4-8显示了投票结果。

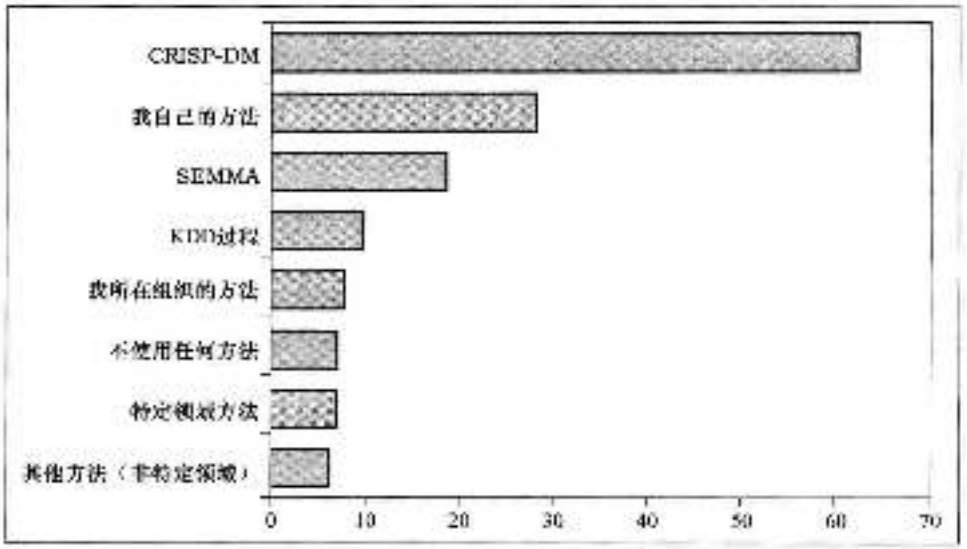


图4-8 数据挖掘方法/过程排名
来源：kdnuggets.com，经许可使用。

4.3 节复习题

1. 有哪些主要的数据挖掘过程？
2. 为什么数据挖掘项目中的早期阶段（例如业务和数据理解阶段）会占据最多的时间？
3. 列举并简单定义CRISP-DM过程中的各阶段。
4. 数据预处理主要包括哪些步骤？简单描述各步骤并举出相关例子。
5. CRISP-DM和SEMMA有哪些不同？

4.4 数据挖掘方法

进行数据挖掘研究的方法有很多种，包括分类、回归、聚类和关联。大多数数据挖掘软件工具对每一种方法都使用多种技术或算法。本节描述了使用最为广泛的数据挖掘方法，并对其代表技术进行了解释。

4.4.1 分类

分类可能是解决现实世界问题时使用最为频繁的数据挖掘技术。作为机器学习技术家族中

最流行的成员。分类从历史数据（预先标记项目、对象或事件的特征信息集合——特性、变量、特征）中学习模式从而将新的实例放到其各自的分组或分类中。例如，可以用分类预测某天的天气是“晴”、“雨”或者“多云”。常见的分类任务包括信贷审批（即信用风险高或低）、商品选址（例如，好、中等、差）、目标营销（例如，可能成为客户、不可能成为客户）、欺诈检测（即是、否）和电信（例如，是否可能转到其他电话公司）。如果所预测的是一个类别标签（例如，“晴”、“雨”或者“多云”），则该预测问题称为分类；而如果所预测的是一个数量值（例如，气温68°F），则该预测问题称为回归。

尽管聚类（另一种常用数据挖掘方法）也可用于确定事物的分组（或者类别属性），但两者之间仍然存在显著区别。分类通过一个监督学习过程学习事物之间的功能特征（即自变量）及其类别属性（即输出变量）。其中，输入变量和输出变量都被提交给分类算法；在聚类中，通过一个无监督学习过程学习对象的分类，提交给聚类算法的只有输入变量。和分类不同，聚类不存在一个监督（或控制）机制来执行学习过程。作为替代，聚类算法使用一个或多个启发式（例如多维距离测度）来发现对象的自然分组。

最通用的分类预测两步法包括模型建立/训练和模型测试/部署。模型建立阶段使用包括实际分类标签在内的一组输入数据。在模型经过训练以后，针对测试组样本进行模型的精确度评估测试，最终进行实际应用部署，使用模型对新的数据实例（类别标签未知）进行类别预测。模型评估要考虑的因素如下：

- **预测精度** 模型对新的或者前所未见的数据类别标签进行正确预测的能力。预测精度是最通用的分类模型评估因素。在计算该指标时，将测试数据集的实际类别标签和由模型预测的类别标签进行匹配。然后，根据模型在数据集样本上的正确分类比率就可以计算出正确率，作为精确度的值。本章后面部分还将给出更多的相关材料。
- **速度** 模型生成和应用的计算成本，越快越好。
- **鲁棒性** 在给定数据存在噪声或缺失错误值时，模型进行合理精确预测的能力。
- **延展性** 给定相当大量的数据时，高效构建预测模型的能力。
- **可解读性** 模型的理解洞察水平（例如，模型对某个预测给出怎样的结论）。

4.4.2 分类模型正确性估算

分类模型正确性估算的主要来源是混淆矩阵或者称分类矩阵、列联表。图4-9显示了一个二分分类问题的混淆矩阵。从左上角到右下角的对角线数目表示正确的决策，而对角线以外的数目表示错误。

图4-9 二分分类结果表格的混淆矩阵示例

表4-5给出了常见的分类模型正确性指标公式。

表4-5 常见的分类模型正确性指标

指 标	描 述
正确的正例类别比例 $=\frac{TP}{TP+FN}$	正例总数除正确分类的正例数目的比例（即命中率或攻击率）
正确的负例类别比例 $=\frac{TN}{TN+FP}$	负例总数除正确分类的负例数目的比例（即错误警报率）
正确性 $=\frac{TP+TN}{TP+TN+FP+FN}$	实例总数除正确分类的实例数目（正例和负例）的比例
精度 $=\frac{TP}{TP+FP}$	正确分类的正例数目加上不正确分类的正例数目之和，除正确分类的正例数目的比例
取消率 $=\frac{TP}{TP+FN}$	正确分类的正例数目加上不正确分类的负例数目之和，除正确分类的正例数目的比例

对于非二元分类问题，混淆矩阵更大（矩阵的大小由分类标签的唯一编号确定），正确性指标仅限于分类准确率和总体分类准确率。

$$(\text{实际分类比率})_i=\frac{(\text{正确分类})_i}{\sum_{i=1}^{n}(\text{错误分类})_i}$$

$$(\text{总体分类精度})_i=\frac{\sum_{i=1}^{n}(\text{错误分类})_i}{\text{实例总数}}$$

对监督学习算法推导的分类模型进行正确性估算是非常重要的。原因有如下两点：首先，可以用于估计未来预测的正确性，这意味着对预测系统输出预测结果的信心水平；其次，可以用于从给定集合中选择分类器（从很多经训练的分类模型中识别出最好的）。以下是一些使用最广泛的分类数据挖掘模型估计方法。

简单拆分 简单拆分（或者测试样本估计）将数据分割成2个互斥的子集，分别称为训练集和测试集（或对照集）。通常的做法是选定数据中的2/3作为训练集，剩下的1/3作为测试集。建模者使用训练集，然后在测试集上对所建立的分类器进行测试。当使用人工神经网络作为分类器时，情况有所不同。此时，数据被分割成3个互斥子集：训练集、验证集和测试集。在建模中，验证集被用于防止过度拟合。（有关人工神经网络的更多信息参见第6章）。图4-10说明了简单拆分方法。

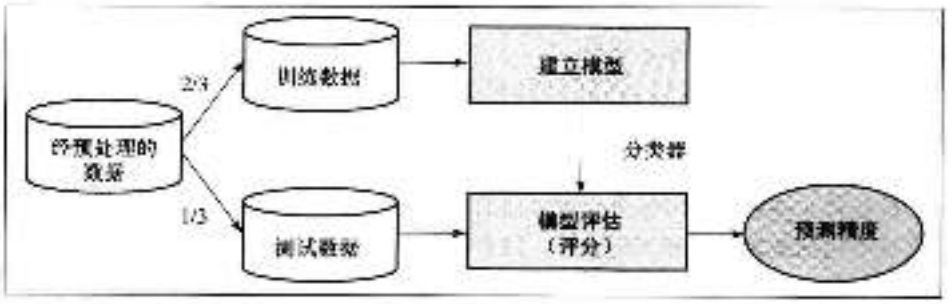


图4-10 简单随机数据拆分

对该方法的主要批评是其假设2个数据子集中的数据属于同一类别（即具有完全相同的属

性）。由于分割是一个简单的随机过程，所以对于数据可能对分类变量有偏斜的大多数实际数据集，这一假设很难满足。为了改善这种状况，提出了分层取样，将样本分层作为输出变量。尽管这对简单拆分有所改进，但仍然存在由于简单随机分割而带来的偏差。

k折交叉确认 在比较两种或多种方法的预测精度时，为了最大限度地减少与训练集和测试集随机取样相关的偏差，可以使用**k折交叉确认方法**。k折交叉确认，也称为轮回估测，将整个数据集随机分成k个大小近似相等的互斥子集。分类模型经过k次训练和测试。每次使用除了一个以外的所有其他折数据进行训练，然后在剩下的一折数据上进行测试。模型的k折交叉确认总体正确性可以由k个正确性指标进行简单平均计算得到，参见以下公式所示：

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

其中，CVA代表交叉检验正确性，k为采用的折数，A为每折的正确性指标（例如，命中率、敏感度、特异性）。

其他分类评估方法 其他常见的评估方法包括如下几种：

- **留一法** 留一法和k值为1的k折交叉确认类似。也就是说，模型数量和数据点数量是相等的，每个数据点在模型上测试一次。这种方法很耗时，但对于较小的数据集来说，有时也是一个可行的选择。
- **拔靴复制法** 从初始数据中提取固定数目的实例作为训练样本，数据集剩下部分用于测试。根据需要，将此过程重复多次。
- **折刀法** 与留一法类似。在折刀法计算精度时，估算过程的每一次迭代都省略一个样本。
- **ROC曲线下面积** ROC曲线下面积是一种图形评估技术，在Y轴上绘制真阳性率，在X轴上绘制假阳性率。ROC曲线以下的面积确定了一个分类器的精度：1表示精度极好；而0.5表明精度等于随机概率。实际的精度值范围在两个极端值之间。例如，在图4-11中，A的分类表现好于B，而C则并不比投硬币的随机概率结果更好。

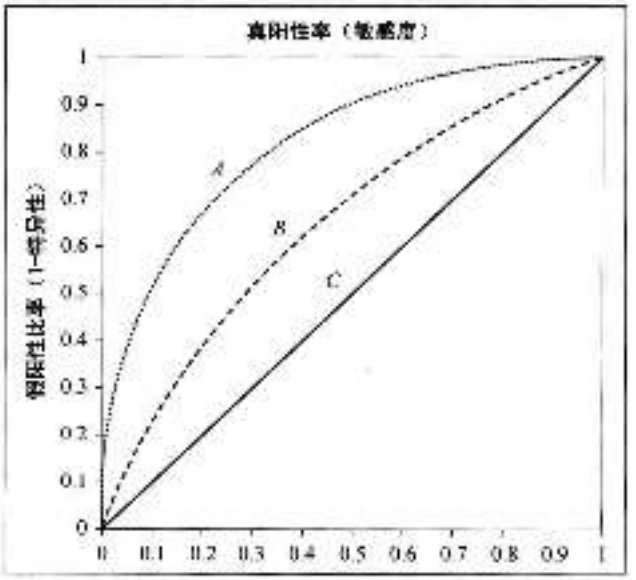


图4-11 一个ROC曲线样本

分类技术 以下是一些用于分类建模的技术或算法：

- **决策树分析** 决策树分析（一种机器学习技术）毫无疑问是数据挖掘领域使用最广泛的分类技术。下一节给出了这种技术的详细描述。
- **统计分析** 在机器学习技术出现以前的很多年，统计技术是最主要的分类算法。统计分类技术包括逻辑回归和判别分析，两者都假设输入和输出变量之间是线性关系、数据趋于正态分布，且变量之间不存在关联，彼此独立。这些假设存在的缺点使得机器学习技术成为趋势。
- **神经网络** 是用于解决分类问题时最常用的机器学习技术。本章后面将对该技术进行详细描述。

- **基于实例的推理** 该方法应用历史实例数据，识别出其共同点，从而将新实例指派到最可能的类别范畴。
- **贝叶斯分类器** 该方法基于事件的过去发生情况，应用概率论建立分类模型，能够将新的实例指派到最可能的类别范畴。
- **遗传算法** 模仿自然演化过程，基于定向搜索机制建立数据样本分类机制。
- **粗糙集** 该方法考虑预定义范畴中的部分类别标签，建立分类问题规则集模型。

对所有分类技术的完整描述超出了本书范围，因此这里仅介绍使用最广泛的几种分类方法。

决策树 在描述决策树的细节以前，我们需要先讨论一些简单术语。首先，决策树包括很多可能影响不同模式分类的输入变量，通常称为属性。例如，建立贷款风险模型要基于两个特征——收入和信用等级，这两个特征就是属性，而输出结果则为分类标签（例如，低风险、中等风险、高风险）。其次，一棵树包括分支和结点。分支表示利用属性得到的一个模式分类测试结果。树末端的叶子结点表示一个模式（从根结点到叶子结点的链条，可以用一个复杂的条件语句来表达）的最终类别选择。

决策树的基本思想在于其对训练集进行递归划分，直至每一部分样本全部或绝大部分都属于同一类别为止。决策树的每一非叶子结点都包含一个分割点，通过测试一个或多个属性来确定如何进一步划分数据。一般而言，决策树算法最初建立树的叶子结点都是单一的，需要进行剪枝以增加其泛化程度，从而提高测试数据的预测正确性。

在决策树的生成阶段，通过递归分割数据直到每一部分都是单一的（即包含成员都属于相同类别）或相对较小的。基本思想就是提问“哪个答案能够提供最多的信息”，类似于我们玩“猜猜二十问”的游戏。

用于分割数据的分割点依赖于分割中采用的属性类型。对于一个连续属性 A，分割的形式为 A 的值。

1. 创建一个根结点，并将所有训练数据分配给根结点。
2. 选择最佳的分割属性。
3. 为分割点的每个值在根结点上增加一个分支。沿着分支的特定分割点和模式路线，将数据分割成为彼此不相容（不相重叠）的子集。
4. 对各叶子结点，重复步骤 2 和 3，直到满足停止条件（例如，结点仅代表一个单独的类别标签）。

已经相继提出了很多不同的决策树创建算法。这些算法的主要区别在于其确定分割属性（和分割值）的方式、分割属性的顺序（对相同属性仅分割一次还是多次）、各结点的分支数目（二分还是三分）、递归停止条件和剪枝方式（先剪枝还是后剪枝）。最广为人知的算法包括来源于机器学习的 ID3（其后又出现了 ID3 的改进版本：C4.5 和 C5）、来源于统计学的分类和回归树（Classification and Regression Tree，CART）和来源于模式识别的卡方自动交互检测（Chi-squared Automatic Interaction Detector，CHAID）。

在建立决策树时，每一结点的目标定位在确定属性及其分割点，以使能够将训练记录进行最佳划分，从而使得结点代表的类型是单一的。已经提出了一些分割指数来评估分割是否恰当，其中最为通用的是基尼系数和信息增益。基尼系数用于 CART 和 SPRINT（可扩展的 PaRalleizable 决策树推导）算法。不同版本的信息增益则被用于 ID3（及其更新版本：C4.5 和 C5）。

基尼系数 在经济学中用于度量人类族群的多样性。同样的概念可以用于确定某一特定类别的纯度，该类别是确定一个特定属性或变量的分支结果。最佳的分割点能够增加分割集合的

纯度。Hastie et al.（2009）对基尼系数及其数学表达进行了更详细的描述。

信息增益 是ID3（可能是最广为人知的决策树算法）采用的分割机制。ID3算法由Ross Quinlan于1986年提出，随后他又将该算法发展成为C4.5和C5算法。ID3及其变种的基本思想是使用熵的概念代替基尼系数。熵度量某一数据集的不确定或随机程度。若一个子集中的所有数据都属于同一种类，那么该数据集中不存在不确定性或随机性，其熵为0。该方法的目标在于建立子树，使得最终得到的各子集熵为0（或接近0）。Quinlan（1986）对信息增益及其数学表达进行了更详细的描述。

应用案例4.6 Highmark公司应用数据挖掘管理保险成本

长期以来，位于宾夕法尼亚匹兹堡的Highmark公司为其成员群体提供廉价的优质卫生保健。Highmark于1996年由两个宾夕法尼亚州蓝十字与蓝盾协会的持牌者：宾夕法尼亚蓝盾（现在的Highmark蓝盾）和宾州西部的蓝十字计划（现在的Highmark蓝十字蓝盾）合并而成。目前Highmark是美国最大的健康保险公司之一。

管理保健组织中的数据

流动于Highmark等管理保健组织中的数据量非常巨大。这些过去常被视为耗费存储空间、很难处理的数据，在最近被认为是新知识的来源。数据挖掘工具和技术为病历数据分析提供了实用手段，也解开了以更低的成本实现更好地管理保健医疗这一难题，这是诸多管理保健医疗公司努力实现的使命。

管理保健公司每天都要收到数以百万计的客户数据字段，每条信息都会更新病案历史记录。公司已经认识到这些数据处理后的用处，并开始使用分析软件工具找出成本相对平均水平更为昂贵的病人群体。早先应用计算机技术提取病人相关信息的工作局限于在两种不同疾病间建立联系。例如，软件工具可以扫描数据，给出确保病或冠心病患者治疗费用最为昂贵的报告。但是，要发现病患的病因，或者分析为什么一些病人相对更容易受到某些疾病的负面影响，这些基于报告的软件工具则无能为力。通过多维信息分析，需要总结不同疾病和病人概况之间的关系和关联，数据挖掘工具能够解决一些这样的问题。

管理保健组织被大量数据淹没，为了避免增加复杂性，其中一些企业不愿意添加数据挖掘应用。他们可能出于各种目的希望扫描数据，但没有能力决定为什么或者怎样分析其数据。不过，健康保险组织正在为提高数据效率和数据组织扫清道路，因此情况对病人和企业都愈加光明。

对数据挖掘的需求

市场压力迫使管理保健组织越来越高效，因而要认真对待数据挖掘。客户要求更多更好的服务，竞争愈加残酷，这些都要求以合适的方式设计和传递定制化产品。

客户化促使我们回到大部分医疗费用发生的原因和位置源点。很多组织开始应用数据挖掘软件来预测哪些人更易于生病，哪些人的治疗费用可能最为昂贵。对未来的关注使得组织能够找出花费最昂贵的病人，并通过采取预防措施降低医疗费用。预测研究的另一重要应用是管理保费。雇员较多的雇主群体成本更高，因而其费率也会增加。

基于历史数据，预测模型能够预告哪些病人更可能成为企业负担。例如，一个预测建模程序可能认为一位糖尿病人有在较高的医疗费用增高风险，这条信息本身可能并不构成有价值的线索。但是，Highmark 的数据挖掘工具在糖尿病人和其他病人参数以及环境相关参数之间建立了联系。即，具有特定心脏不适状况的病人可能具有更高的风险罹患糖尿病。建立这种联系是因为服用强心剂可能导致病人在以后患糖尿病。Highmark 正式证实了该事实。他们声称，他们本来可能不会监测使用强心剂的病人，也不会在强心剂和糖尿病之间建立联系。医学研究成功地对病人复杂状况进行了系统编码。数据挖掘为更好地检测和恰当地干预计划奠定了基础。

来源：Based on G. Gillespie, "Data Mining: Solving Care, Cost Capers," *Health Data Management*, November 2004, findarticles.com/p/articles/mi_km2925/is_200411/ai_n8622737 (accessed May 2009); and "Highmark Enhances Patient Care, Keeps Medical Costs Down with SAS," sas.com/success/highmark.html (accessed April 2006).

4.4.3 数据挖掘聚类分析

聚类分析是一种重要的数据挖掘方法，它将物品、事件或者概念分成称为聚类的公共组。该方法广泛应用于生物、医药、遗传、社交网络分析、人类学、考古学、航天、字符识别，甚至管理信息系统开发中。随着数据挖掘越来越流行，相关技术已被用于商业，特别是营销中。聚类分析已被广泛应用于检测欺诈（包括信用卡和电子商务诈骗）和现代 CRM 系统的客户市场分类。随着人们对聚类分析的认识和应用，更多的商业应用在继续发展。

聚类分析是一种用于解决分类问题的探索数据分析工具。其目标是将实例（例如人、事物、事件）分成组或群，使得相同群中的成员关联程度较强，而不同群中的成员彼此关联程度较弱。每个群描述了其成员所属的类。一个简单的一维聚类分析例子是为大学课堂建立分数范围，以根据不同级别分班。这和美国财政部在 20 世纪 80 年代遇到的建立新的应税级别的聚类分析问题非常类似。在 J. K. Rowling 的《Harry Potter》一书中，有一个虚构的聚类实例。分院帽决定霍格沃茨魔法学校的新生进入哪一个分院（例如，宿舍）。还有一个例子就是确定婚礼客人如何排座位。就数据挖掘而言，聚类分析的重要性在于其可以发现数据中的关联和结构，这些关联和结构虽然本来并不明显却合乎情理，一旦被发现就很有用。

聚类分析结果可用于：

- 识别分类计划（例如客户类项）
- 提出人口种族描述统计模型
- 给出新实例的分类规则，以实现识别、定位或诊断目标
- 提供定义和估算指标、大小，替换原本宽泛的概念
- 发现标记和表达类别的典型例子
- 为其他数据挖掘方法降低问题空间的大小和复杂度
- 识别特定领域（例如偶发事件检测）的离群值。

确定聚类的最佳数目 聚类算法通常需要指定所要寻找的聚类数目。若该数目是先前未知的，就需要以某种方式来确定。可是，并不存在一种最佳方法计算该值。因此，已经提出多种不同的启发式方法。其中应用最广泛的有如下几种：

- 将变量的比率看做是聚类数目的函数。也就是说，选择一个值作为聚类的数目，使得增加聚类不会给数据建模带来多大好处。明确地说，若对由聚类解释的变量比率绘图，那么存在一个点使得边际收益下降（图中将出现一个角），即为所选择的聚类数目。
- 令聚类的数目为 $(n/2)^{1/2}$，其中 n 是数据点的数目。

- 应用赤池信息准则（一种基于熵概念的拟合优度测度）确定聚类的数目。
- 应用贝叶斯信息准则（一种基于最大似然估计的模型选择标准）确定聚类的数目。

分析方法 聚类分析可基于以下的一种或多种通用方法：

- 层级或非层级统计方法（例如k-均值、k-模式，等）
- 自组织映射（Self-Organizing Map，SOM）结构神经网络
- **模糊逻辑** （例如模糊c-均值算法）
- 遗传算法

上述各方法一般使用两种通用的分类方法之一：

- **分裂法** 所有项目起始于同一聚类，然后将其分裂开。
- **聚集法** 所有项目起始于各自的聚类，然后将这些聚类合并到一起。

大多数聚类分析方法使用距离测度计算项目对之间的远近。常用的聚类测度包括欧几里得距离（两点之间的普通距离，可用标尺测量）和曼哈顿距离（也称为两点间的直角距离或计程车距离）。这些聚类测度常基于测量的实际距离，但这不是必须的。典型例子如信息系统开发。在建立这些距离时，可以使用加权平均数。例如，在信息系统开发项目中，可以通过输入、输出、流程和特定数据彼此之间的相似度，将各个系统模块关联起来。然后，将这些因素合计，按项目配对，得到单独的距离测度。

k-均值聚类算法 k-均值聚类算法（其中k代表预定义的聚类数目）是无可争议的引用最多的聚类算法。该算法源于传统的统计分析。顾名思义，该算法将各个数据点（客户、事件、对象等）分配到中心（也称为质心）最接近的聚类中。质心由聚类中所有点的平均值来计算。也就是说，其坐标分别为聚类中所有点各维度的算术平均。以下是算法步骤（如图4-12所示）：

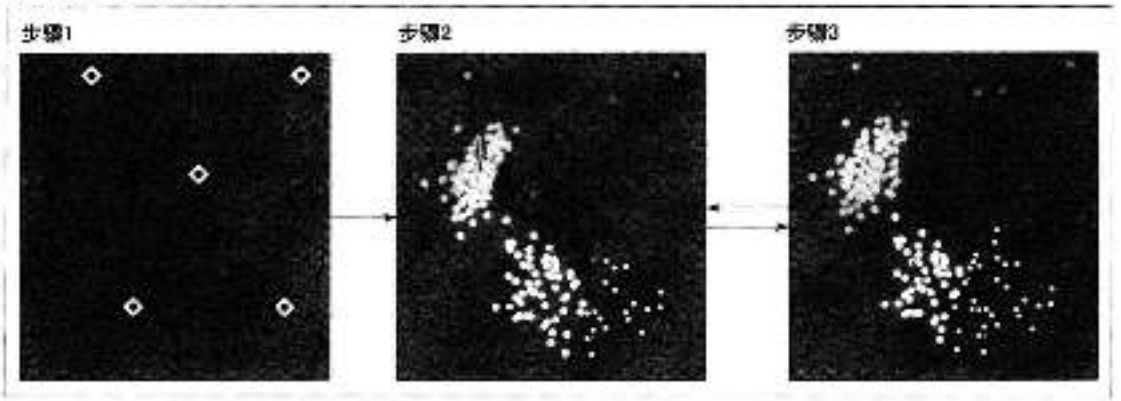


图4-12 k-均值算法步骤的图形说明

初始化步骤：选择聚类数目（即k的值）。

步骤1：随机产生k个任意点作为初始聚类中心。

步骤2：将各个点分配到最接近的聚类中心。

步骤3：重新计算新的聚类中心。

重复步骤：重复步骤2和步骤3，直到满足某综合标准（通常点到聚类的指派变得稳定了）。

4.4.4 关联规则挖掘

关联规则挖掘是一种常用的数据挖掘方法，常被用做例子来向技术背景较弱的读者解释数据挖掘是什么，以及数据挖掘能做什么。大多数读者可能都听说过著名的（或声名狼藉的，取决于如何

看待）在零售店啤酒销售和尿布销售之间发现的关联。据说，某大型连锁超市（可能是沃尔玛，可能不是；究竟是哪家连锁超市并没有一致意见）对客户购买行为做了一次分析，发现啤酒购买和尿布购买在统计上有显著相关性。根据推理，其原因是父亲（假定为年轻男人）会中途在超市停车为宝宝（特别是在礼拜四）购买尿布，而由于没有时间像平常那样再去运动酒吧，他们也会买点啤酒。由于这一发现，据说该连锁超市将尿布放在啤酒旁边，结果是两者的销售量都增加了。

本质上，关联规则挖掘的目标在于发现大型数据库变量（字段）之间的有趣关系（密切关系）。由于其在商业问题中的成功应用，通常将它称为购物篮分析。购物篮分析的主旨在于识别通常在一起购买的（在同一购物篮中一起出现，购物篮可以是杂货铺中的实际购物篮，也可以是电子商务 Web 网站的虚拟购物篮）不同产品（或服务）之间的强关联。例如，购物篮分析可能发现类似模式，"若客户购买笔记本电脑和防病毒软件，那么他也会购买延长 70% 服务时间的延期服务计划"。购物篮分析的输入是简单的销售点交易数据，将很多一起购买的产品或服务（就像一张购物小票的内容）列入一个单独的交易实例表格。分析结果提供的宝贵信息能够用于更好地理解客户购买行为，从而最大限度地提高商业交易利润。商店可以充分利用这些知识：(1) 把这些商品放在彼此旁边，方便客户一起拿取，避免购买其他东西时忘了买某件商品（增加销售量）；(2) 将这些商品打包促销（如果其中的其他商品都在打折，那么不单独销售其中某件商品）；(3) 将这些商品分开放置，客户必须走过通道寻找所要购买的东西，这时客户也可能会看到并购买其他商品。

购物篮分析的应用包括交叉销售、店铺设计、目录设计、电子商务网站设计、在线广告优化、产品定价和促销配置。本质上，购物篮分析帮助商店从客户购买模式中推断出其需求和偏好。在商业世界之外，管理规则也成功地应用于发现症状和疾病、诊断和病人特征及疗法（用于医学决策支持系统）、基因和其功能（用于基因组项目）等之间的关系。

关于关联规则挖掘能够发现的模式/关系，应该问的一个很好的问题是，"所有的关联规则都是有趣且有用的吗？"关联规则挖掘使用两个通用指标来回答这个问题：支持度和置信度。在定义这些术语前，让我们先稍微专业一些，看看一条关联规则看起来是什么样的：

$$X \Rightarrow Y[S\%, C\%]$$

$$[\text{笔记本电脑, 防病毒软件}] \Rightarrow [\text{延期服务计划}][30\%, 70\%]$$

式中，X——产品或服务；称为左手边（Left-Hand Side，LHS），或前因；Y——产品或服务，称为右手边（Right-Hand Side，RHS），或后果；这里，X 和 Y 关联。S 为特定规则的支持度，C 为置信度。规则的支持度用于度量相关产品或服务（即左手边 + 右手边 = 笔记本电脑、防病毒软件和延期服务计划）在同一交易中共同出现的频率，也就是数据集中包含特定规则中提到的所有产品和服务的交易比例。在本例的假想商店数据库中，大约有 30% 的交易在单张销售小票中包含所有 3 种产品。规则的置信度度量右手边（后果）和左手边（前因）中的产品和服务共同出现的频度，也就是包含 LHS 同时也包含 RHS 的交易比例。换句话说，就是在规则左手边 LHS 已经存在的情况下，交易中发现规则右手边 RHS 的条件概率。

有多种算法可用于生成关联规则。最著名的算法包括 Apriori、Eclat 和 FP-Growth。这些算法只做了一半工作，即识别出数据库中的频繁项集。频繁项集指的是在一次交易（例如一个购物篮）中频繁共同出现的任意数目的项目。一旦识别出频繁项集，就需要将其转换成包括前因和后果两部分的规则。从频繁项集中确定规则是一个简单的匹配过程，但该过程对于大型交易数据库可能非常耗时。尽管规则的每部分都可以包含很多项目，但实际中的后果部分通常仅包含一个单独项目。下一节将解释一种最流行的频繁项集识别算法。

Apriori 算法 Apriori 算法是最通用的一种关联规则发现算法。给定一组项集（例如零售交易集合，购买的每样商品），算法试图发现至少有最低数量公共项集的子集（即符合最低限度的

支持度）。Apriori 使用自底向上的方法，频繁项集每次扩展一项（一种候选集生成方法，其中频繁项集的大小从一项子集增长到二项子集，然后是三项子集，如此继续），每一级的候选组对最小支持度数据进行测试。当不能进一步成功找到扩展时，算法终止。

考虑以下实例。某杂货店通过 SKU（库存单元）跟踪销售交易，了解适宜那些商品会在一起购买。图 4-13 显示了交易数据库以及识别频繁项集的后续步骤。交易数据库中的每一库存单元对应一个产品，例如“1 = 黄油”，“2 = 面包”，“3 = 水”等。Apriori 算法的第一步是计算各项目（一项项集）频率（即支持度）的总和。在这个简化的例子中，设最小支持度为 3（或 50%，表示若某项集在数据库的每 6 条交易中至少出现 3 次，则可作为频繁项集）。由于所有一项项集的支持度都至少等于 3，所以将其都作为频繁项集。不过，若任一一项项集为非频繁项集，其也不会成为二项项集成员。Apriori 通过这种方式对所有可能的项集树进行了剪枝。如图 4-13 所示，所有可能的二项项集使用一项项集产生，并通过交易数据库计算其支持度。因为二项项集 {1，3} 的支持度小于 3，所以其将不会被包括在产生下一级项集（三项项集）的频繁项集中。该算法看起来很简单，其实仅仅是对小数据集而言才这样。对于较大的数据集，特别是当其包含大量很少出现的项目，或者包含少量多次出现的项目时，搜索和计算将变成一个计算密集过程。

图 4-13 Apriori 算法中的频繁项集识别

4.4 节复习题

1. 指出至少 3 种主要的数据挖掘方法。
2. 请举例说明何种情形适合使用分类数据挖掘技术，何种情形适合使用回归数据挖掘技术。
3. 列出并简单定义至少两种分类技术。
4. 比较和筛选最佳分类技术的标准有哪些？
5. 简单描述使用决策树算法。
6. 给出基尼系数的定义，并说明基尼系数如何进行度量？
7. 举例说明何种情形适合使用聚类分析数据挖掘技术。
8. 说明聚类分析和分类的主要区别。
9. 有哪些聚类分析方法？
10. 举例说明何种情形适合使用关联数据挖掘技术。

4.5 数据挖掘中的人工神经网络

在其他技术产生的解决方案不能令人满意时，神经网络已经成为一种先进的数据挖掘工具。顾名思义，在信息处理时，神经网络具有生物启发建模能力（表示类似于人脑）。由于具有从数据中“学习”的能力、非参数特性（即没有严格假设）和概括能力（Kaykin，2009），所以在很多预测和商业分类应用中，神经网络已经被证明是一种很有前途的计算系统。神经计算是指机

器学习中的一种模式识别方法。神经计算产生的模型结果常被称为**人工神经网络**（Artificial Neural Network，ANN）或**神经网络**。神经网络是数据挖掘工具包的关键组件。神经网络大量应用于金融、营销、制造、运营管理、信息系统和社会行为分析等领域中。

生物神经网络由很多大型互联神经元组成。每一个神经元都有轴突和树突，这些指状突起通过收发生物化学信号使得神经元能够和其相邻的其他神经元通信。和生物神经网络多少有些类似的是，人工神经网络主要为人工神经元的简单互联处理单元（Processing Element，PE）组成。和生物神经元类似，人工神经网络中的处理单元共同并行处理信息。人工神经网络拥有一些和生物神经网络类似的理想特性，例如学习能力、自组织能力和支持容错的能力。图4-14显示了生物神经网络和人工神经网络之间的类似之处。

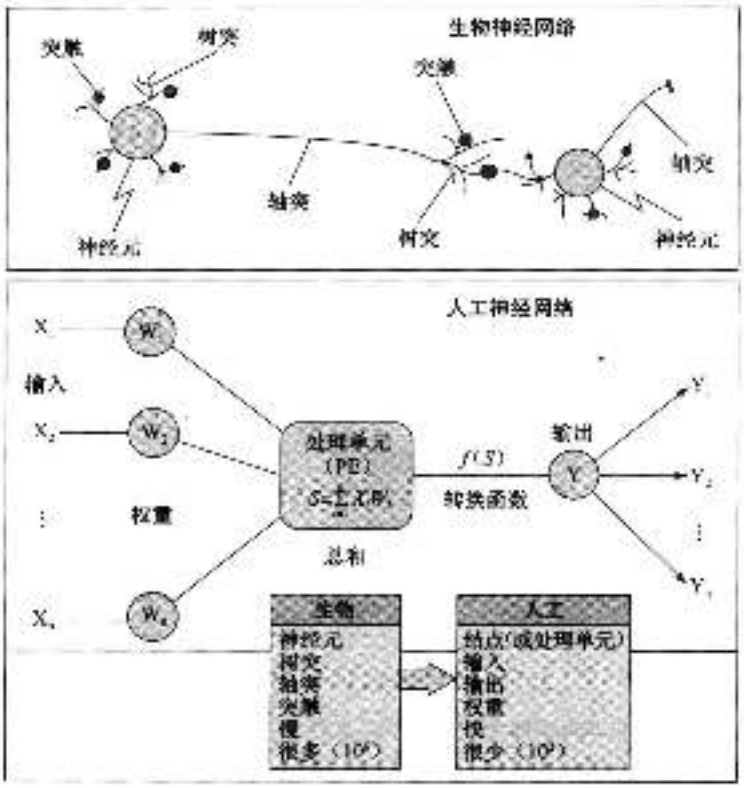


图4-14 生物神经网络和人工神经网络类比

4.5.1 人工神经网络的要素

处理单元 人工神经网络的处理单元本质上是人工神经元。和生物神经元类似，每一处理单元接收输入，进行处理后，传递一个输出，如图4-14下半部分所示。输入可以是原始输入数据，也可以是其他处理单元的输出。输出可以是最终结果，也可以作为其他神经元的输入。

信息处理 神经元接收到的输入要经过两步处理，得到输出：求和函数和转换函数（见图4-14的下半部分）。求和函数产生输入及其连接权值的和。转换函数取求和函数的生成值，进行一个非线性函数（常为S型函数）运算，然后生成该神经元的输出值。

网络结构 每个人工神经网络由一些分层神经元（或PE）组成。图4-15显示了一个典型的神经网络分层结构。如图所示，包括3层：输入层、中间层（隐含层）、输出层。隐含层中的神

经元取前一层的输入，将其转换为输出，再进一步处理。在输入层和输出层之间可以有多个隐含层，但通常只使用一个隐含层。这种分层神经网络结构通常称为**多层感知器**（Multi-Layered Perceptron，MLP）。MLP 能够得到高度精确的分类和回归预测模型。除了 MLP，人工神经网络还有其他结构，例如 Kohonen 自组织特征映射（常用于聚类问题）、Hopfield 网络（用于解决复杂计算问题）、循环网络（和正反馈相反，该结构也允许后向连接）和概率型网络（权值可基于由训练数据得到的统计尺度进行调整）。

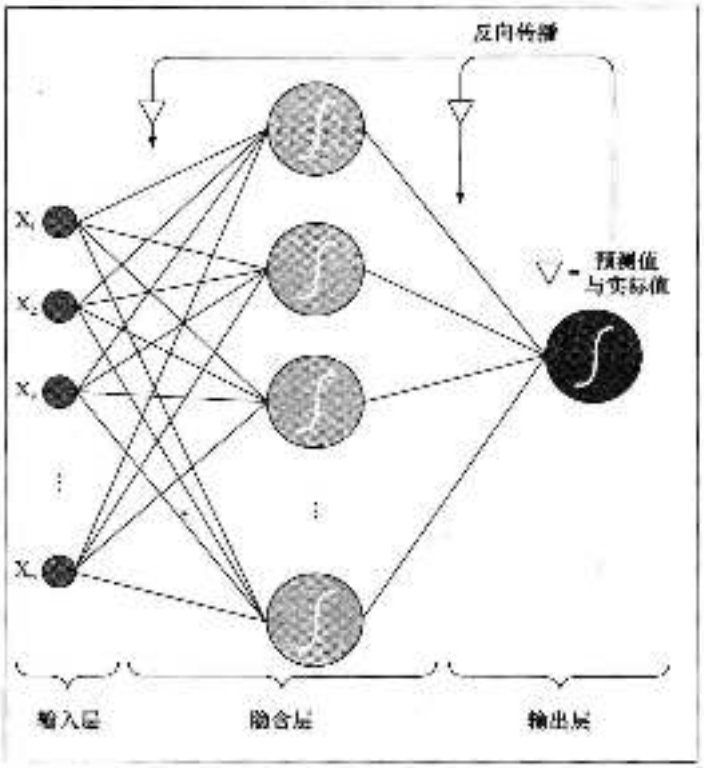


图 4-15 支持反向传播学习算法的多层感知神经网络

反向传播 反向传播是一种前馈 MLP 网络学习机制。这种学习机制遵循一个迭代过程，网络输出和理想输出之间的差异被反馈到网络，用以调整网络权值，从而得到更接近实际值的输出结果。

4.5.2 人工神经网络应用

由于其能够对高度复杂的现实世界问题建模，所以学术界和业界已经发现了人工神经网络的很多用途。很多应用已经为过去认为不可解的问题带来了解决方案。在最高概念层次，神经网络应用一般可以分为 4 类（对应于数据挖掘的一般处理任务）：

1. **分类** 神经网络可以经训练预测一个类别（即类别标签）输出变量。在数学意义上，这涉及将一个 n 维空间分成不同区域，且给定空间中的一点，应该能够确定其属于哪个区域。很多现实世界的模式识别应用都采用了这种思想，每个模式被转换为一个多维点，并分类成特定组，分别表达一个已知模式。分类任务使用的神经网络类型包括前馈网络（例如支持反向传播学习的多层感知神经网络）、径向基函数和概率神经网络。应用案例 4.7 介绍了一个有趣的利用神经网络预测特征分析和改善啤酒风味的案例。

应用案例4.7 库尔斯公司利用神经网络改善啤酒风味

位于英格兰啤酒之都特伦特河畔伯顿的库尔斯酿酒公司，以拥有英国最好的啤酒品牌、20%的市场占有率、多年的经验和行业最优秀的人才为荣。热门品牌包括Carling（国内最畅销的啤酒）、Grolsch、库尔斯精细淡啤酒、Sol和Knrenwolf。

问题

关于喝什么饮料，今天的客户有很多种选择。客户的选择依赖于各种因素，包括情绪、观点和场合。库尔斯公司的目标是无论何种情形，确保客户选择库尔斯品牌产品。

按照库尔斯公司的说法，创意是长期成功的关键。要成为客户选择的品牌，库尔斯需要创造性地预测客户变化无常的情绪。对啤酒来说，很重要的一个问题在于风味；各种啤酒都有自己独特的风味。这些风味多数由感官组评定来确定，只是这种评定要耗费时间。如果库尔斯能够仅仅通过化学成分来了解啤酒风味，那么就可以为制作迎合客户预期的啤酒开辟新途径。

目前还不是很清楚化学分析和啤酒风味之间的关系。有大量啤酒化学成分和感官分析的数据，库尔斯需要一种机制将两者联系起来。神经网络被用于建立化学成分和感官分析之间的联系。

解决方案

多年以来，库尔斯公司积累了数量可观的最终产品分析数据，并就之以由经训练的内部测试小组提供的感官数据。下表列出了一些分析数据输入和感官数据输出：

分析数据：输入	感官数据：输出	分析数据：输入	感官数据：输出
酒精	酒味	乙酸异丁酯	烧焦
颜色	酸味	丁酸乙酯	啤酒花
苦味波计	麦芽味	乙酸异戊酯	乳脂味
乙酸乙酯	颗粒	己酸乙酯	甜

首先使用的是限于单一品质和风味的单神经网络，对分析数据和感官数据间关系建模。该神经网络基于NeuroDimension公司（nd.com）提供的解决方案包，由一个包括两个隐含层的多层感知器架构组成。数据在网络内被标准化，从而能够对各感官输出进行比较。神经网络通过相关输入/输出的组合来达成进行训练，学习输入和输出之间的关系。当网络错误100步以内没有明显改善时，训练自动终止。训练被执行50次，以保证能够计算得到相当数量的普通网络错误用于比较。在每次训练运行以前，先通过随机化数据源记录提出一个不同的训练交叉检验数据集，以消除任何偏差。

这种技术产生的结果很差，主要源于两个因素。首先，集中于单一产品品质滋味的数据波动非常少，神经网络不能从数据中提取出有用关系。其次，所给出的输入很可能只有一个子集对所选啤酒风味有作用。神经网络性能受到了由对风味没有作用的输入产生的“噪声”影响。

为了解决第一个问题因素，在训练范围中包括了更多样化的产品范围。识别最重要的分析输入更具有挑战性。该挑战是通过使用一个能够对输入的所有可能组合进行神经网络训练的软件开关解决的。软件开关并不禁止重要输入；如果重要输入被禁止，则可以预期网络错误将增加。若被禁止的输入无关紧要，则网络错误可能维持不变，也可能因为噪声被除去而减少。这种方法被称为穷举搜索，因为其评估所有可能的组合。虽然这种技术概念简单，但在计算上却由于输入数量巨大而不切实际；所以每种风味的可能组合数目达到1 670万。

需要一种更高效的方法搜索相关输入。遗传算法可以解决该问题。遗传算法能够利用不同的输入开关对神经网络的误差项做出响应。遗传算法的目标在于最大限度地减少网络误差项。当网络误差项最小时，开关设置可以识别出最有可能预测风味的分析输入。

结果

在确定有哪些相关输入后，就有可能识别出哪些风味可以进行更精确的预测。使用原先识别出的相关输入对网络进行多次训练。在每次训练开始以前，网络数据被随机化，保证使用的训练数据集和交叉检验数据集不同。每次训练进行完毕后，网络错误被记录下来。用于评估训练网络性能的测试集包括样本数据中的大约80条记录。神经网络使用化学输入精确预测出了一些风味。例如，"烧焦"风味的预测相关系数为0.87。

目前，分析数据正被用于预测有限数量的风味。由于潜在的相互作用和高度可变的灵敏度阈值，感官数据极度复杂。标准仪器分析往往倾向于总体参数，而且由于实际经济原因，对很多风味活跃化合物样本不测量。只有将大量的风味作用分析物都考虑进来，才能对风味和分析数据之间的关系有效建模。此外，除了明显的风味活跃物质以外，总体感官概况中还应当考虑口感和身体因素。

来源：Based on C. J. Wilson and L. Threapleton, "Application of Artificial Intelligences for Predicting Beer Flavours from Chemical Analysis," *Proceedings of the 29th European Brewery Congress*, Dublin, Ireland, May 17 – 22, 2003, neurosolutions.com/resources/apps/beer.html (accessed January 2010); R. Nischwitz, M. Goldsmith, M. Lees, P. Rogers, and I. MacLeod, "Developing Functional Malt Specifications for Improved Brewing Performance," The Regional Institute Ltd., regional.org.au/au/abts/1999/nischwitz.htm (accessed December 2009).

2. **回归** 可以训练神经网络用于预测数值型（即实数或整数）输出变量。若一个网络在对已知值序列建模时适应得很好，则其也可用于对未来结果进行预测。一个明显的回归任务例子是预测股票市场指数。用于回归任务的人工神经网络类型包括前馈型网络（例如支持反向传播学习的多层感知器）和径向基函数。

3. **聚类** 在有些情况下，数据集非常复杂，不存在明显的数据分类方法。人工神经网络可以用于识别这些数据的特征，并且在缺少数据先验知识的情况下将其分为不同的类别。这种技术对于识别商业和科学问题中的事物自然分组很有用。用于解决聚类问题的人工神经网络类型包括自适应共振理论网络和自组织映射神经网络。

4. **关联** 神经网络可以被训练"记住"很多独特的模式。这样，当特定模式被歪曲时，网络可以将其和记忆中最接近的模式关联，恢复该模式原来的形式。当数据包含噪声或不完整时，这对于恢复噪声数据和识别隐藏对象或事件非常有用。用于解决关联问题的人工神经网络类型有Hopfield网络。

4.5 节复习题

1. 什么是神经网络？
2. 生物神经网络和人工神经网络之间有哪些共同点？有哪些区别？
3. 什么是神经网络结构？有哪些常用的神经网络结构？
4. 多层感知器神经网络是如何学习的？

4.6 数据挖掘软件工具

很多软件商都提供了强大的数据挖掘工具。数据挖掘软件供应商包括SPSS（PASW Modeler）、SAS（Enterprise Miner）、StatSoft（Statistica Data Miner）、Salford（CART、MARS、TreeNet、

RandomForest）、Angoss（KnowledgeSTUDIO，KnowledgeSeeker）和 Megaputer（PolyAnalyst）。可以看出，大多数热门工具都是由大型统计软件公司（SPSS、SAS 和 StatSoft）开发的。大多数商务智能工具供应商（例如，IBM Cognos、Oracle Hyperion、SAP Business Objects、MicroStrategy、Teradata 和 Microsoft）在其软件产品中也会集成某种程度的数据挖掘功能。这些 BI 工具主要关注的仍然是多维建模和数据可视化，因此并不被看成是数据挖掘工具软件商的直接竞争者。

除了这些商业工具以外，还可以使用一些开源或免费的数据挖掘软件工具。Weka 怀卡托智能分析环境（Waikato Environment for Knowledge Analysis，Weka）可能是最流行的免费开源数据挖掘工具，由新西兰 Waikato 大学的研究者们开发（该工具可以从 cs. waikato. ac. nz/ml/weka/下载）。Weka 包括很多支持不同数据挖掘工作的算法，且用户界面很直观。RapidMiner（由 Rapid-I 开发，可从 rapid-i. com 下载）也是一个新近发布的免费（非商用）数据挖掘工具。该工具具有图形增强用户界面、支持相当多的算法、且集成了多种数据可视化特征，这些都使其在各种免费工具中别具一格。商业工具（如 Enterprise Miner、PASW 和 Statistica）和免费工具（如 Weka 和 RapidMiner）之间的主要区别在于其计算效率。对涉及特大数据集的数据挖掘任务，免费软件花费的时间可能要远远超过商业软件。在有些情况下，使用免费软件甚至是不可行的（即由于计算机存储空间的低效而造成崩溃）。表 4-6 列出了一些主要产品及其 Web 网址。

在数据挖掘研究中，Microsoft 的 SQL Server 已经成为越来越热门的商务智能功能套件。其中，数据和模型都存储在同一关系型数据库环境中，使得模型管理相当容易。Microsoft 企业联盟在全世界范围内为教学和科研提供 Microsoft SQL Server 2008 软件包访问服务。该联盟是为了让全世界的大学能够访问企业技术而不必在校园内进行必需的软硬件维护建立的。联盟提供各种各样的商务智能开发工具（例如，数据挖掘、建立立方体和商业报告）和大量源自山姆会员商店、狄乐百货和泰森食品的大规模实际数据集。图 4-16 显示了 SQL Server 2008 商务智能开发套件中客户流失分析决策树的开发。Microsoft 企业联盟是免费的，只能用于学术目的。企业系统主机位于阿肯色大学的山姆沃尔顿商学院，允许联盟成员和其学生使用简单的远程桌面连接访问这些资源。通过网址 Enterprise. waltoncollege. uark. edu/mec/可以得到如何成为联盟成员的细节信息、易于学习的指导手册及相关示例。

表 4-6 部分数据挖掘软件产品

产品名称	网址（URL）
Clementine	spss. com/Clementine
Enterprise Miner	sas. com/technologies/bi/analytics/index. html
Statistica	statsoft. com/products/dataminer. htm
Intelligent Miner	ibm. com/software/data/iminer
PolyAnalyst	megaputer. com/polyanalyst. php
CART，MARS，TreeNet，RandomForest	salford-systems. com
Insightful Miner	insightful. com
XLMiner	xlminer. net
KXEN（Knowledge eXtraction Engines）	kxen. com
GhostMiner	fqs. pl/ghostminer
Microsoft SQL Server Data Mining	microsoft. com/sqlserver/2008/data mining. aspx
Knowledge Miner	knowledgeminer. net
Teradata Warehouse Miner	ncr. com/products/software/teradata_mining. htm

（续）

产品名称	网址（URL）
Oracle Data Mining（ODM）	otn.oracle.com/products/bi/9idmining.html
Fair Isaac Business Science	fairisaac.com/edm
DeltaMaster	bissantz.de
iData Analyzer	infoarmen.com
Orange Data Mining Tool	ailab.si/orange/
Zementis Predictive Analytics	zementis.com

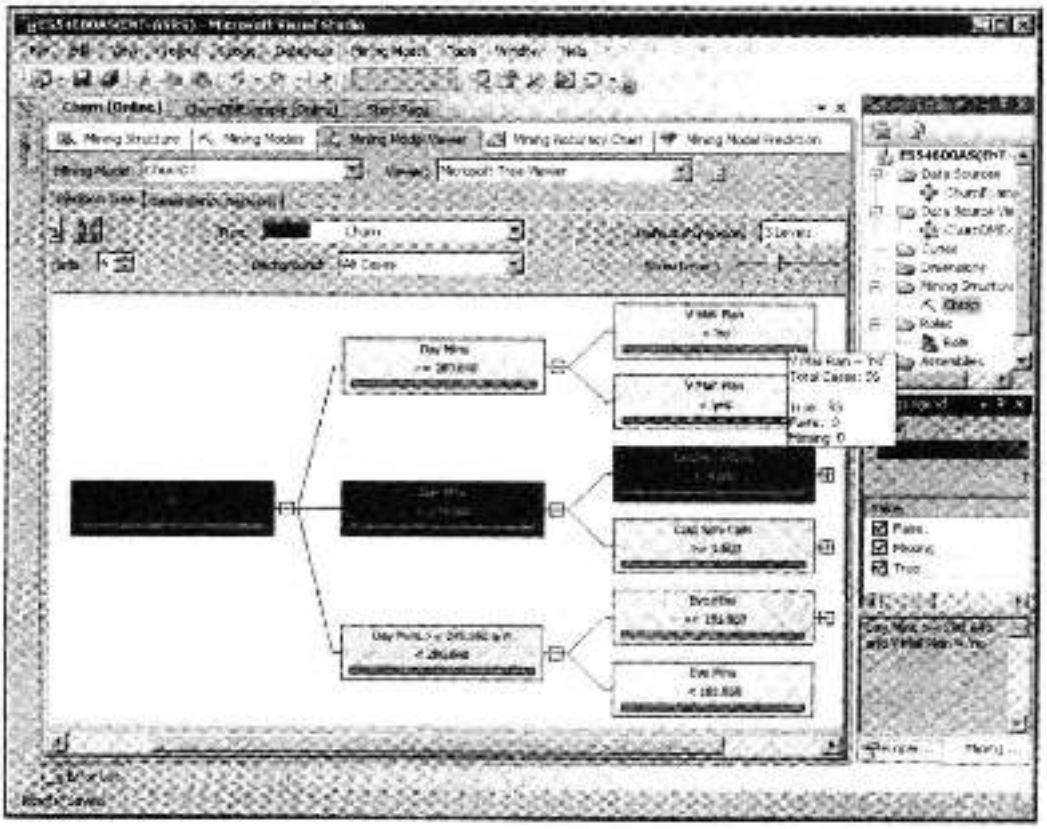


图4-16 SQL Server 2008中的决策树开发环境截图

来源：Microsoft公司提供的Microsoft SQL Server 2008；经Microsoft允许使用。

2009年5月，kdnuggets.com就如下问题对数据挖掘研究群体进行了一次调查："在过去的6个月中，你使用哪种数据挖掘工具做实际项目（不仅仅是为了评估）？"为使调查结果更具代表性，来自软件供应商的投票被剔除了。往年的经验表明，在SPSS Clementine、SPSS Statistics以及SAS Enterprise Miner、SAS Statistics之间具有强相关性；因此，这两组工具系列的投票被合并到一起。总共有364张不同的投票结果被计数排名。最热门的工具是SPSS PASW Modeler、RapidMiner、SAS Enterprise Miner和Microsoft Excel。和往年的调查结果（见kdnuggets.com/polls/2008/data-mining-software-tools-used.htm的2008年数据）相比，在商业工具中，SPSS PASW Modeler、StatSoft Statistica和SAS Enterprise Miner表现出最强劲的增长势头；在免费软件中，RapidMiner和Orange增长最快。调查结果，如图4-17所示。

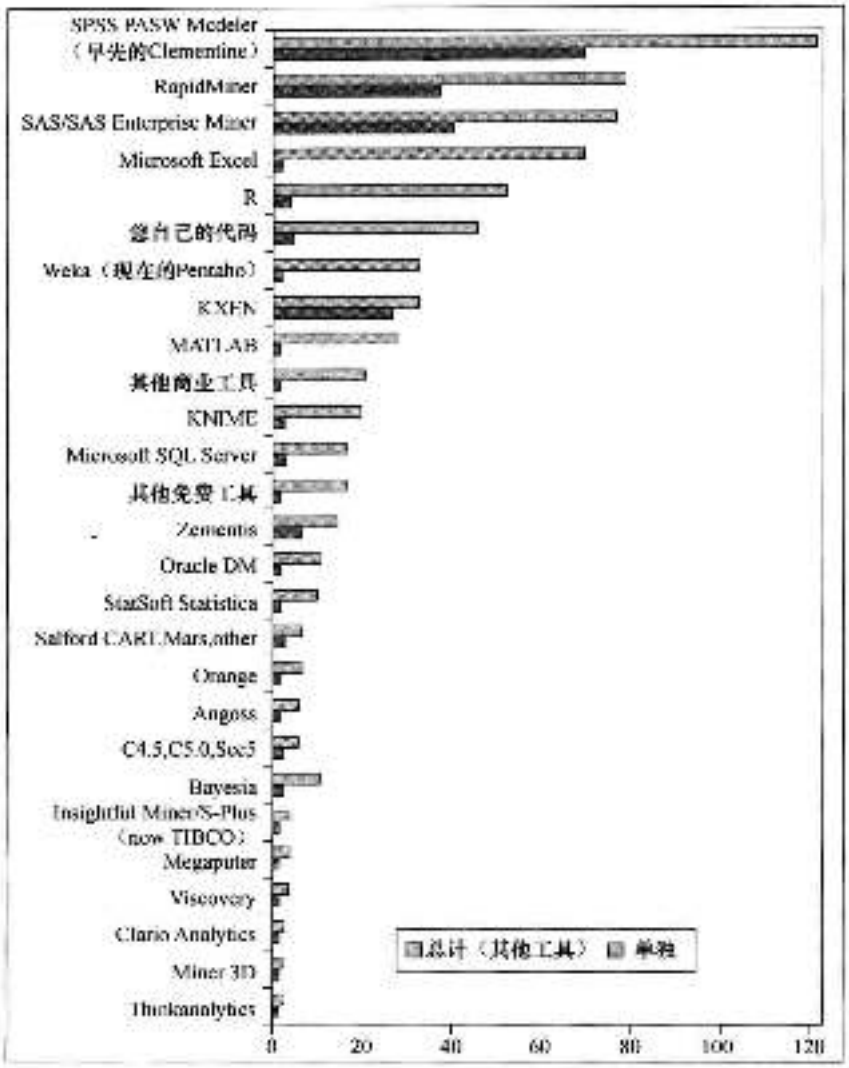


图4-17 常用商业数据挖掘软件工具

来源：kdnuggets.com，经允许使用。

应用案例4.8 客户流失预测——不同工具的竞争

2003年，杜克大学/NCR Teradata咨询中心想要寻找最好的预测建模技术来帮助无线通信供应商解决一个困惑问题：客户流失。尽管其他行业同样面临客户转向竞争对手的问题，但无线服务零售客户更换服务供应商的速度大约是每年25%或每月25例。在20世纪90年代早期，新用户增长比率达50%，通信公司关注重点在于获取新客户而不是保持已有客户。但是，在增长速度放缓（仅为10%）的新时期，很显然客户保持对于总体利润率至关重要。

客户保持的关键在于预测哪些客户投奔竞争对手的风险最高，并为其提供价值激励以留住客户。要有效执行该策略，必须进行准确预测（流失记分卡）以使客户保持工作所针对的客户对象是合适的。

数据

数据由一家大型通信公司提供，使用的是其2001年下半年的客户记录。所提供的账户汇总数据属于已在该公司至少6个月的100 000名客户。为了帮助建模，对流失客户（在随后60天中离开该公司的客户）进行了采样，使得样本中有一半是流失客户，另外一半是随后60天后仍然留在该公司的客户。采用171个各种各样的潜在预测变量，跨越标准服务供应商拥有的全部常规数据类型。预测数据包括：

- **人口统计资料**：年龄、位置、号码、子女年龄等。
- **财务状况**：信用评分和信用卡所有权。
- **产品规格**：手机费用、手机功能等。
- **电话使用情况**：各种通话的号码和时长。

评价标准

数据用于支持预测建模。参与者（数据挖掘软件公司、大学研究中心和其他非盈利咨询公司）被要求使用其最好的模型来预测两组不同客户的流失概率：来自2001年下半年的51 306个"当前"样本和来自2002年第一季度的100 462个"未来"样本。一般认为预测"未来"数据比较难，因为外部因素和行为模式都会随着时间的推移而改变。在真实世界中，预测模型总是被用于未来的数据，比赛组织者想要复制一个类似的场景。

比赛中的每一个竞争者都被要求将当前和未来的评分样本按其流失概率降序排列。比赛组织者利用所掌握的实际流失状态，对各预测模型计算两大性能指标：总体基尼系数和前10位提升。对当前和未来两种样本都计算这两个指标值，这样每位参赛者都有4个性能评分值。包括比赛Web站点在内的多处网址都详细说明了该评价标准。前十分提升很容易解释：测量一个模型所捕获的最有可能流失的客户中实际流失的客户数目。

结果

参赛者可以在一定时间范围内，根据评价标准建立并优化模型。在所有类别中的优胜者都是Salford Systems。Salford Systems使用它的TreeNet软件建立模型。TreeNet是以建立精确分类模型而著称的增强决策树分析的一种创新形式。裁判发现，在预测流失时，决策树和逻辑回归方法一般是所有参赛作品中最好的，尽管他们也承认比赛中并非所有方法都具有合适的代表性。

Salford的TreeNet模型普遍捕获了最多的流失客户，并且发现了在171个预测变量中，哪些对于预测客户流失是最重要的。在前10%客户中，TreeNet发现的流失客户比竞争对手的平均水平多35%到45%，比随机样本中可以发现的数目多3倍。对于用户群很大的公司，这个结果可以转换成每月多识别出上千的潜在流失客户。对这些客户采取合适的客户保持策略，每年可以为公司节省上百万美元。

来源：Salford Systems, "The Duke/NCR Teradata Churn Modeling Tournament," salford-systems.com/churn.php (accessed April 20, 2009); and W. Yu, D. N. Jutla, and S. C. Sivakumar, "A Churn-Strategy Alignment Model for Managers in Mobile Telecom," *Proceedings of the Communication Networks and Services Research Conference*, IEEE Publications, 2005, pp. 48–53.

4.6 节复习题

1. 最流行的商业数据挖掘工具有哪些？
2. 为什么最流行的工具都是由统计公司开发的？
3. 最流行的免费数据挖掘工具有哪些？
4. 商业数据挖掘软件工具和免费数据挖掘软件工具有哪些主要区别？
5. 您在选择数据挖掘工具时，考虑的前5个准则是什么？请解释。

4.7 关于数据挖掘的一些谎言和谬误

数据挖掘是一种强大的分析工具，它使得企业经理能够通过描述过去属性预测未来。数据挖掘帮助市场营销人员解开客户行为模式之谜。数据挖掘的结果可以用于增加收益、减少开支、识别欺诈和发现商机，从而为获得竞争优势开辟了新领域。作为一个正在发展和成熟的领域，数据挖掘常常伴随着诸多谬误说法，包括如下一些（Zmirn，2003）：

谬误	实际
数据挖掘提供快速的类似占卜预言的预测	数据挖掘是一个多步骤过程，需要精心主动的设计和应用
数据挖掘对商业应用还不可行	目前数据挖掘的发展水平几乎可以用于任何业务
数据挖掘需要一个独立的专用数据库	由于数据库技术的发展，不需要专用数据库，虽然有时有一个专用数据库会更好
只有高学历的人才能进行数据挖掘	新的基于Web的工具使得任何教育水平的管理者都可以进行数据挖掘
数据挖掘只适用于客户数据很多的大公司	只要数据能够正确反映业务或客户，企业就可以应用数据挖掘

具有远见卓识的数据挖掘人员能够理解这些谬误不过是奇谈怪论，因此他们获得了巨大的竞争优势。

以下是实际中常犯的10大错误（Skalak，2001；Shultz，2004），应当尽力避免：

1. 未能正确定义数据挖掘要解决的问题。
2. 忽视赞助商对于数据挖掘以及数据挖掘能做什么，不能做什么的看法。
3. 数据准备时间不够充分。数据准备工作需要的时间比一般理解得更多。
4. 只关注汇总结果，忽视单个记录。IBM的DB2 IMS能够突出显示感兴趣的单个记录。
5. 对数据挖掘过程和结果的跟踪过于草率。
6. 忽视发现结果中的疑点而继续进行挖掘过程。
7. 盲目反复运行数据挖掘算法。更重要的是，要努力思考数据分析下一步要做什么。数据挖掘是一种很需要亲自实践的活动。
8. 轻信被告知的所有一切数据信息。
9. 轻信被告知的关于自己进行的数据挖掘分析的所有一切。
10. 采用和赞助商不同的方法来测量结果。

4.7节复习题

1. 关于数据挖掘有哪些最常见的谬误看法。
2. 产生这些数据挖掘错误看法的原因是什么？
3. 有那些数据挖掘错误是最常见的？如何消除或最大限度地避免这些错误？

本章重点

- 数据挖掘是一个从数据库中发现新知识的过程。
- 数据挖掘采用的数据源可以是简单的平面文件，也可以是数据仓库。
- 数据挖掘有很多其他名称和定义。
- 数据挖掘是很多学科的交叉，包括统计学、人工智能和数学建模。
- 企业使用数据挖掘加强对客户的理解，优化企业运作。
- 在企业和政府几乎所有领域都可以见到数据挖掘应用，包括健康保健、金融、营销和国土安全。
- 数据挖掘任务可以分为3大类：预测（分类或回归）、聚类和关联。
- 和创建其他信息系统类似，要取得成功，数据挖掘项目必须遵循系统的项目管理流程。

- 已经确立的数据挖掘过程包括：CRISP-DM、SEMMA 和 KDD 等。
- CRISP-DM 为实施数据挖掘项目提供了一种系统的顺序方法。
- 数据挖掘项目前面的几步（即理解领域和相关数据）耗费整个项目时间的绝大部分（常占总时间的 80%）。
- 数据预处理对于成功进行任何数据挖掘研究都是至关重要的。好的数据会产生好的信息；好的信息会产生好的决策。
- 数据预处理包括 4 个主要步骤：数据整合、数据清洗、数据转换和数据简化。
- 分类方法通过包括输入和分类标签结果的过去例子学习，一旦正确训练，就能够对未来的实例进行分类。
- 聚类将记录划分为自然分组，各分组内部成员特征类似。
- 数据挖掘可以是假设驱动或发现驱动的。假设驱动的数据挖掘开始于用户的提议。发现驱动的数据挖掘结果则更广泛。
- 有很多不同算法被广泛应用于分类。商业实现算法包括 ID3、C4.5、C5、CART 和 SPRINT。
- 决策树通过按不同属性划分数据，每一叶子结点具有同一类别的所有模式。
- 基尼系数和信息增益（熵）是确定决策树分支选择的两种常用方法。
- 基尼系数度量样本的纯度，若其样本中所有样例都属于同一类别，那么其基尼系数值为 0。
- 用于度量分类模型预测精度的评估技术有多种，包括简单拆分、k-折交叉确认、拔靴复制法和 ROC 曲线下面积。
- 当数据记录不存在预定义的类别标识符（即未知某一特定记录属于哪一个类别）时，使用聚类算法。
- 聚类算法计算类似度，从而对类似实例进行聚类。
- 聚类分析中最常用的类似度尺度是距离。
- 最常用的聚类算法是 k-均值和自组织映射。
- 关联规则挖掘用于发现总在一起的两个或多个项目（或事件、概念）。
- 关联规则挖掘常常被当做购物篮分析。
- 最常用的关联规则挖掘算法是 Apriori，该算法通过自底向上的方法识别频繁项集。
- 基于其支持度和置信度，对关联规则进行评估。
- 有很多商业的和免费的数据挖掘工具。
- 最常用的商业数据挖掘工具是 SPSS PASW 和 SAS Enterprise Miner。
- 最常用的免费数据挖掘工具是 Weka 和 RapidMiner。

关键术语

adaptive resonance theory 自适应共振理论
algorithm 算法
Apriori algorithm Apriori 算法
area under the ROC curve ROC 曲线下面积
Artificial Neural Network（ANN，人工神经网络）
associations 关联
axone 轴突
backpropagation 反向传播
bootstrapping 拔靴法
business analyst 业务分析师
categorical data 分类数据
chromosome 染色体
classification 分类
clustering 聚类
confidence 置信度
connection weight 连接权值
CRISP-DM 跨行业数据挖掘标准过程
data mining 数据挖掘
decision trees 决策树
dendrites 树突
discovery-driven data mining 发现驱动数据挖掘
distance measure 距离测度
entropy 熵
fuzzy logic 模糊逻辑

genetic algorithm 遗传算法
Gini index 基尼系数
heuristics 启发式
hidden layer 隐含层
hypothesis-driven data mining 假设驱动数据挖掘
information gain 信息增益
interval data 区间数据
k-fold cross validation k折交叉验证
Knowledge Discovery in Databases（KDD，数据库知识发现）
Kohonen's self-organizing feature map Kohonen自组织特征映射
learning algorithm 学习算法
link analysis 链接分析
machine learning 机器学习
Microsoft Enterprise Consortium Microsoft 企业联盟
Multi-Layered Perceptron（MLP，多层感知器）
neural computing 神经计算
neural network 神经网络
neurons 神经元
nominal data 名称数据
numeric data 数值数据
ordinal data 序数数据
pattern recognition 模式识别
prediction 预测
Processing Elements（PE，处理单元）
Rapidminer 快速挖掘
ratio data 比率数据
regression 回归
result（outcome）variable 结果变量
SAS Enterprise Miner SAS 企业版挖掘工具
SEMMA 抽样、探索、修正、建模、分析
sensitivity analysis 灵敏度分析
sequence mining 序列挖掘
sigmoid function S型函数
simple split 简单拆分
SPSS PASW Modeler SPSS 预测分析软件建模
summation function 求和函数
supervised learning 监督学习
support 支持度
Support Vector Machines（SVM，支持向量机）
synapse 突触
transformation function 转换函数
unsupervised learning 无监督学习
Waikato Environment for Knowledge Analysis（Weka，怀卡托智能分析环境）

讨论题

1. 给出数据挖掘的定义。为什么数据挖掘有很多名称和定义？
2. 近来数据挖掘广泛应用的原因有哪些？
3. 讨论：组织在决策购买数据挖掘软件之前，应该考虑哪些问题？
4. 如何区别数据挖掘和其他分析工具技术？
5. 讨论主要的数据挖掘方法。这些方法之间存在哪些根本差别？
6. 数据挖掘主要有哪些应用领域？讨论这些领域的共同点。
7. 为什么需要标准化的数据挖掘流程？应用最广泛的数据挖掘流程有哪些？
8. 讨论两个应用最广泛的数据挖掘流程之间的差别。
9. 数据挖掘流程仅仅是活动的顺序集合吗？
10. 为什么需要数据预处理？数据预处理的主要任务和相关技术有哪些？
11. 讨论分类模型评估背后的逻辑。
12. 分类和聚类的主要区别是什么？用具体例子说明。
13. 抛开本章讨论的内容，还有哪些地方可以使用关联规则挖掘？
14. 组织在进行数据挖掘软件购买决策前，应考虑哪些问题？
15. 什么是人工神经网络 ANN？试比较生物神经网络和人工神经网络。

练习

Teradata 大学和其他动手练习

1. 访问 teradatastudentnetwork.com，找出关于数据挖掘的案例，并描述该领域的最新进展。

2. 转到 teradatastudentnetwork. com 或老师给出的 URL，找到数据挖掘相关网络课程。特别地，观看由 C. Imhoff 和 T. Zouges 给出的网络课程。然后回答下列问题：
   a. 数据挖掘有哪些有趣的应用？
   b. 组织开始一项数据挖掘项目会有哪些类型的收益和支出？
3. 本练习的目标是建立一个能够识别输入（预测变量），从而区分高风险客户和一般客户的模型（基于过往的客户模式），然后使用这些输入预测新的高风险客户。这样例在本领域是很典型的。
   练习用到的样本数据可以从在线文件 CreditRisk. xlsx 的 W4. 1 中找到。该数据集包括 425 个实例、过往的 15 个变量和用于各种原因从银行贷款的当前客户。数据集包含客户相关信息，例如财务状况、贷款理由、职业、人口统计信息和作为输出结果的因变量——信用状况，基于机构过去的经验，将每个实例按好坏分类。
   取 400 个实例作为训练实例，其余 25 个实例留出来用于测试。建立学习问题特征的决策树，然后在其余 25 个实例上测试模型性能，并给出模型学习和测试性能报告。起草一份报告，对决策树模型、训练参数和测试集性能结果进行识别。可以使用任意决策树软件。
   （本练习经 Statsoft 公司准许，基于来自 ftp. ics. uci. edu/pub/machine-learning-databases/statlog/german 的数据集，以 CreditRisk 重命名并进行了改动。）
4. 在本练习中，您需要复制本章开篇场景中说明的票房预测模型。下载在线文件 MovieTrain. xlsx，在 W4. 2 中可以找到训练数据集。它包括 184 条记录，格式是 Microsoft Excel 文件。使用开篇场景中给出的数据描述理解该领域及所要解决的问题。仔细选择自变量，并建立至少 3 种分类模型（例如，决策树、逻辑回归和神经网络）。使用 10 折交叉确认和百分比分割技术对结果精确度进行比较。其中要用到混淆矩阵，并对结果进行评述。在测试集上对建立的模型进行测试（详见在线文件 W4. 3，MovieTest. xlsx，包含 29 个记录）。应用不同模型对结果进行分析，指出哪个分类模型是最好的，并给出支持该结论的分析结果。

小组作业和角色扮演

1. 调查新的数据获取工具，例如无线射频识别（Radio Frequency Identification，RFID）标签是如何帮助组织精确识别客户和对客户进行分类的，以进行精准营销的。很多这种应用都牵涉到数据挖掘。浏览文献及 Web 网页，然后提出 5 种新的潜在 RFID 技术数据挖掘应用。如果国家法律要求在每个人体内都植入这种标签，以建立国家身份识别系统，将会出现哪些问题？
2. 采访你所在大学的管理人员或您所在组织的主管，了解数据仓库、数据挖掘、在线分析处理和可视化 BI/DSS 工具是如何帮助管理的。撰写报告，描述你发现结果，并给出成本估算和效益分析。
3. 在网址 ics. uci. edu/ ~ mlearn/MLRepository. html 有一个非常好的数据资源库，已被用于测试很多机器学习算法的性能。有些数据集是为了测试当前机器学习算法的局限性，并与新的学习算法进行性能比较。不管怎样，可以使用其中的一些小数据集研究任意一个数据挖掘软件或本书配套软件（例如 Statistica DataMiner）的功能。从该数据资源库中下载至少一个数据集（例如信贷审查数据库或住房数据库），尝试使用决策树算法或聚类算法。基于你的结果编写一份报告。（其中的一些练习甚至可以作为一学期的项目，用于撰写学期论文。）
4. 考虑以下数据集，其中包括 3 个属性和 MBA 项目录取结果分类：

GMAT	GPA	GMAT 数量成绩（百分数）	录取结果	GMAT	GPA	GMAT 数量成绩（百分数）	录取结果
650	2. 75	35	否	400	3. 85	45	否
580	3. 50	70	否	640	3. 50	75	是
600	3. 50	75	是	540	3. 00	60	?
450	2. 95	80	否	690	2. 85	80	?
700	3. 25	90	是	490	4. 00	65	?
590	3. 50	80	是				

a. 使用显示的数据，手工建立你自己的专家决策制定规则。

b. 使用基尼系数建立决策树。可以使用手工计算或电子表格进行基本运算。

c. 使用自动决策树软件程序为同组数据建立决策树。

5. 本练习的目的是建立模型，使用多个地图度量指标预测森林植被类型。给定数据集（在线文件 W4.1）包括北科罗拉多罗斯福国家森林公园的4个荒野地区数据。总共12个地图度量指标被用作自变量；7种主要类型的森林植被被用作因变量。下面的表格对这些自变量和因变量进行了简略描述。

这是一个很好的多类分类问题例子。数据集相当大（有581 012个不同实例），并且功能丰富。建模者可以采取必要的决策进行数据预处理，并找到最有可能的预测因素。使用你最喜欢的工具建立模型，并以书面报告形式记录过程细节和你的经验。在报告中使用屏幕截图阐释重要的有趣发现。讨论并解释您在此过程中所做出的决策。

序号	名称	描述
		自变量
1	海拔	海拔（米）
2	方位	方位角度方向
3	坡度	斜坡度
4	水文水平距离	地表水最近水平距离
5	水文垂直距离	地表水最近垂直距离
6	公路水平距离	公路最近水平距离
7	坡面阴影（上午9点）	夏至日上午9点坡面阴影指数
8	坡面阴影（正午）	夏至日正午坡面阴影指数
9	坡面阴影（下午3点）	夏至日下午3点坡面阴影指数
10	着火点水平距离	最近野火着火点水平距离
11	荒野保护区（4个二元变量）	荒野保护区标记
12	土壤类型（40个二元变量）	土壤类型标记
序号		因变量
1	植被类型（7种不同类型）	植被类型标记

a 数据集更多细节（变量和观察值）可从在线文件中找到。

重复使用该数据集不受限制，只需保留 Jock A. Blackard 和科罗拉多州立大学的版权声明。

网络练习

1. 访问 ca. ualberta. ca/ ~ aixplore/的人工智能探索博物馆。点击“Decision Tree”（决策树）链接。阅读篮球赛统计资料描述，仔细观察数据并建立决策树。给出你对决策树精确度的印象报告。同时，研究不同算法的效果。
2. 从 fairisaac. com 和 egain. com 开始，调查一些数据挖掘工具和供应商。查阅 dmreview. com，找出一些本章未提及的数据挖掘产品和服务供应商。
3. 找出最近的一些数据挖掘成功应用案例。访问几个数据挖掘供应商 Web 站点，查找成功案例。撰写报告，总结5个新的案例研究。
4. 访问供应商 Web 站点（特别是 SAS、SPSS、Cognos、Teradata、Statsoft 和 Fair Isaac），查找 BI（OLAP 和数据挖掘）工具的成功应用案例。这些不同成功案例间有什么共同点？有什么区别？
5. 登录 statsoft. com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web 挖掘技术？
6. 登录 sas. com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web 挖掘技术？
7. 登录 spss. com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web 挖掘技术？
8. 登录 nd. com。下载至少3份神经网络客户成功应用研究案例。这些不同的成功案例有何共同点？有何差别？
9. 登录 teradata. com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web 挖掘

技术？

10. 登录 fairisaac.com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web挖掘技术？
11. 登录 salfordsystems.com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web挖掘技术？
12. 登录 rulequest.com。下载至少3份应用白皮书。哪些应用采用了本章讨论过的数据/文本/Web挖掘技术？
13. 登录 kdnuggets.com。探索关于应用和软件的章节。找到至少3种数据挖掘和文本挖掘的额外软件包。

本章结尾应用案例

数据挖掘帮助通信公司为客户定制产品组合

背景

argonauten 360°咨询集团帮助企业建立并完善成功的客户关系管理战略。该公司采用关联营销促进和相关客户的对话，从而创造价值。BMW、Allianz、Deutsche 银行、Gerling 和 Coca-Cola 等很多其他企业都是该咨询集团的客户。

问题

作为电信等行业领先的咨询公司，argonauten 360°日常工作的一个常规部分就是应用先进有效的分析技术进行客户评分、聚类和客户终身价值计算。由于每个项目都会提出一组新的特定的情况、数据情景、障碍和分析挑战，所以要求分析工具既灵活又强大是一种苛求。因此，需要由尖端有效而灵活的数据挖掘功能来弥补现有工具集。另一个重要的考虑因素是希望解决方案能够快速产生投资回报。方案必须易于应用，具有快速的学习曲线，以使分析师能快速掌握即使是最先进的分析过程。

解决方案

公司需要一组统一的、易于使用的分析工具集，具备大范围的建模功能和简单的配置选项。不同的建模任务需要学习不同的工具，这会严重阻碍咨询的效率和效果。因此，该公司偏向统一的解决方案环境，功能范围从对任意媒介（例如，数据库、在线数据资源库、文本文件和 XML 文件）的数据访问，到在大范围 BI 系统中部署复杂数据挖掘解决方案。

经过12个月的大量数据挖掘工具评估，该公司选择了 Statistica Data Miner（由 StatSoft 公司提供）。据该公司行政主管说，这是因为这种工具提供了理想的功能组合，能够满足几乎所有分析师的需求，而且用户界面友好。

一个创新项目例子

在欧洲，所谓的“预付费电话”服务在手机和普通电话用户中非常流行。这种规划方案对基本服务不收费或收费很低，主要按照实际通话时长收费。这种业务很具竞争力，预付费通信服务供应商在很大程度上依赖其每分钟通话费率的吸引力。这些费率排名被广泛刊登。关键在于要在费率表最低排名所前5位的同时，获得最高的利润。由于这种形势造成的竞争环境，大众普遍认为“在这个市场上实际上不存在价格弹性（供应商能够获得哪怕是最少的附加收益而不流失客户）；即使存在这种价格弹性，也肯定不能预测。”然而，argonauten 360°的咨询顾问使用 Statistica 数据挖掘工具对现有数据的分析正好：这种普遍看法是错误的！事实上，他们的成功分析为 argonauten 360°赢得了预付费服务业务的供应商领先地位。

分析

分析行为基于每分每秒电话流量的描述数据。具体来说，分析针对一年中的通话分钟销售。为了获得最好的分析效果，建立了20种不同种类的评估模型组合。每个模型使用一种回归类型的数学表达函数来预测长期趋势，然后在更高层次的元模型中结合个别模型。所有的具体通话时间段都被仔细分析，识别出各时间段的价格敏感度和竞争压力。

2个月后的结果

在将数据挖掘所得模型投入应用以前，启发式“专家观点”首先被用来预测后续2个月的通话时长。

使用 Statistica Data Miner，这些预测的精确度得到了显著改善，同时错误率下降了一半。由于分钟通话流量（时长）数量非常巨大，这清楚地证明了先进分析策略在解决这种类型问题时的效果和潜在好处。

在客户站点实施解决方案

目前，预付费服务供应商正在使用这种方案对最佳的通话费率进行预测和模拟。该系统被 argonauten 360°公司配置成为一个完全的交钥匙（只需按一下按钮）方案。使用这种方案，预付费服务提供商能够以高得多的精确度来预测价格弹性敏感的市场需求，给出"正确的"费率，从而获得关键竞争优势。

在下一阶段，一种类似仪表盘的系统将进一步完善该系统，使其能够自动比较预测结果和观测数据。argonauten 360°保证该系统在必要时能够更新模型参数估算，以适应市场变化。这样，预付费服务供应商不需要任何分析技能，就能够实现可靠的复杂需求预测和费率模拟系统。这在以往被认为是不可能的。这是应用数据挖掘技术，在高度竞争商业环境中获得竞争优势的一个很好的典范。

本章结尾应用案例的问题

1. 为什么咨询公司更有可能使用数据挖掘工具和技术？他们的具体价值诉求是什么？
2. 为什么对 argonauten 360°公司来说，选用一种具备所有建模功能的综合工具非常重要？
3. argonauten 360°公司帮助预付费服务供应商解决了什么问题？
4. 你还能想出其他数据挖掘可以解决的电信企业问题吗？

来源：StatSoft，"The German Consulting Company argonauten 360° Uses Statistica Data Miner to Develop Effective Product Portfolios Custom- Tailored to Their Customers," statsoft. com/company/success _ stories/pdf/argonauten360. pdf（accessed on May 25，2009）.

参考文献

Bhandari, I., E. Colet, J. Parker, Z. Pines, R. Pratap, and K. Ramanujam. (1997). "Advanced Scout: Data Mining and Knowledge Discovery in NBA Data." *Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp. 121–125.

Bolton, R. J. (2002, January). "Statistical Fraud Detection: A Review." *Statistical Science*, Vol. 17, No. 3, p. 235.

Chan, P. K., W. Phan, A. Prodromidis, and S. Stolfo. (1999). "Distributed Data Mining in Credit Card Fraud Detection." *IEEE Intelligent Systems*, Vol. 14, No. 6, pp. 67–74.

CRISP-DM. (2009). "Cross-Industry Standard Process for Data Mining (CRISP-DM)." **crisp-dm.org** (accessed January 2010).

Davenport, T. H. (2006, January). "Competing on Analytics." *Harvard Business Review*.

Delen, D. (2009). "Analysis of Cancer Data: A Data Mining Approach." *Expert Systems*, Vol. 26, No. 1, pp. 100–112.

Delen, D., R. Sharda, and P. Kumar. (2007). "Movie Forecast Guru: A Web-based DSS for Hollywood Managers." *Decision Support Systems*, Vol. 43, No. 4, pp. 1151–1170.

Delen, D., G. Walker, and A. Kadam. (2005). "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods." *Artificial Intelligence in Medicine*, Vol. 34, No. 2, pp. 113–127.

Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. (1996). "From Knowledge Discovery in Databases." *AI Magazine*, Vol. 17, No. 3, pp. 37–54.

Gillespie, G. (2004, November). "Data Mining: Solving Care, Cost Capers." *Health Data Management*. **findarticles.com/p/articles/mi_km2925/is_200411/ai_n8622757** (accessed May 2009); and "Highmark Enhances Patient Care, Keeps Medical Costs Down with SAS." **sas.com/success/highmark.html** (accessed April 2006).

Hastie, T., R. Tibshirani, and J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Haykin, S. S. (2009). *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Kohonen, T. (1982). "Self-organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics*, Vol. 43, No. 1, pp. 59–69.

Liebowitz, J. "New Trends in Intelligent Systems." Presentation made at University of Granada, **docto-si.ugr.es/seminario2006/presentaciones/jay.ppt** (accessed May 2009).

Nemati, H. R., and C. D. Barko. (2001). "Issues in Organizational Data Mining: A Survey of Current Practices." *Journal of Data Warehousing*, Vol. 6, No. 1, pp. 25–36.

Nischwitz, R., M. Goldsmith, M. Lees, P. Rogers, and I. MacLeod. "Developing Functional Malt Specifications for Improved Brewing Performance." The Regional Institute Ltd., regional.org.au/au/abts/1999/nischwitz.htm (accessed December 2009).

Quinlan, J. R. (1986). "Induction of Decision Trees." *Machine Learning*, Vol. 1, pp. 81–106.

Salford Systems. "The Duke/KDD Tendata Churn Modeling Tournament." **salford-systems.com/churn.php** (accessed April 20, 2009).

SEMMA. (2009). "SAS's Data Mining Process: Sample, Explore, Modify, Model, Assess." **sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html** (accessed August 2009).

Sharda, R., and Delen, D. (2006). "Predicting Box-office Success of Motion Pictures with Neural Networks." *Expert Systems with Applications*, Vol. 30, pp. 243–254.

Shultz, R. (2004, December 7). "Live from NCDM: Tales of Database Buffoonery." **directmag.com/news/ncdm-12-07-04/index.html** (accessed April 2009).

Skalak, D. (2001). "Data Mining Blunders Exposed!" *DB2 Magazine*, Vol. 6, No. 2, pp. 10–13.

StatSoft. (2006). "Data Mining Techniques." **statsoft.com/textbook/stdatmin.html** (accessed August 2006).

Thongkam, J., G. Xu, Y. Zhang, and F. Huang. (2009, in press). "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space." *Expert Systems with Applications*.

Tseng, W. S., H. Nguyen, J. Liebowitz, and W. Agresti. (2005, January)."Distractions and Motor Vehicle Accidents: Data Mining Application on Fatality Analysis Reporting System (FARS) Data Files." *Industrial Management & Data Systems*, Vol. 105, No. 9, pp. 1188–1205.

Wilson, C. L., and L. Threapleton. (2003, May 17–22). "Application of Artificial Intelligence for Predicting Beer Flavours from Chemical Analysis." *Proceedings of the 29th European Brewery Congress*, Dublin, Ireland, **neurosolutions.com/resources/apps/beer.html** (accessed January 2010).

Wilson, R., and R. Sharda. (1994). "Bankruptcy Prediction Using Neural Networks." *Decision Support Systems*, Vol. 11, pp. 545–557.

Yu, W., D. N. Jutla, and S. C. Sivakumar. (2005). "A Churn-Strategy Alignment Model for Managers in Mobile Telecom." *Proceedings of the Communication Networks and Services Research Conference*, IEEE Publications, pp. 48–53.

Zaima, A. (2003). "The Five Myths of Data Mining." *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, Chatsworth, CA: Data Warehousing Institute, pp. 42–43.

Zdanowic, J. S. (2004, May). "Detecting Money Laundering and Terrorist Financing via Data Mining." *Communications of the ACM*, Vol. 47, No. 5, p. 53.

文本挖掘与 Web 挖掘

学习目标

- 介绍文本挖掘并理解文本挖掘的应用
- 区分文本挖掘和数据挖掘的差别
- 理解文本挖掘的不同应用领域
- 了解文本挖掘实现的过程
- 理解描述文本型数据结构的不同方法
- 介绍 Web 挖掘的含义、目的和优势
- 理解 Web 挖掘的 3 个不同的分支
- 理解 Web 内容挖掘、Web 结构挖掘和 Web 日志挖掘

本章综合性地描述了与商务智能和决策支持系统有关的文本挖掘和 Web 挖掘的概念。Web 挖掘和文本挖掘本质上都是数据挖掘的派生词。因为文本数据和网络流量数据在同一量级的容量上比数据库中结构化的数据增加的速度要快，所以了解一些用于处理海量非结构化数据的技术非常重要。

开篇场景：文本挖掘与安全和反恐

假设你是美国大使馆里解救人质的一个决策者，你正设法查清“非是恐怖分子的首领”，“本次恐怖袭击背后的阴谋是什么”以及“这个组织是否有可能袭击其他的大使馆”。尽管你有获得大量信息的渠道，但是在这种情形下你很难有效利用如此大量的信息并做出更好的决策。在这个取决于准确性和实时智能的危急关头，计算机怎么发挥作用呢？美国国防部高级研究计划局（Defense Advanced Research Project Agency，DARPA）整体情报识别项目的一个子课题——Genoa[⊖]，利用先进工具和技术快速分析与当时情形相近的信息，从而支持更好的决策。尤其是，Genoa 提供了知识发现工具，用于从相关信息资源中进行更好的“挖掘”，从而发现可做出响应的信息模式（例如，相关知识领域）。

Genoa 所面临的一个挑战是使最终用户容易地利用从分析工具中发现的知识，并将它以简明有用的形式嵌入智能产品中。一个为公众利益服务的非盈利创新研究组织 MITER（mitre.org），被委任从事开发文本挖掘的基础软件系统来迎接这个挑战。这个系统允许用户选择各种文本挖掘工具，并且使得用户在点击几下鼠标的情况下，就可以创建一个复杂的过滤器，该过滤器能够实现任何所需的知识发现功能。过滤器用于输入信息并将其转换为更简要更有用的形式。过滤器也可以清除信息里与自己研究的内容不相关的部分。

例如，在前面所讨论的危机情形下，分析家可以利用文本挖掘工具从搜集到的大量新闻资源中挖掘重要的信息块。文本挖掘工具的这种应用可以理解为查看 TopCat，TopCat 是 MITRE 开发的一个系统，能够识别一些文件内容的不同主题，并且为每个话题找出核心词。TopCat 利用规则挖掘技术来识别人、组织、位置和事件之间的关系（在图 5-1 中分别叫做 P、O、L 和 E）。

⊖ Genoa 课题起始于 1997 年，2003 年转变为 Genoa II，所属的整体情报识别项目 2003 年更名为 Topsail，两个项目都因为是政府主导的侵犯个人隐私和人权的阴谋案而遭到非议。

通过创建"话题簇"将这些关系分组，就像图5-1中分为三组，它们是将6个月的全球新闻分类，共包含了印刷品、广播、视频等60 000多条新闻事件。

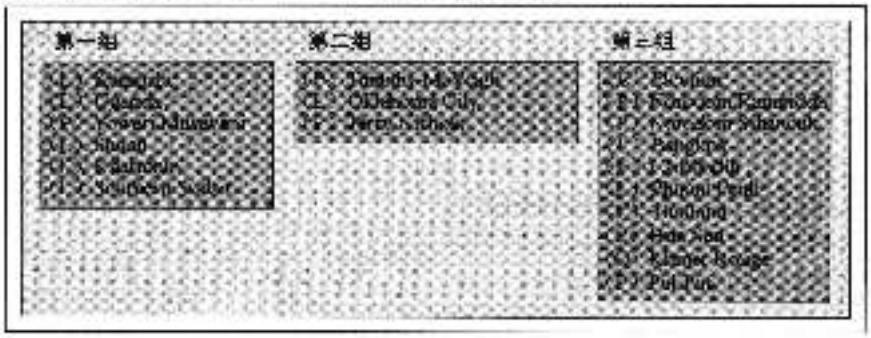


图5-1 从60 000个新闻事件提取出的不同主题组

来源：Mitre Corporation，www.mitre.org（accessed May 20，2009）

这个工具可以帮助分析家发现知识，例如，分析恐怖事件中恐怖分子之间的联系，就像"McVeigh和Nichols是同一个组织的"，从而可以进行更深入的分析。反过来，这个工具还可以起到发现新知识的作用，形成分析模型，来预测一个特别的恐怖分子组织是否有可能在未来的几天袭击其他地方。同样的，第三组话题可以揭示柬埔寨选举中的重要人物，所发现的这个信息可以帮助预测柬埔寨的情况是否会引发危机从而对美国人在当地的利益产生潜在影响。

假设用户想在上段提到的第3个话题（柬埔寨选举）中更多地了解这些重要人物，分析家就利用一个主题探测过滤器和人物传记摘要过滤器（就像TopCat中识别人的关系），从各种相关主题中搜索到重要人物的行为，而不是通过翻读各种相关主题的成千上万的文字来搜集信息。这个组合的结果将产生一个短的主题点睛总结。

摘要过滤器是由DARPA基金支持开发的，它利用有效的语法分析、使用同义词典和一些简单的自然语言处理技术，从大量文件中对人的描述进行分类和聚合。它还可以根据现有人名、住址和文件中一些类似的条件，比如出现的频率及与其他条件之间的关系，从一些资料中提取出和这些人相关的重要语句。

TopCat中的摘要过滤器还具有和MITRE新闻广播导航器相类似的功能。利用这个功能可以连续不断地从新闻中获取信息，提取命名实体和关键词，并识别它们中有趣的改写本和句子。摘要过滤器具有详细说明目标长度和下降率的参数，这样就可以将信息概括为不同长度的摘要。例如，长的摘要有可能包含其他人物（如Pol Pot）的信息。

这个例子说明了如何利用现在的知识挖掘工具进行文本挖掘，如TopCat摘要过滤器可以发现不同层次信息之间的重要联系。实施TopCat中使用的组件开发方法可以很容易地将这些过滤器集成为智能产品，例如自动形成智能报表、简报和仪表盘。这些摘要过滤器可以链接到简报网页上的一个特定部分，这些网页可以被相互合作的分析家分享。当一个文件或者一个文件夹被某个特定部分的过滤器发现，这个过滤器处理文件中的文本和数据，在该特定区域出现文字概要或者以可视化的图像展现的信息。

开篇场景的问题

1. 在危机情况下如何利用文本挖掘技术？
2. 什么是Genoa项目？Genoa项目背后的动机是什么？
3. 什么是TopCat？TopCat做什么？
4. 什么是摘要过滤器？

5. 阐述在反恐中未来的文本挖掘工具。

我们从开篇场景中能学到什么

近十几年，文本挖掘工具已经成为国家智能措施的一部分，如几十年来的整体情报意识项目。在开篇场景中，DARPA 和 MITRE 相互合作来开发自动化过滤器，并使文本型的信息资源及时转化为可付诸行动的信息。利用基于组件的架构，使这个复杂系统中的部分结构在独立于其他部件的情况下得以修改、使用和重用。通过联系、归类、聚集分析这些基于文本的文件分析工具，展现了从大量新闻中获取知识的力量。智能领域所获取的成就标志着在不久的将来知识发现工具和技术的发展潜力。

来源：MITRE Corporation, mitre. org(accessed on May 20,2009); J. Mena, *Investigative Data Mining for Security and Criminal Detection*, Elsevier Science, Burlington, MA, 2003.

5.1 文本挖掘的概念和定义

我们所处的信息时代，大量数据和信息呈现快速增长的特点，这些数据和信息通过电子媒介进行收集、存储、获得。大量的商业数据存储在非结构化的文本文件中。Merrill Lynch 和 Gartner 的一个研究表明，获取和存储的企业数据中有 85% ~90% 是非结构化的（McKnight，2005）。相同的研究也表明这些非结构化的数据每 18 个月在数量上就会翻一倍。在今天的商业世界中，知识就是力量，而且知识来源于数据和信息，商务需要有效地挖掘文本数据资源，将必要的知识用于更好的决策和领导，相对于其他落后的商务，这些商务占据了优势。这就是现今商务需要文本挖掘的原因。

文本挖掘（也叫文本数据挖掘或文本型数据库中的知识发现）是指半自动化地从大量的非结构化数据资源中提取模式（即有用的信息或知识）的过程。回想数据挖掘是指从存储在结构化的数据库的数据中识别出有效的、新颖的、潜在有用的、最终可理解的模式的过程，这些数据以分类的、顺序的、连续变量的结构组织为记录的形式。文本挖掘和数据挖掘一样，它们具有相同的目的并利用相同的处理过程。但是对于文本挖掘，处理过程的输入是非结构化（或者少量的结构化）数据文件，这些数据文件包括 Word 文献、PDF 文件、文本摘录、XML 文件等。实质上，文本挖掘可以被看做是这样一个处理过程（包括两个主要步骤），首先是将这些基于文本的数据资源进行结构化处理，然后利用数据挖掘的技术和工具从这些结构化的文本数据中提取相关的信息和知识。

文本挖掘在涉及大量文本型数据的领域占有很大优势，例如，法律（法庭命令）、学术研究（研究论文）、财务（年报）、医学（诊断结果）、生物学（分子间相互作用）、技术（专利文件）和市场（顾客意见）。比如顾客之间以自由形式存在的相互的抱怨（或赞扬）和索赔要求的文本资料，可以客观地看出产品的好坏和服务态度的满意程度，这些资料不是很完美但是可以促进产品和服务质量的改进。同样，市场推广计划和焦点小组也能产生大量的数据。不用编纂形式限制产品和服务反馈，换句话说，顾客能够用自己的话来体现他们认为的公司产品和服务意见。非结构化文本处理产生重要影响的另一个重要领域是电子通信和电子邮件领域。文本挖掘不仅可以对垃圾邮件进行分类和过滤，还可以根据邮件的重要程度自动地将邮件优先处理，还可以进行自动回复（Weng and Liu，2004）。下面是文本挖掘技术的主要应用领域：

- **信息提取** 通过类型匹配，利用事先定义的顺序，识别出文本中的关键的短语和关系。
- **主题跟踪** 根据用户的形象和对用户一些行为的记载，可以预测出该用户感兴趣的一些文件。
- **摘要** 一个文件的摘要可以节省读者的时间。
- **分类** 了解每个文献的主题，并根据主题将这些文件放在之前定义好的分类中。
- **聚类** 将没有分类的文件放入到类似的分类中。

- **概念关联** 将具有相同概念的文件联系在一起，这样可以帮助用户找到利用传统方法所没有找到的信息。
- **问题解答** 通过知识类型匹配找到所给问题的最好答案。

技术前沿5.1解释了文本挖掘中的一些术语和概念，应用案例5.1描述了专利分析中文本挖掘的使用。

技术前沿5.1 文本挖掘术语

下面是文本挖掘中所涉及的常见术语：

- **非结构化数据（相对于结构化数据）** 结构化数据具有定义好的格式，通常是简单的数据以记录（分类的、顺序的、连续变量）的形式存储在数据库中。相反，非结构化数据没有预先定义好的格式，存储在文本文件中。实际上，结构化数据用于计算机处理而非结构化数据用于人类处理和理解。
- **语料库** 字面上，语料库是指大量有组织的文本集合（现在通常是以电子化的方式存储和处理），用于进行知识发现。
- **术语** 术语是指从利用自然语言的形式直接从特殊领域的语料库中提取出来的一个词或多词短语。
- **概念** 概念是利用人工、统计并基于一定规则或者多种分类的方法，从大量文件中得到的结果。和术语相比，概念是更高级别抽象的结果。
- **词根** 将词语的变形转化为原型（或者基本形式、根源）。如词语 stemmer、stemming、stemmed 的词根都是 stem。
- **无用词** 无用词（或噪声词）是指通过自然语言数据（如，文本）处理之前或之后过滤掉的词语。尽管没有列出来这些无用词，但是大多数自然语言处理工具列出了冠词（如，a、an、the、of 等）、附属动词（如，is、are、was、were 等）和专业性词语，这些类型的词语没有差异。
- **同义词和多义词** 同义词是指在语句构成上不同（如，拼写不同），但是在意义上相同或相近的词语（如电影和影片）。相比之下，多义词，也叫异动同名词，是指在构成上相同（如拼写一样），但是意义不同的词语（如 bow 可以解释为"向前弯曲"、"船头"、"射箭武器"或者"系东西的丝带"）。
- **标记处理** 一个标记是指一个句子中的文字分类块。文字分类块是根据功能标志进行分类。这项任务对字块的意义就是标记处理。标记只是结构化文本中一个有用的部分，它可以表现为任何形式。
- **术语词典** 对一个专业领域的术语进行汇总，可用于对语料库中提取出的术语的范围进行限定。
- **词频** 一个词语在一个特殊文献中出现的次数。
- **词性标注** 一个文本中词的标注要和词语的词性（如，名词、动词、形容词和副词）相关，词性取决于词语的意思、上下文环境。
- **构词法** 语言学和研究词语结构（词语的构成形式）的自然语言领域的一个分支。
- **文献－术语相关矩阵（频数矩阵或术语文献矩阵）** 根据术语出现的频率和文献之间的关系制作成的以表格形式体现的模式，行代表术语，列代表文献，单元格的内容代表术语在文献中出现的频率。
- **奇异值分解（潜在语义索引）** 一种降维简化方法，指利用一个类似于主成分分析的矩阵处理方法，用一个大小合适频数表示方式代替文献术语矩阵中的频数。

应用案例5.1 专利分析中的文本挖掘

专利是指国家赋予一项公开发明创造的首创者一定期限内所拥有的独享权益权利（不同的国家专利授予的程序、专利权所有人的要求、独享权利的范围都不同）。这些公开的发明对将来的科学和技术有促进作用。行知分析的话，专利文件可以帮助人们认识新兴的技术、鼓励新颖的解决方法、促进共生合作、增强企业的能力和局限的全面意识。

专利分析是指利用分析技术从专利数据库中提取有用的知识。国家或国家组织主要的专利数据库（美洲、欧洲、日本等）每年都增加上千个专利。处理如此庞大的半结构化数据（专利数据库中通常包括部分结构化和部分文本型数据）几乎是不可能的。而利用半自动化的软件工具是简化处理这些大型数据库的一种方式。

专利分析的典型案例

Eastman Kodak 雇佣了世界范围内5 000多个科学家、工程师和技术人员。20世纪，这些知识工作者和以前的知识工作者们声称世界上专利数量排名前10位的企业几乎拥有20 000个专利。由于商业界连续不断地变化，企业要想成功（或者仅仅保持生存）就要依靠自己的能力来应用超过一个世纪的有价值的科学技术，并将这些技术应用到新领域，同时利用专利技术保护这些新应用。

Kodak 欣赏专利的价值，他不仅自己发明专利并且还研究他人发明的专利。通过依靠分析家和现有的软件工具（包括 ClearForest 公司的专业文本挖掘工具），Kodak 总是深入到各种数据资源（专利数据库、新的成就和产品发布）中，从而形成具有竞争力的思想。就像 Kodak，对专利的合理分析可以给企业带来以下优势：

- 使企业具备有竞争力的智力。了解竞争对手的情况可以帮助企业做出反击。
- 可以帮助企业做出关键性的商业决策，例如，生产什么产品，如何安排生产线，深入开发哪种技术或者想要什么样的合并和并购方式。
- 帮助企业发现和招聘最好最聪明的新人才，名字出现在专利开发上的那些人对企业的成功起到重要的作用。
- 帮助企业发现那些非法利用自己专利的行为，采取行动保护自己的专利。
- 帮助企业认识到互补的专利，从而帮助企业找到合作伙伴或者促进企业间的合作关系。
- 可以防止企业竞争者生产类似的产品，保护企业的专利诉讼权。

将专利分析作为企业丰富知识的手段和战略的武器（起到防御和攻击的双重作用），Kodak 不仅仅能够生存下来，而且通过创新和不断的改进在市场上占领了优势。

来源：P. X. Chiem, "Kodak Turns Knowledge Gained About Patents into Competitive Intelligence," *Knowledge Management*, 2001, pp. 11－12; Y-H. Tseng, C-J. Lin, and Y-I. Lin, "Text Mining Techniques for Patent Analysis," *Information Processing & Management*, Vol. 43, No. 5, 2007, pp. 1216－1245.

5.1 节复习题

1. 什么是文本挖掘？文本挖掘和数据挖掘有何不同？
2. 为什么文本挖掘作为一个 BI 工具越来越重要？
3. 文本挖掘应用于哪些重要的领域？

5.2 自然语言处理

早期的一些文本挖掘应用表现为一个叫做字袋模型的简单形式，字袋模型是一个基于文本文件的集合，它将这些文件分为两种或更多预先测定的种类，或者将它们进行自然分类。在这个字袋模型中，文本比如句子、段落或者整个文件表现为词语的集合形式，而不考虑语法和词语的顺序。在一些简单的文本分类工具中，字袋模型仍被采用。例如，垃圾邮件过滤中，电子邮件信息被看做是一些无序词语的集合（一个字袋模型），这些词语和之前定义好的两个不同的"袋"进行对比。其中一个袋中是组成垃圾邮件的词语，另外一个袋中是组成合法邮件的词语。尽管一些词语在这两个袋中都可以找到，但是垃圾邮件包中包含的与垃圾邮件相关词语（如库存、伟哥和购买）的出现的频率比合法邮件袋中所包含的词语（如用户的朋友或工作地点）出现的频率要高。将一个电子邮件的"字袋模型"和这两个袋之间进行对比，叙述语决定这个电子邮件是属于垃圾邮件还是合理邮件。

当然，我们不会利用一些没有规律和结构的词语，我们将词组成句子是要符合语义和语法结构的。因此，自动化的技术（如文本挖掘）超出了字袋模型所能处理的能力并且在字袋模型中引入了越来越多的语义结构。现有文本挖掘技术包含了许多先进的功能，这些功能通过自然语言处理体现出来。

可以看出，字袋模型的方式在文本挖掘工作（分类、聚集、联系等）中不能提供令人满意的信息。这可以用基于证据的医学示例说明。基于证据的医学中的一个关键类作是将最有利用价值的研究发现利用到诊所诊所程序中，包括评估收集的印刷资料的有效性和相关性。马里兰大学的几名研究者利用字袋模型方法对收集到的资料进行评估（Lin and Demner，2005）。他们利用了著名的机器学习法，对从MEDLINE（医学文献分析与检索在线系统）收集的50万篇研究论文进行处理。在他们的模型中，将每个摘要作为一个字袋模型，在字袋模型中，每个主要术语表示一个特征。尽管这种著名的分类方法是经过试验证实了的，但是预测结果并不比简单的猜想好，这表明字袋模型在这个领域中不能获得令人满意的处理结果；因此，需要利用如自然语言处理这种先进的技术进行处理。

自然语言处理（Natural Language Processing，NLP）是文本挖掘中一个重要的组成部分，也是人工智能和机器语言领域的一个部分。它将人类语言描述（如文本型文件）转变为更容易被计算机程序所能操作的形式（数字或符号格式的数据），从而“理解”人类自然语言。NLP的目的是将语法驱动文本操作的处理（通常叫做“文字统计”）变为真正的自然语言理解和处理，这种自然语言具有语法、语义和上下文的约束。

“可被理解”的深度和广度是NLP讨论的主要主题。如果人类自然语言表达不明确，而且理解时需要对所涉及的主题有一定程度的了解（不仅仅要知道这些词语、语句、段落是什么意思），那么计算机能够像人类一样准确地理解自然语言吗？可能不会！从简单的文字统计到NLP需要许多处理过程，要真正理解自然语言就需要更多的处理过程。下面是一些与实现NLP有关的一些挑战性问题：

- **部分词性标注** 由于涉及词性（如名词、动词、形容词和副词），增加了专业术语标记成文本的难度。因为词性不仅与专业术语的含义有关，而且还与所在的上下文有关系。
- **文本分割** 在一些书面语言中，如中文、日文和泰文，字与字之间没有边界。在这种句子中，文本解析过程需要对文字的边界进行界定，这通常很困难。相同的断句问题也会出现在口语分析中，因为说话的时候一系列词语连接在一起。
- **词语歧义** 许多词语不仅有一种意思，而且还需要通过查看上下文使用的词语，才能判断出哪种意思使句子更有意义。
- **语法含糊不清** 自然语言的语法含糊不清，也就是说，我们要考虑到多种句子结构。对句子的结构做出正确的判断，需要将语义和上下文信息结合起来。
- **有缺陷或不规则的输入** 说外语或方言、发音问题以及印刷或语法上的文本错误都会使语言的处理工作变得更加困难。
- **说话的方式** 句子通常代表说话人的行动，但是单独的句子结构不可能包含足够的信息用来确定行动。例如，“你能通过这门课吗？”仅需要回答“是或否”，但是“你能把盐递给我吗？”还需要肢体语言的配合。

在人工智能领域中，实现自动阅读能力的算法以及从文本中获取知识还需要很多处理过程。斯坦福大学NLP实验室的研究人员利用学习算法解析文本找到了一些方法，借助这些方法可以从文本中自动识别一些概念以及发现这些概念之间的关系。通过对这些大量文本进行独特的处理，他们的算法可以帮助他们从这些本文中自动地获得许多成百上千的知识条目，利用这些方法还可以在某种程度上提高词汇网络库。词汇网络是指一个包含词语、词语含义、近义词集合，

以及同义词语义之间关系的手工编制的复杂数据库。词汇网络是NLP应用的主要部分，但是建立和手工维护词汇网络的成本非常昂贵。因此，通过将知识自动地引入到词汇网络中，使词汇网络成为NLP一个强大的、丰富的资源。

NLP发挥显著优势并取得了很大成果的领域是客户关系管理。一般地说，CRM的目标是对顾客需求或察觉到的用户需求做出更好的理解并做出积极的回应，从而实现顾客价值最大化。NLP对CRM有重要影响的地方是情绪分析。语义分析是一个从大量文本型数据资源（网页上记录的顾客反馈信息）中发掘顾客对产品或服务所持有的满意或不满意的观点的技术。应用案例5.2就是一个在CRM领域，应用文本挖掘的成功例子。

应用案例5.2 文本挖掘技术帮助Merck更好地理解和服务于顾客需求

Merck Sharp&Dohme（MSD）是一个全球性的研究药品的德国公司，所研究的药品主要用于满足全球人类的健康需要。MSD成立于1891年，该公司通过了解、研究、生产和销售疫苗及药品，解决人类健康所面临的问题。

作为世界上最大的药品制造公司，MSD主要的投入是医生对病人提供的帮助，预期的产出是那些得了艾滋病、骨质疏松症、心脏衰弱、偏头疼、哮喘以及其他病的病人获得更好的帮助。

MSD许多年前就意识到了知识发现的重要性，并研发了一个利用数据挖掘和文本挖掘技术来丰富自己的数据和信息集合的分析型项目。MSD利用SPSS中的文本挖掘技术分析搜集到的来自于各种资源的信息，然后利用这些信息找出能够满足内科医生和病人需求的诊断程序。

挑战

和其他职业一样，在医疗保健行业的医生也有自己的看法和观点，这就是MSD所面临的挑战。要能够及时获得医生做出的诊断信息并能将这些诊断信息发送到产品研发部门，从而生产出更好的药品并为这些药品做好市场营销活动。MSD要不断地适应不同的目标用户，这件工作是非常困难的。一方面，"敢于创新"的医生有鲜明的洞察力，研究结果很快地由科学理论转变为现实。另一方面，"保守个性"的医生遵循传统的方法，任何事情都遵照书本，将大量的时间放在研究治疗方法上，这些方法来源于专家的论文或者是同事之间相互讨论的结果。MSD想要找到适合不同类型医生的方法，就需要先判断医生所属的类型。因此，MSD需要分析来自不同来源的大量的资料（包括内部数据和外部人员提供的数据），才能做出正确的判断。

解决方案

MSD利用文本挖掘和SPSS的定量分析工具来更好地理解调查得到的数据，其中的一些资料来源于各种小组交流讨论，然后向市场部提供有价值的信息。调查内容包括医生从医的年数、医生的病人数量以及一些问答资料生成的无限制的文本反应。一旦发现有用的数据，专业分析人员就会分析数据的各种意义及其之间的联系从而进行深入的研究。同时，MSD将收集到的数据用于分析。分析工具允许MSD将医生作为类型学的处理对象。利用市场部提供的指标实现了医生分类，这样MSD就能针对不同的目标用户采取不同的行动。

结果

对MSD而言，文本挖掘技术——分析非结构化的文本型数据——是必不可少的。文本挖掘建立在文本自然语法分析的基础上。它不仅仅是对关键字的搜索，还是对语法的分析和内容的"理解"。这样，企业就可以获得有用的信息，从而增强了自身的竞争力。

MSD和Gesellschaft für Konsumforschung公司的专题讨论小组（研究顾客行为）相互合作，用医生的日常工作记录分析都用到了哪些药物，包含了药物产品哪些方面的信息，以及将来处方中还会不会继续使用这些药物。通过对医生谈话和处方中体现的医药信息，进行文本挖掘分析，使得MSD优化了其产品和市场活动，并提高了市场代表的谈判能力。利用SPSS和它的文本挖掘工具，通过和医生谈话，MSD认识到药物哪方面的属性和信息容易被人理解，该术语用于市场活动中时需要进行提炼。

来源：SPSS"Munck Sharp & Dohme,""stories http://www.spss.com/success/template_view.cfm?Story_ID=185 (accessed May 15,2009).

语义分析为各种应用提供了巨大的机会。比如，在竞争力分析、市场分析和危机事件的谣言处理中发挥了巨大作用。IBM研究人员设计了一种语义分析方法，从收集到的资料中的某个专题（产品或服务）中找出涉及支持和反对两种观点。（Kanayama and Nasukawa，2006）语义分析主要问题就是识别文本中的观点是怎么表述的，这种表述所代表的是支持还是反对的态度。为了提高分析的准确性，找出观点表述和主题之间的语义关系是非常重要的。利用语法分析器和情绪词汇进行语义分析，IBM公司的系统能够从网页和新文章中分析观点，达到很高精确度（达到75%～95%，依赖于数据的多少）。

NLP已经成功地利用计算机程序自动地处理以前只能由人类处理的自然语言，并在各种领域中得到应用。下面是一些主要的应用领域：

- **信息检索** 查找相关的文献，从中找出专业知识并产生出这些内容的元数据。
- **信息提取** 信息检索的目的之一是自动地提取结构化信息，例如，从某个机器可识别的非结构化领域中得到已分类的、内容连贯的、有特定语义的、定义明确的数据。
- **命名实体识别** 与实体识别和实体提取一样，信息提取的另外一个目的是从文本中找出实体并进行分类，如姓名、组织、地址、时代表达式、数量、货币、百分比等。
- **问题解答** 自动回答用自然语言提问的问题，就是当给出一个用人类语言的形式表述的问题时，能够产生人类语言类型的答案。计算机程序从结构化数据库或者收集的自然语言文件（例如万维网中的文本语料库）中找到问题的答案。
- **自动总结** 利用计算机程序生成一个包含文献最重要观点的缩略版本。
- **自然语言生成** 系统可以将计算机数据库中的信息转化为人类可读的语句。
- **自然语言理解** 系统可以将人类语言转化为计算机程序更容易操作的更正式的表示法。
- **机器翻译** 自动地将一种人类语言翻译为另一种语言。
- **阅读外语** 计算机程序能够帮助非本语言读者用正确的发音和口音阅读不同语言部分的外语。
- **书写外语** 计算机程序能够帮助用户用外语书写。
- **语音识别** 将口语转化为机器可读的输入。根据人的发音，系统产生文本型文字。
- **文本到语音** 也叫语音合成，计算机程序自动地将自然语言文本转化为语音形式。
- **文本校对** 计算机程序对校稿或文件中的错误进行核查和改正。
- **视觉识别** 将手写稿的图片、打字机稿或打印文件（通常通过电子扫描仪获取）转化为机器可识别的文本型文件。

文本挖掘的成功和普及在很大程度上依赖于NLP的进步和对人类语言的理解。NLP实现了从非结构化文本中提取信息，这使得数据挖掘技术被用于知识（新颖有用的形式和关系）提取。简而言之，文本挖掘是NLP和数据挖掘的结合。

5.2 节复习题

1. 什么是自然语言处理？

2. NLP和文本挖掘之间有什么联系？
3. NLP有哪些好处和挑战？
4. NLP有哪些主要应用？

5.3 文本挖掘应用

随着组织机构收集的非结构化数据的增长，文本挖掘工具的价值和普及也呈上升的趋势。许多组织机构意识到利用文本挖掘工具从基于文献的数据中提取知识变得非常重要。下面是文本挖掘工具的一部分应用类型。

5.3.1 市场营销应用

文本挖掘通过分析客户服务中心的非结构化数据来提升交叉销售和追加销售的业绩。利用文本挖掘算法，从客户服务中心获得的数据及顾客的交谈记录中，提取顾客对企业产品和服务新颖的、可采取行动的信息。另外，顾客在曾经浏览的网页上对产品的评述、博客以及讨论区是企业发觉顾客意见的好方式。对大量的信息进行合理分析，可以提升顾客的满意度和终身顾客的价值（Coussement and Van den Poel，2008）。

文本挖掘为客户关系管理提供了非常宝贵的价值。企业利用文本挖掘对大量非结构化数据进行分析，结合对从组织机构数据库中的提取的结构化数据，来预测顾客的需求和购买行为。Coussement and Van den Poel（2009）利用文本挖掘预测顾客消极态度（例如，顾客数量减少），因此，企业可以认识到顾客放弃本企业产品的可能性，并明确地知道让顾客保留下来的方法。

Ghani et al.（2006）文献将文本挖掘工具应用于推断产品的显性和隐性属性，从而提高零售商分析产品数据库的能力。把商品看做是各种有价值的属性的集合，而不仅仅是原子的堆积，这能使产品在许多商业应用上更有价值，如需求预测、最优化决策、产品推荐、零售商和供应商的对比以及产品供应商的选择等应用中。系统通过利用监督学习或半监督学习技术从零售商的网站上了解产品的属性。这样企业在花费少量的人工成本的基础上，就能突出产品的属性和属性价值。

5.3.2 安全应用

文本挖掘在安全领域最重要的应用是ECHELON监视系统。正如传说中的一样，ECHELON系统能够识别的内容有电话呼叫、传真、电子邮件信息及其他类型的数据，以及通过卫星、公用电话网和微波传送拦截到的信息。

2007年，欧洲刑警组织EUROPOL开发了一个集成系统，这个系统集成了市场上最新的数据和文本挖掘技术，通过获取、存储和分析大量的结构化和非结构化数据来追踪国际组织的犯罪行为，该系统称为综合智能支持分析系统。该系统使得欧洲警察组织在国际上执行法律效力取得了很大成果（EUROPOL，2007）。

美国联邦调查局（Federal Bureau of Investigation，FBI）和美国中央情报局（Central Intelligence Agency，CIA），在国家安全部门的帮助下，联合开发了一个超级计算机数据和文本挖掘系统。该系统主要作用是创建一个大型的数据仓库，该数据库包含联邦政府、国家和地方法律部门知识挖掘所需要的各种数据和文本挖掘模型。在此之前，美国联邦调查局和美国中央情报局有各自独立的数据库，这些数据库中之间有少量甚至没有关联。

文本挖掘在安全领域中的另外一个应用是欺诈行为侦察。Fuller et al.（2008）研发了一种鉴别出欺诈行为的模型，对大量真实世界犯罪（嫌疑犯）相关的资料进行文本挖掘处理。该模型利用从文本中提取的大量线索，对抽样进行预测，准确率达到70%。由于线索仅仅是从文本型资料（没有语言和视觉上的资料）中提取的，所以能达到这种准确率意义非同寻常。而且，和

其他欺诈行为侦察技术相比，例如测谎器，这种方法能够做到不打草惊蛇，并能广泛应用于文本型数据和录制的音频中。应用案例5.3更详细地介绍了基于文本的欺诈行为侦察。

应用案例5.3 谎言挖掘

随着网络信息的增长和全球化的趋势，计算机通信慢慢渗入到人们的日常生活中，欺骗犯罪行为也出现了新的形式。聊天信息、即时信息、文本信息和在线社区活动信息迅速增长。甚至连电子邮件的使用也越来越频繁。随着文本交流信息的大幅度增加，人们通过计算机通信进行欺诈的行为也越来越多，并造成严重的损失。

令人遗憾的是，人们对欺骗行为的检测结果不是很好，文本通信使得检测欺骗变得更加困难。大量的欺骗检测（如置信度评定）研究包含面对面的交流和访问。因此，随着基于文本通信的增加，基于文本的犯罪行为侦察技术是必要的。

成功侦查欺骗行为（也就是谎言）的技术已经得到普遍应用。法律案件中利用决策支持工具和技术进行犯罪调查，机场安检，监控恐怖分子嫌疑犯通信信息。人力资源部门可以利用欺骗侦查技术对职位申请人进行调查。公司办公人员用这些工具和技术检查电子邮件信息，发现欺诈或不正当的行为。尽管有些人认为自己可以识别出那些不可信的人，但是欺骗行为研究结果表明：平均54%的人能够做出准确的判断（Bond and Depaulo，2006）。如果涉及从文本信息中查出欺骗行为，那么这个数据可能更低。

Fuller et al.（2008）将文本挖掘和数据挖掘技术联合起来，对军事犯罪中嫌疑犯个人陈述资料进行了分析。这些资料来自于嫌疑人和目击者用自己的语言写下的对事件的回忆。军事执法人员通过核对相关信息，判断这些陈述是真实的还是虚假的，这些判断建立在有效的证据和决议的基础上。一旦材料被证实为真或假，执法人员就会将这些判定结果和陈述资料交给研究小组。最终，共有371个陈述文件用于分析。以上Fuller et al.（2008）文献采用的基于文本的欺骗检测方法是利用了一种称为“信息特征挖掘”的处理过程，该过程依靠数据元素和文本挖掘技术。图5-2对这个过程进行了简单描述。

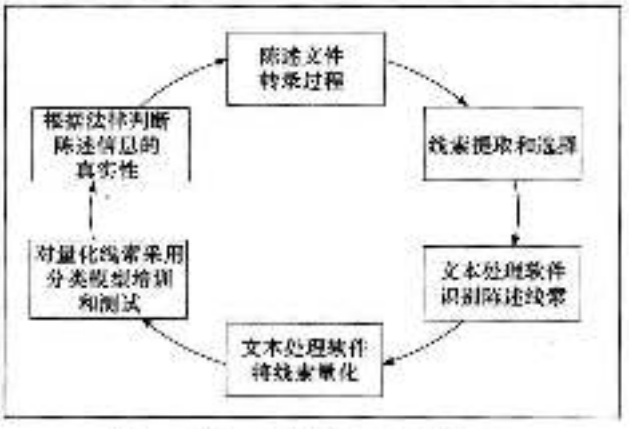


图5-2 基于文本的欺诈行为侦察过程

来源：C. M. Fuller, D. Biros and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, January 2008, Big Island, HI, IEEE Press, pp. 80–99.

首先，调查人员准备数据。原始的手写陈述文件要转录到Word文档中。然后，识别特征（也就是线索）。调查人员辨别出31类重要特点，这些重要特点的类型和语言种类与原文件相互独立，并且可以用自动化的方法进行分析。如，第一个人的发音可以通过所说的话，如"I or me"进行判断，而不需要对文本内容进行分析。表5-1列出了这个研究中的重要信息的种类并进行举例说明。

表5-1 欺骗检测语言特征中用到的种类和示例

序号	类别	示例
1	数量	动词个数、名词短语个数等
2	复杂度	从句的平均个数、句子的平均长度等
3	不确定性	情态词、情态动词等
4	间接性	消极的声音、客观因素等
5	表达方式	感性
6	多样性	词汇的多样性、冗余等
7	非正式性	印刷错误比例
8	专业性	时空信息、直觉信息等
9	倾向	积极倾向、消极倾向等

来源：Based on C. M. Fuller, D. Biros, and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, 2008, Big Island, HI, IEEE Press, pp. 80–99; C. F. Bond and B. M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Reports*, Vol. 10, No. 3, 2006, pp. 214–234.

将陈述文件中提取到的特征放到一个简单的文件中，以便以后使用。用十折交叉确认法和其他几种特征选择方法，研究人员比较了3种流行数据挖掘方法的预测精确度。结果表明神经网络模型在测试样本数据时精确度最高，达到73.46%；决策树次之，精确度为71.60%；回归分析方法精确度最差，为65.28%。

以上结果表明利用文本信息对欺骗行为进行自动化的侦查能够帮助调查人员从文本信息中找出欺骗信息，并能应用到真实世界数据中。尽管这些技术仅适用于文本型线索，但是它们的精确度高于绝大多数其他的欺骗行为侦查技术。

5.3.3 生物医学应用

文本挖掘在医药和生物医学领域有重要应用，主要由于以下几个原因。首先，该领域的文献和出版物（特别是开放性资源期刊的出现）在数量上以指数的形式上涨。其次，和其他领域相比，医药领域的文献更加规范、标准，使其成为一个"可挖掘的"信息资源。最后，该领域专业术语相对固定且格式标准。下面是文本挖掘技术在生物医学文献中提取新模式的典型应用。

实验技术，如DNA序列分析、基因表达系列分析（Serial Analysis of Gene Expression，SAGE）和蛋白质质谱分析，生成大量和基因、蛋白质相关的数据。和其他实验科方法一样，需要对这些研究中的大量生物体数据进行分析。该领域的相关文献可以为实验验证和实验解释提供大量的信息资料。因此，将自动化的文本挖掘工具应用于实验解释是目前生物学研究所面临的一个主要挑战。

知道蛋白质在细胞内的分布有助于分析它在生物学和在药物中的作用。文献中介绍了许多蛋白质分布预测系统，有些文献针对某些特定的细胞进行研究，有些文献则涉及大量的细胞。Shatkay et al.（2007）提出了一个基于多种序列分析和文本特征的蛋白质存储单元综合预测系统。该系统的创新之处在于它选择文本资料的方法，以及用序列分析法将这些资料进行集成分析。他们分别用以前的数据集和新的数据集来测试系统的预测能力，结果表明他们的系统总是优于以前的判断结果。

Chun et al.（2006）利用美国国立医学图书馆（MEDLINE）的文献，设计了一个发现疾病和基因之间关系的系统。他们记录了6个公共数据库中的疾病和基因并组成了一个字典，通过字典

匹配发现其中的关联。由于字典匹配产生大量错误的信息，所以他们利用基于机器学习的命名实体识别技术（Named Entity Recognition，NER），过滤出错误的疾病-基因命名。他们发现能否发现疾病和基因之间的关系主要取决于命名实体识别过滤器的性能，在使用在过滤器的基础上精确度提高了26.7%，并减少了没用的信息。

图5-3简单描述了文献（Nakov et al.，2005）中，从生物医学文献中发现基因-蛋白质关系（或蛋白质-蛋白质关系）的多级文本分析过程。该示例对生物医学文本中一个简单句子中最主要部分（第三层底部）的词性标记出来，并进行浅层句法分析。用层次表示法对标记的生物学内容（词语）进行分析，从而得到基因-蛋白质之间的关系。将这种方法（或者这种方法的改进）应用到生物医学文献中，对人类基因组计划中的复杂解码发挥了重大的作用。

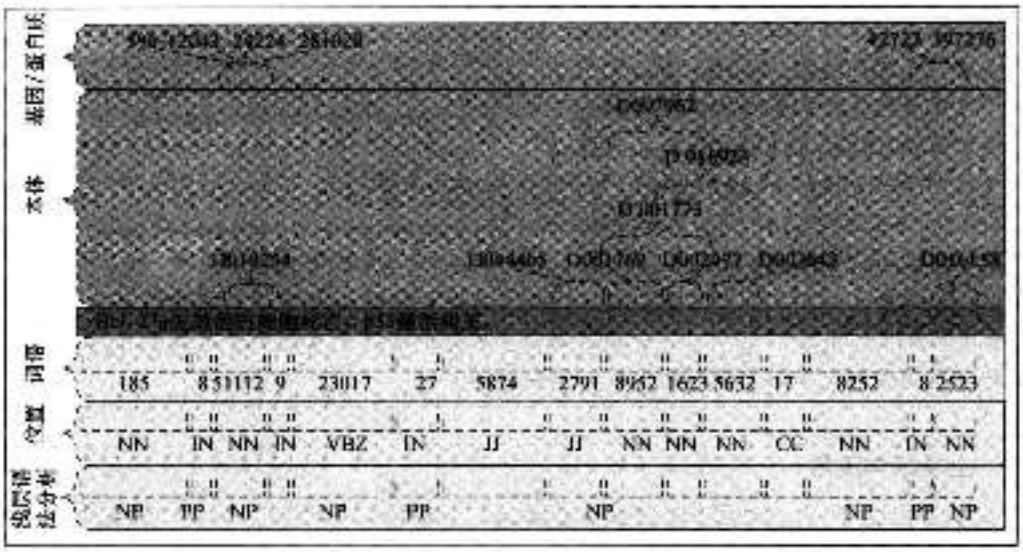


图5-3 多级文本分析过程的基因-蛋白质之间关系

来源：P. Nakov, A. Schwartz, B. Wolf, and M. A. Hearst, "Supporting Annotation Layers for Natural Language Processing," *Proceedings of the Association for Computational Linguistics (ACL)*, interactive poster and demonstration sessions, 2005, Ann Arbor, MI, pp. 65–68.

5.3.4 学术应用

利用文本挖掘对信息制定索引，便于信息检索的需要，对于拥有大量数据库的出版商具有重要作用。特别是科研领域，因为科研领域的文献中存储着大量的专业信息。有些措施已经开始实施，如《Natural》期刊的文本挖掘接口和美国国立卫生研究院的公共期刊的文献类型定义（Document Type Definition，DTD），DTD是指给机器提供语法提示，回答对文献中所包含内容进行的特殊查询，而无需消除出版商对公众访问的限制。

学术机构也开始了文本挖掘行动，如曼彻斯特大学和利物浦大学合作建立的国家文本挖掘中心，对学术领域提供了定制化工具、研究设备及对学术团体文本挖掘方面的建议。文本挖掘首先应用于生物和生物医学领域，现在已经扩展到社会科学领域。美国加利福尼亚大学伯克利分校的信息学院正在研发一个叫做生物文本的项目，在生物科学研究中引入利用文本挖掘和分析工具。

本节介绍了文本挖掘在不同领域的广泛应用。应用案例5.4将介绍利用文本挖掘技术分析航

空业空气腐蚀，从而提高安全性的应用。

应用案例5.4 穿过文本飞行

文本挖掘被认为是从数字化格式存储的手写资料中提取出有用的信息一种重要工具。分析家利用文本挖掘软件运用模式识别方法研究某些重要领域。例如，航空公司用文本挖掘技术研究事故报告信息，提高了行业的知识水平。通过文本挖掘工具可以及时地研究客观的、组织的和行为方面的问题。

航空公司对企业的运营进行彻底、系统的分析。一旦发生的事情能够引发事故的话就会形成一个事故报告文件。文本挖掘技术可以从大量的事故报告文件中识别出重要的问题。人对航空公司的大型数据库的理解具有局限性，计算机专业术语和日常术语也是不同的。

爱尔兰航空公司（aerlingus.com）通过对1998年1月到2003年12月的事故报告文件进行分析，发现了潜在的模型并找出模型之间的关系。爱尔兰航空公司利用Megaputer公司（megaputer.com）的综合数据和文本挖掘工具PolyAnalyst，目标是开发出一个能够通过分析事故的类型、地点、时间及其他信息来发现模型和模型之间关系的程序，该程序能够满足研究人员的日常需要。

PolyAnalyst找出事故报告文件中最常用的术语，并形成一个术语词典，虽然这个词典不包含全部的术语，但它可以作为文本分析一个有价值的开端。PolyAnalyst也可以从数据中提取出关键术语（或者这些术语的同义词）。这样就形成一个常见术语报表（或语义表），这个报表包含了术语项及其使用的频率，其目的是识别有意义的术语聚类。一个叙述性概要文件包含了一系列关键词组，根据这些关键词将文件分成不同意义的组。如，泄漏这个关键词可能和食物、燃料、化学药品和洗手间这4个关键词有联系。从关键词角度来看，在语义上食物和咖啡、茶或饮料有联系。因此，食物作为一个类别，将和泄漏相关的不同的产品报告归为食物类别。

将文本挖掘技术应用于航空公司事故报告中，可以找出影响安全性能的因素。在大量事故报告文件中，应用文本挖掘技术可以验证指定的理论并获取一般常识性知识，同时还能获得新的模式并将其添加到知识库中。

来源：J. Fruehlich, S. Ananyan, and D. L. Olson, "Business Intelligence Through Text Mining," *Business Intelligence Journal*, Vol. 10, NO. 1, 2005, pp. 43 – 50.

5.3节复习题

1. 列出并简要说明文本挖掘的应用。
2. 在安全领域和反恐方面，如何应用文本挖掘技术？
3. 在生物医学领域文本挖掘技术有哪些应用前景？

5.4 文本挖掘过程

文本挖掘的研究需要依靠基于最佳实践的坚实方法论才能取得成功。一个标准的处理模型是跨行业数据挖掘过程标准CRISP-DM，也就是数据挖掘的行业标准。虽然跨行业数据挖掘过程标准基本上可以应用于文本挖掘项目中，但是还需要包括复杂数据预处理活动的特殊的文本挖

掘处理模型。图5-4描述了（Delen and Crossland，2008）文献提出的一个典型的文本挖掘过程的顶层语境图。这个语境图展示了处理的范围，分析了与外界大环境的接口。实际上，处理过程的边界可以明确地说明文本挖掘过程中所包含（或不包含）的内容。

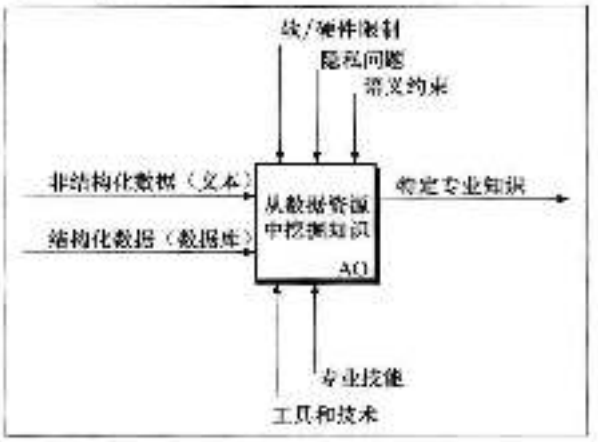


图5-4 文本挖掘过程的语境图

从图中可以看出，这个基于文本的知识发现过程的输入（和矩形框左边界内部连接的部分）包括了可用于该过程的非结构化和结构化数据的收集、存储和处理。输出（矩形框右边界的外部延伸部分）的是用于决策支持的特定知识。控制，也叫约束（和矩形框顶部连接的内部）包含了软硬件的要求、隐私问题以及与该文本处理过程相关的自然语言的表现格式。机制（和矩形框底部相连的内部）包括适当的技术、软件工具和专业技能。文本挖掘（包含于知识发现上下文中）的主要目的是对非结构化（文本型）数据（也包括和相关问题有关联的结构化数据）进行处理，挖掘出有意义的可用的内容，从而有利于决策的制定。

在最高级别中，文本挖掘的处理可以划分为3个连续的任务，每一项任务都要求特定的输入并产生固定的输出（如图5-5）。如果由于某种原因，其中的一项输出不是用户所希望的，就要返回到上一个步骤的执行中。

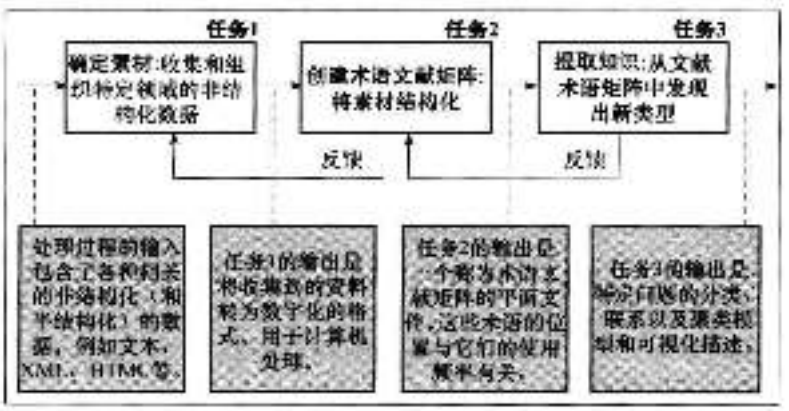


图5-5 文本挖掘处理过程的3步骤

5.4.1 任务1：确定素材

第一项任务的主要目的是收集和所要研究的内容（感兴趣的领域）相关的各种文献，包含文本文件、XML文件、电子邮件、网页和短摘要。除了以上的文本型数据外，音频数据可以通过语言识别程序转化，成为文本数据。

收集好文本数据之后，要将这些文本数据转化并重新组织为相同的格式（如美国信息交换标准代码ASSCII文本文件的格式），便于计算机处理。文件的组织形式可以看做是存储在文件夹中的数字化摘录，或者是某个特定领域的网页链接。许多商用的文本挖掘工具可以将这种重组后的文件作为输入，然后转换为一个平面文件以便处理。或者，该平面文件可以不由文本挖掘软件生成，将它作为文本挖掘应用的输入。

5.4.2 任务2：创建文献术语矩阵

该任务是利用数字化的组织好的文献（语料库）形成文献术语矩阵（Term-Document Matrix，TDM）。在TDM矩阵中，行代表文献，列代表术语。术语和文献之间具有标记与被标记（如可以用术语在文献中出现的次数来表示）的关系。图5-6是文献术语矩阵的一个典型示例。

任务2的目标是将一系列组织好的文献（语料库）转化为TDM矩阵，矩阵单元格中是合适的指数。假设前提是文献可以用其中的一些术语及其出现的次数来表示。所有的术语在表现文献内容上都同样重要吗？很显然，不是。有些术语，如冠词、辅助动词以及语料库中几乎所有的文献都用到的术语，没有重要的影响。因此，在索引过程中应该把这些词排除在外。这些术语一般叫做无用词，而且这些词在某些特定的研究领域是专业的，由该领域的专家来识别。另一方面，选择预先定义为文献索引的词作为术语（这些术语叫做包含词或词典）。而且同义词（被看做相同的词）和专业短语（如埃菲尔铁塔）应当包含到这些术语中，这样索引项就更加精确。

文献 \ 术语	投资风险	项目管理	软件工程	并发
文档1	1			1
文档2		1		
文档3			3	
文档4		1		
文档5			2	1
文档6	1			1

图5-6 一个简单的文献术语矩阵

创建索引的另外一种过滤方法是词干提取，将文献中的词汇进行缩减，从而得到词干，提取词干的时候，不同语法格式偏差的动词被认为是相同的词。如，提取词干的时候modeling和modeled都看做是model。

第一代TDM矩阵包含了语料库中除无用词之外所有术语（作为列），所有文献（作为行）以及每个术语在文献中出现的次数（单元格的值）。通常情况下，语料库中包含了大量的文献，因此TDM矩阵包含大量的术语。处理如此大的矩阵需要耗费大量的时间，更重要的是有可能得不到正确的模式。因此，必须考虑以下两个问题：（1）找出最有代表性的指数；（2）怎样将矩阵的大小减少到合适的规模？

代表性的指数 一旦输入的文献被检索并计算出其出现的次数，那么就要做一些额外的转换工作对获取的信息进行总结和聚集。原词出现的次数反映了该词语在一个文献中的重要性程度。尤其是出现次数多的词语更能代表文献的内容。然而，假设词语出现的次数和它在文献中的重要性成比例是不合理的。如，一个词语在文献A中出现了一次，在文献B中出现了3次，不能表明该词语在文献B中的重要性是文献A中重要性的3倍。为了找出一个前后一致的TDM矩阵做一进步的分析，需要将最初指数标准化。不直接用词语的出现次数，而是用一些方法将术语与文献之间的数值表示进行标准化处理。下面是一些常用的标准化处理方法（StatSoft，2009）：

- **对数频率法** 对词语出现的次数取对数，这种转换会减少词语出现的次数，并减小词语的出现次数对后期分析结果的影响。

$$f(wf) = 1 + \log(wf) \quad (wf > 0)$$

式中，wf指原词（或术语）出现的次数，$f(wf)$是取对数的结果。这种方法适用于任何词语出现次数大于0的术语文献矩阵。

- **二进制频率法** 只要一次简单的转换，这种方法就可以确定一个词语是否在一篇文献中出现过。

$$f(wf) = 1 \quad (wf > 0)$$

这个计算结果用1和0来确定文献中出现和未出现的各个词语，而且这种方法能够减少

原词出现频率对后期计算和分析的影响。

- **逆文献频率法** 包含同一个术语出现在不同文献的频率（记为 df）是需要仔细考虑，而且这对更深入的分析有重要影响。如，词语猜想可能在所有的上下文中频繁出现，而软件这个词在少数的几个文献中出现。因为，不管什么主题的文献中都有可能做出各种"猜想"，而软件这个词只可能出现在和计算机软件相关的文献中。一个既考虑了词语的专业性（文献出现的次数），又考虑了词语出现频率（某个术语出现的次数）的方法是逆文献频率法，可以用以下公式表示第 i 个词语和第 j 篇文献的转换（Manning and Schutze，2009）：

$$idf(i,j) = \begin{cases} 0, & \text{如果 } wf_{ij} = 0 \\ (1 + \log(wf_{ij}))\log\dfrac{N}{df_i}, & \text{如果 } wf_{ij} \geq 1 \end{cases}$$

该式中，N 表示所有文献的数目，df_i 表示包含第 i 个词语的文献的个数。因此，该公式可以通过取对数的方式减少简单词语出现的次数，并且如果每个文献中都包含同一个词语，那么该词语的权重为0（也就是 $\log(N/N)=0$），如果一个词只出现在一个文献中，则为最大权重（因为 $\log(N/1)=\log(N)$）。很容易看清这种转换方法是如何通过词语出现的相对频率它们在文献分析中的语义演变这两个方面来形成指数的。这是最普遍使用的一种方法。

减小矩阵的规模 因为文献术语矩阵规模通常很大，而且词语比较分散（大部分词语出现的次数为0），所以"怎样将矩阵的规模减小到合适的大小"成为一个重要的问题。下面是处理矩阵规模的几种方法：

- 专家分析所有的术语，淘汰对研究内容意义不大的词语（这是一个手工且劳动强度比较大的处理过程）。
- 淘汰只在少量文献中出现而且出现次数很少的词语。
- 用奇异值分解（Singular Value Decomposition，SVD）方法对矩阵进行处理。

奇异值分解类似于主成分分析，将矩阵输入（由提取的术语个数决定的文献个数）的规模减小到较小的维度空间中，每个连续的维度代表变量的最大可能范围（术语和文献之间）（Manning and Schutze，2009）。理想情况下，研究人员确定两三个包含了绝大多词语和文献变量（差别）的主要维度。一旦确定了矩阵的维度，就可以得出文献所包含（讨论、描述）的隐含主题。特别地，矩阵 A 是一个 $m \times n$ 矩阵，m 指输入的文献个数，n 指分析中用到的术语的个数。奇异值分解计算出正交矩阵 $U=m \times r$、正交矩阵 $V=n \times r$ 和正交矩阵 $D=r \times r$，因此 $A=UDV'$，r 是指矩阵 $A'A$ 的特征值。

5.4.3 任务3：提取知识

利用准备好的术语文献矩阵和其他潜在的一些结构化的数据，就可以从大量的专业内容中挖掘出新的模式。主要的知识提取方法包括：分类、聚类、关联和趋势分析。下面对这几种方法进行简单介绍。

分类 对复杂的数据进行知识发现的最常见的方法大概可以说是对某一对象的分类。分类是指将给出的数据实体分为不同的类别（或种类）。在文本挖掘中叫做文本分类，也就是利用包含文献和文献分类的数据集所产生的模型，从已有的分类（对象、主题或概念）和许多文本文献中找出正确的主题（对象或概念）。如今，自动化的文本分类已经应用到各个环境中，包括自动化或半自动化（交互的）创建文本索引、过滤垃圾邮件、根据分级目录进行网页分类、自动生成元数据和基因检测等。

文本分类的两个主要方法是知识工程和机器学习（Feldman and Sanger，2007）。知识工程方法是将有关分类的专家知识，以描述的形式或程序分类规则的方式编码到系统中。机器学习方法是指通过学习一系列的重新分类示例，利用一个普遍适用的归纳程序创建类别。随着文献的数量呈指数的增加，以及专家很难掌握所有的知识，所以这两种方法中机器学习法更加流行。

聚类 聚类是指将对象归类到“自然的”集合的一个无监督的数据处理方法。和分类相比，聚类利用预归类训练示例的数值特征创建一个新的未标签组，聚类将未标记的对象（例如，文献、顾客评语、网页）分为不包含以前知识的有意义的集群。

聚类的应用很广泛，从文献检索到实现网页信息搜索都应用到聚类。事实上，聚类的一个典型应用是对大量的文本（如网页）进行分析和导航。假设相关的文献比不相关的文献更具有相似性，如果这个假设成立，那么基于内容相似性的文献聚类就可以提高文献搜索的有效性（Feldman and Sanger，2007）：

- **改进搜索查全率** 由于聚类是基于相似性而不是文献的单个术语，所以当某个查询和一个文献相匹配时，就会返回整个分组，因此聚类可以改进搜索查全率。
- **提高搜索的准确性** 聚类可以提高搜索的准确性。由于文献的数量正不断增长，所以从大量的文献中找出和搜索相匹配的文献非常困难。聚类可以将相关文献分为数量上更小的组，并根据文献的相关性进行排序，仅返回和搜索内容最相关的组。

最常见的两种聚类方法是分散/收集聚类和查询特定聚类：

- **分散/收集聚类** 对于不能得到一个明确结果的搜索，这种聚类方法能够提高人们查阅文献的效率。该方法动态地生成一个文献内容摘要列表，而且这个列表随着人们查阅的内容不同而自动地进行调整和更改。
- **查询特定聚类** 这是一种分层次的聚类方法，确定文献的相关性程度，与搜索内容相关程度最大的文献放在一个小的集群中，而相关程度小的文献放在一个大的集群中。这种方法在大规模的文献收集方面一直表现很好。

关联 在第4章中详细介绍了关联。产生关联规则（或解决购物篮问题）的主要思想是找出频繁出现的组合。

在文本挖掘中，关联特指概念（术语）或概念集合之间的直接关系。关联规则 $A \Rightarrow C$，包含了两个概念集 A 和 C，这个关联规则可以用支持度和置信度这两种方法来量化。置信度是指包含 C 中所有概念的文献占同一文献集中包含 A 中所有概念的文献的百分比，支持度是指包含 A 和 C 两个概念集中所有概念的文献百分比（或数量）。如一个文献库中，概念“软件实施错误”在与“企业资源规划”和“顾客关系管理”概念相关的文献中频繁出现，支持度是4%，置信度是55%，这说明有4%的文献同时包含这3个概念，而且包含“软件实施错误”的文献中有55%的文献也包含“企业资源规划”和“顾客关系管理”这两个概念。

关联规则的数据挖掘用来分析网络文学（发表在网络上的新闻和学术性文章）以便跟踪禽流感的爆发和进展情况（Mahgoub et al. 2008）。其思路就是要从地理区域、硬币的散布和对策（处理）中自动识别其中的关联关系。

趋势分析 文本挖掘中的趋势分析基于这样的思路，各种类型的概念分布是文献汇集的功能，也就是说对同一个概念集，不同的文献汇集导致不同的概念分布。因此，两种分布除了来自于不同文献的之外其他条件都相同，可以对这两种分类进行比较。这种分析方法的显著特点是两个不同的文献集都来自于同一个资源（如都来自于同一个学术期刊），但是文献的时间不同。Delen and Crossland（2008）在大量的学术文章（出版量最大的3个学术期刊）中应用趋势分析，研究一些重要的概念在信息系统领域的演变。

综上所述，文本挖掘包含各种方法。应用案例5.5介绍应用不同的技术对大量文献进行分析。

应用案例5.5 文本挖掘在文献调查研究中的应用

研究人员在搜索和评审相关文献时，面临的任务量和复杂性越来越大。要想对相关知识进行拓展，在对文献中承载的信息进行收集、组织、分析和吸收方面加大投入是非常重要的，特别是研究人员本身的学科更是如此。随着相关领域以及之前被认为不相关领域的重大研究工作的深入，研究人员要想进行详尽的研究，工作量大得惊人。

在新领域，研究人员的工作更加乏味和复杂。从其他文献中查找出相关知识，特别是需要对大批量出版的文献进行手工查找是非常困难甚至是不可能的。即使利用大量的研究生或同事来查阅所有的相关出版物也是很困难的。

每年都会举行许多学术会议，作为会议当前所关注的知识补充，组织者还举办小型探索和研讨会。许多情况下，这些增加的项目主要是向参会者介绍相关领域的重大研究方向，以及研究感兴趣和关注的“下一个热点”。确定下一个小型探索和研讨会的主题通常是主观决定，而不是依据现有的研究。

最近的研究中，Delen and Crossland（2008）通过利用文本挖掘工具实现了对大量的文献进行半自动化分析，从而在很大程度上提高了研究人员的效率。利用数字图书馆和在线搜索引擎，用户可以从管理信息系统领域的3个主要期刊中下载到所有有用的文献，这3个期刊包括：《MIS Quarterly》（MISQ）、《Information System Research》（ISR）和《Journal of Management Information System》（JMIS）。为了使3个期刊保持相同的时间间隔（进行纵向比较研究），将期刊的数字出版物发表时间作为研究的开始时间（如，《Journal of Management Information System》文献从1994年开始有数据的格式）。找出每个文献的标题、内容摘要、作者、关键字、卷、出版号和出版年，然后将文献下载到一个简单的数据库中。同时，数据集中还包含了每个文献的期刊类别，以使用于差异分析。这些文献中不包含编辑附注、研究笔记和文献管理状况。表5-2以表格的形式概括了文献中包含的信息。

表5-2 数据集中的项目列表

期刊	年份	作者	标题	卷/期	页	关键词	摘要
MISQ	2005	A. Malhotra, S. Gosain 和 O. A. El Sawy	供应链环境吸收能力：资产负债的合作伙伴使加识创造市场	29/1	145－187	知识管理、供应链、吸收能力、组织间信息系统、架构方法	频繁的价值创新使得供应链从交易处理转变为合作伙伴的合作平台
ISR	1999	D. Robey 和 W. C. Boudreau	信息技术在组织的多重应用：理论指导和方法应用		165－185	组织变革、技术影响、组织理论、研究方法论、组织凝聚力、电子交流、管理信息系统实施、文化系统	尽管当前认为先进的技术在组织变革中起着决定性的作用，但是经验研究发现一些不一致的决定性因素，本文分析矛盾的……
JMIS	2001	R. Ann 和 E. K. Clemon	信息产品质量投资和信息产品自我保护投资的优化平衡		65－88	信息产品、互联网广告、产品定位、信号传输、信号博弈	当产品（服务）的生产商不能够满足用户的需求时，他们就要考虑广告的作用……

在分析阶段，Delen 和 Crossland 用文献的内容摘要作为提取信息的来源。他们之所以没有选主要基于以下两个原因：（1）正常情况下，文献的内容摘要中包含了关键字，如果利用

关键字中包含的词进行搜索就造成了重复，这样做没有意义；(2) 作者在关键词中使用的是和研究内容相关的词语（并不一定是文献内容所包含的词语），所以用关键词可能对文献的内容分析产生偏差。

首先，对这3个期刊进行纵向研究（如，随时间的变化研究内容的变化）。为了进行纵向研究，他们将这3个期刊12年（从1994年到2005年）的文献分为4个部分，每个部分包含了3年的期刊。这样，12个相互独立的数据集形成了12个实验组。对这12个数据集进行文本挖掘，从每个数据集中文献的内容摘要中提取出该文献中最具有代表性的术语。将3个期刊上的术语按时间顺序列成表。

其次，利用所有的数据集（包含3个期刊的4个时间段的文献）进行聚类分析。聚类可能是文本挖掘技术中最常用的分析方法。该研究用聚类的方法将文献进行自然分组（将它们分到不同的组中），然后找出能代表每个分组的术语。他们用奇异值分解来减小术语文献矩阵的大小，然后用最大期望值算法创建分组，并通过实验来确定最佳分组数。最后决定将文献分成9个组，然后从以下两个方面对这些分组进行分析：(1) 期刊类型（见图5-7）；(2) 时间。目的是找出这3个期刊之间的区别和共同点以及这些组之间的区别；也就是说，回答了“这些分组是否能够体现出每个期刊所代表的研究主题的不同？”以及“这些分组之间是否呈现了随时间而变化的特征？”。它们利用表格和图解的表现形式发现和分析了多个有趣的模型（更多信息请查看 Delen and Crossland，2008）。

来源：D. Delen and M. Crossland, "Seeding the Survey and Analysis of Research Literature with Text Mining," *Expert Systems with Applications*, Vol. 34, NO. 3, 2008, pp. 1707 – 1720.

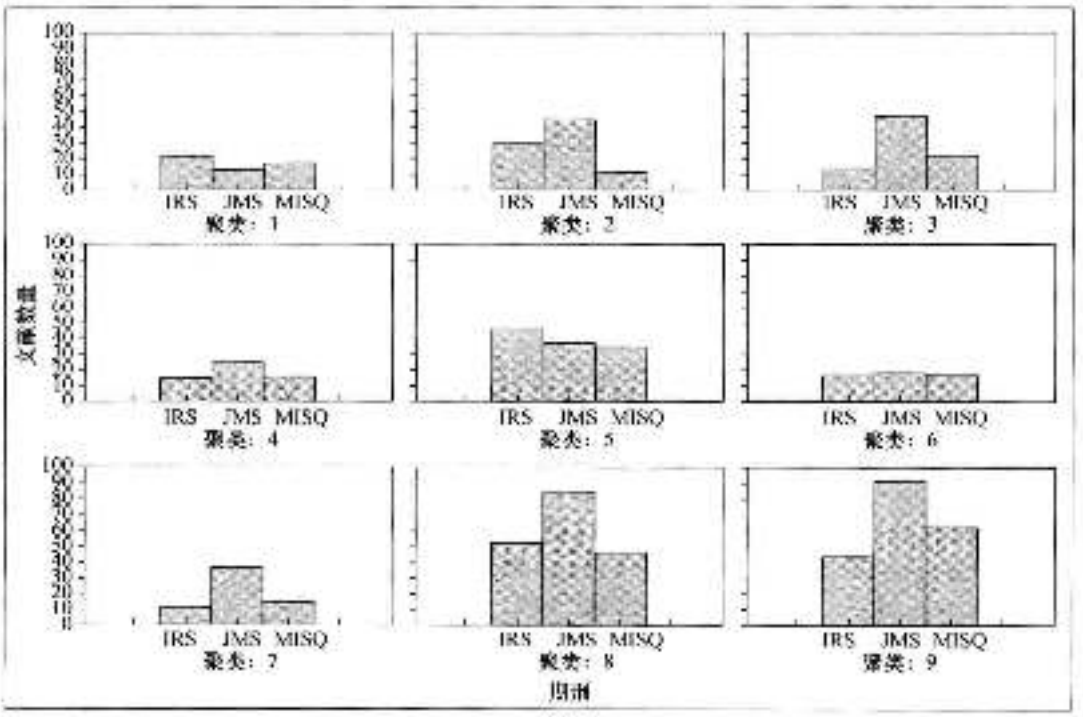


图5-7 用9个聚类对6个期刊文献的数量进行分析

来源：D. Delen and M. Crossland, "Seeding the Survey and Analysis of Research Literature with Text Minging," *Expert Systems with Applications* Vol. 34, No. 3, 2008, pp. 1707 – 1720.

5.4 节复习题

1. 文本挖掘的主要步骤有哪些?
2. 将词语的出现率规范化的原因是什么?将词语出现率规范化的方法有哪些?
3. 什么是单一价值分解?在文本挖掘是如何应用的?
4. 从资料中提取知识的主要方法有哪些?

5.5 文本挖掘工具

越来越多的组织开始注意到文本挖掘的价值,软件厂商和一些免费的文本挖掘工具呈现增长的趋势。下面是按照商业软件工具和免费软件工具进行分类的一些流行的文本挖掘工具。

5.5.1 商业软件工具

下面是一些著名的文本挖掘工具,许多公司在它们的网站上提供了产品的演示版本。

1. ClearForest 提供了文本分析和可视化工具(clearforest.com)。
2. IBM 智能数据挖掘套件,包括数据和文本挖掘工具,现如今已经完全集成到 IBM 的 InfoSphere 数据仓库中(ibm.com)。
3. Megaputer 的文本分析提供了自由格式的文本语义分析、概括、聚类、导航和具有重新定位动态搜索的自然语言检索(Megaputer.com)。
4. SAS 的文本挖掘提供了丰富的文本处理和分析工具(sas.com)。
5. SPSS 的文本挖掘工具从呼叫中心记录、博客、邮件和其他非结构化数据中提取关键概念、观点及相互之间的关系,并将它们转化为结构化的格式用于预测模型的建模(SPSS.com)。
6. Statistica 的文本挖掘提供了包含可视化功能的文本挖掘工具(statsoft.com)。
7. VantagePoint 提供了各种交互图形视图和分析工具,具有强大的从文本数据仓库空中发现知识的功能(vpvp.com)。
8. Provalis Research 的 WordStart 分析模型实现对文本信息的分析,如开放式问题的回复和采访信息(provalisresearch.com)。

5.5.2 免费软件工具

一些免费软件工具是开源的,适用于非营利组织:

1. GATE 是一个开源文本挖掘工具包,具有免费的开源框架(或 SDK)及图形处理环境(gate.ac.uk)。
2. RapidMiner 的社区版包括文本挖掘模块(rapid-i.com)。
3. LingPipe 是一个 Java 编译的套件,对人类语言进行语义分析(alias-i.com)。
4. S-EM(Spy-EM)是一个文本分类系统,该系统通过一些确定的未标记的示例进行学习(cs.uic.edu/~liub/S-EM-download.html)。
5. Vivisimo/Clusty 是一个网络搜索和文本聚类引擎(clusty.com)。

5.5 节复习题

1. 最流行的文本挖掘工具有哪些?
2. 为什么一些流行的文本挖掘工具都是由统计公司提供的?
3. 选择一个免费的文本挖掘工具和选择一个商业工具的利弊各有哪些?

5.6 Web挖掘概述

万维网服务（或即时网络）作为一个十分庞大的数据和信息库，几乎存储了人们可以想象出的所有东西。Web可能是世界上最大的数据和文本库，它里面存储的数据量每天都在快速地增长着。在网上人们可以发现许多有趣的信息：谁的主页链接到其他主页了，多少人链接到了某一特殊的网站页面，以及一些独特的网站是怎样建立的。另外，网站的每一次访问，搜索引擎的每一次搜索，任何链接的每一次点击，电子商务网站上的每一笔业务处理等，都会创造出更多的数据。虽然用HTML或XML编写的，以网页源码形式存在的非结构化文本数据是网站的主要内容，但网络上也包含了许多超链接信息（链接到其他网页）和使用信息（访问者与网站间的互动记录），这些信息都为知识发现提供了丰富的数据。对于这些信息的分析可以使我们更好地利用网站，并且能够帮助我们增强与网站访问者间的关系，提升网站的价值。

然而，Han和Kamber在2006年提出，在高效、有效的知识挖掘方面，网络也面临着十分严峻的挑战：

- **对于高效的数据挖掘来说网络还是过于庞大** 网络是如此的庞大、成长的如此迅速，甚至确定网络的大小都十分的困难。由于网络规模的急速增长，建立一个数据仓库用来复制、存储、整合网络中所有的数据是不可行的，这使得数据收集和整合成为了一个极大的挑战。
- **网络太复杂** 一个网页的复杂性远远大于传统的一页文本记录。网页缺乏统一的结构。网页包含了比任何书籍、文章或者其他传统的基于文本的文献都丰富很多的自主风格创意和内容。
- **网络动态性过强** 网络是一个动态性很高的信息源，不仅网页增加的速度快，而且网页内容更新的也很频繁。网页上的博客、新闻、股市结果、天气预报、体育信息、价格、企业广告和大量的其他信息都在不停地更新。
- **网络涉及领域广** 网络服务于不同的领域，连接上亿个工作站。网络用户有不同的背景、兴趣和应用目的。多数用户对信息网络结构没有清楚地认识，当他们想搜索到自己想要的信息时，可能会遇到许多困难。
- **网络包容万象** 对某些用户（或某些应用）来说，网页上的信息只有小部分是和自己相关的或有用的。据说对99%的用户来说，有99%的信息是没用的，我们有可能不能很明显地感觉出这种现象，但是一般用户只会对网页上的小部分信息感兴趣，而且其他的信息会对用户产生困扰。找出和某些用户或某些应用相关的内容是网络搜索中一个重要的问题。

以上问题促进了研究人员对如何提高网页上的数据集挖掘和应用的有效性和效率的研究。许多基于索引的网络搜索引擎可以对网络上的信息进行搜索，并根据关键词索引到相应的网页。利用这些搜索引擎，一些经验丰富的用户根据重要的关键词或词组得到自己想要的文献。然而，基于简单关键词的搜索引擎也存在一些问题。首先，任何一个主题可能和成千上万的文献相关，这样搜索引擎就会返回给用户大量的文献，而只有少量的文献和用户真正想要的相关。其次，和用户主题相关的文献可能不包含用户使用的关键词。和网页关键词搜索引擎相比，Web挖掘是一个从本质上提高网页搜索引擎的优秀的（和更有挑战性的）方法，因为Web挖掘可以鉴别出可靠的网页，对网页文献进行分类，以及解决网络搜索引擎中有歧义或细微差别的问题。

Web挖掘（或Web数据挖掘） 是指从Web数据中挖掘出本质关系（例如，用户感兴趣的和有用的信息）的过程。这些Web数据通常表现为文本信息、链接信息或使用信息。Web挖掘一词首次由Etzioni（1996）提出，现如今，许多会议、期刊和书中涉及Web数据挖掘。Web挖

掘通常应用于技术和商业领域。图5-8是Web挖掘的涉及的3个主要领域：Web内容挖掘、Web结构挖掘和Web使用挖掘。

图5-8 3个主要的Web挖掘领域

5.6 节复习题

1. Web知识挖掘面临的主要问题有哪些？
2. 什么是Web挖掘？和传统的数据挖掘有什么区别？
3. Web挖掘涉及到的3个主要领域是什么？

5.7 Web内容挖掘和Web结构挖掘

Web内容挖掘是指从网页上提取出有用的信息。由于网页上的文件资料是机器可读的，所以就能利用自动化的工具获取网页上的信息。**网络爬虫**能够自动地阅读网页上的信息，这些信息包含了类似文本挖掘中用到的文献特征，但其中也包含了其他一些概念，如文献的层次结构。Web内容挖掘能够提高搜索引擎的效果。如，Turetken and Sharda（2004）提出了一个可视化的系统，该系统阅读搜索引擎谷歌结果中的前100个文件，利用IBM的智能文本挖掘工具对这些文件进行聚类分析，并将结果以图表的格式展示出来。

除了文本信息，网页还包含了超链接，实现从一个网页链接到另一个网页上，超链接包含了许多隐藏的信息，这些信息可帮助自动推断出权威的概念。如果一个网页开发人员将网页链接到其他网页，链接页可以看做是其他网页的授权页。通过一个网页被不同开发人员链接的次数，就可以看出该网页的重要性并能够对授权网页进行挖掘（Miller，2005）。因此，大量的网页链接信息提供了网页信息的相关性、质量和网页内容的结构，也是Web挖掘一个丰富的资源。

在网页上搜索一个特殊主题时，只能得到少量的相关的高质量的网页，大部分网页都是没用的。利用以上的授权式网页的形式（或其他的办法）能够提高搜索结果的质量并将这些结果进行排列。这种授权（或授权式网页）的思想源于较早的信息检索工作，该工作利用期刊文献的引文来评价研究论文的意义（Miller，2005）。尽管网页链接采用了这个原始的思想，但是研究论文引文和网页超链接之间有很大区别。首先，并不是所有的网页都用签注的形式进行超链接（有些用导航条或付费广告的形式实现链接）。但是，如果大部分网页用签注的形式进行超链接，这种原始的思想就会被接受。其次，由于商业竞争，一个网页的链接很少会链接到相同领域的竞争对手的网页上。如，Microsoft公司不可能将自己的网页链接到Apple公司网页上。最后，授权网页上的信息很少是描述性的信息，如Yahoo！的主页上的信息不包含直接描述性的信息，而是一个网页搜索引擎。

网络超链接的结构引出了另外一个叫做中心的一类重要网页，中心网页是指某个或某些可以链接到其他各个授权网页的超链接集合。虽然中心网页可能不是很显著，该页上仅包含少数的几个链接，但是它提供了该领域内用户感兴趣的各种链接信息。一个中心网页可能链接个人

主页，也可能是某个课程网页推荐的参考文献，或者是关于某个话题的专业性资料，中心网页起到隐含对链接目标领域授权的作用。实际上，一个好的中心网页与被链接的网页之间存在着相互依存的关系，一个中心之所以是好的，是因为它链接到许多好的授权网页，而一个好的授权网页之所以好，是因为它被许多好的中心链接到。它们之间的这种关系使得从网页上自动检索到高质量的信息成为可能。

一个最熟悉和经常被提到的计算中心和授权网页之间的算法叫做**超链接主题搜索**（Hyperlink-Induced Topic Search，HITS），该算法最初由 Kleinberg（1999）提出，后来许多研究人员对它进行了改进。HITS 是一个利用网页里的超链接信息来评价网页的链接分析算法。在网络搜索中，HITS 算法为某个特定的查询找到基本的文献集。然后对每个文献的链接权威度和内容权威度进行回归分析。将这些基本文献集集中起来，就可以实现从一个搜索引擎中获得某个查询的最基本的集合。对每个被检索文献来说，原始文献和该文献所连接到的文献都被添加到文献集中。对这些文献将定进行回归分析和链接分析，直到链接权威度和内容权威度两个权值聚集到一点。根据这些权值对某个特定查询的文献制定索引并按重要性进行排序。

Web 结构挖掘是指从网络文件中的关联关系中挖掘出有用信息的过程。用于确定权威网页和中心网页，是网页质量排名算法的基础，该排名是一些重要搜索引擎（如 Google 和 Yahoo!）的核心竞争力。仅看链接到一个网页的情况就可以看出这个网页的知名度（或权威性），通过网页内的链接（或单个网站）情况可以看出该网页上对某个话题的分析的深度。对于分析大量网页之间的关联关系来说链接是一个重要的分析方法，能够帮助用户更好地理解某个特定社区网页、网络群体或网络团体内的关系。应用案例 5.6 讲述了一个借助 Web 内容挖掘和 Web 结构挖掘来更好分析美国极端主义群体内部联系的事例。

应用案例 5.6　网络追捕

我们一般从当前的外部环境中搜索相关问题的答案，然而，在搜索答案的过程中我们通常遇到许多问题。在处理全球恐怖分子问题上，通常一些本国的恐怖分子组织被忽视。其实，本国的恐怖分子对美国的威胁更大，因为他们了解更多的本国信息，而且他们的规模在不断壮大，通过网络他们能够找到国外的其他恐怖分子组织。

在网络上对相关内容实施有效的监视是很困难的，研究人员和权威人士需要更先进的工具对这些恐怖分子组织进行分析和监控。在国家安全部门和其他政府机构的帮助下，亚利桑那大学的研究人员研发了一种 Web 挖掘方法，该方法依靠因特网来发现和分析本国的恐怖分子组织的网站。恐怖分子组织通过因特网进行交流、访问个人信息并进行在线筹集资金。

该方法首先搜集重要恐怖分子成员和恐怖组织网站，再利用超链接连接到其他的恐怖分子和恐怖组织网站。与其他网站的相互连接是估算各个群体的目标和似法的一个重要因素。下一步是进行内容分析，这是在基本属性（例如，交流信息、集资、意识形态）分析基础上，对这些网站进行进一步规格。

基于链接分析和内容分析，研究人员已经获得了 97 个美国恐怖分子组织和仇恨群体的网站。通常，从这些组织之间的相互链接上，看不出任何合作关系。然而，每个组织之间的链接关系能帮助研究人员找出在同一领导下的组织。更深层次的数据挖掘研究用于解决全球性问题，找出全球的恐怖分子组织和美国恐怖分子组织之间的联系。

来源：Based on Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, Vol. 20, No. 5, September/October 2005, pp. 44–51.

5.7 节复习题

1. 什么是Web内容挖掘？Web内容挖掘和文本挖掘有什么区别？
2. 什么是Web结构挖掘？说明Web结构挖掘和Web内容挖掘的区别。
3. Web结构挖掘的目的是什么？
4. 什么是链接权威度和内容权威度？什么是HITS算法？

5.8 Web使用挖掘

Web使用挖掘是指从网页访问和交易中产生的数据中提取出有用的信息的过程。Masand et al.（2002）指出，网页访问产生的数据至少有以下3种类型：

1. 存储在服务器上的访问日志、引用日志、代理日志和客户端文字信息。
2. 用户特征。
3. 元数据，如网页属性、内容属性和使用数据。

分析Web服务器上的信息可以帮助我们更好地理解用户的行为特征，这种分析叫做点击流量分析。通过利用数据挖掘和文本挖掘技术，企业可以从点击流量中挖掘出有用的模型。如，可以了解到有60%的用户在搜索“毛伊岛宾馆”之前搜索了“飞向毛伊岛的航线”。这种信息能够帮助企业决定如何放置广告。点击流量分析还能够帮助我们了解用户的访问时间。例如，一个企业认识到有70%的用户下载软件的时间是从晚上7点到11点，这样企业可以在这段时间里提供更好的客户服务和更好的网络带宽。图5-9解释了从点击流量数据中获取知识和将知识用于改善服务、改善网页质量的过程，更重要的是提高客户价值。Naeraoui（2006）指出了Web挖掘的应用：

1. 决定客户终身价值。
2. 制定产品市场战略。
3. 评估促销宣传。
4. 根据用户的访问类型制定电子广告和优惠券。
5. 基于之前的学习规则和用户特征预测用户的行为。
6. 基于用户的兴趣和特征，向用户提供动态信息。

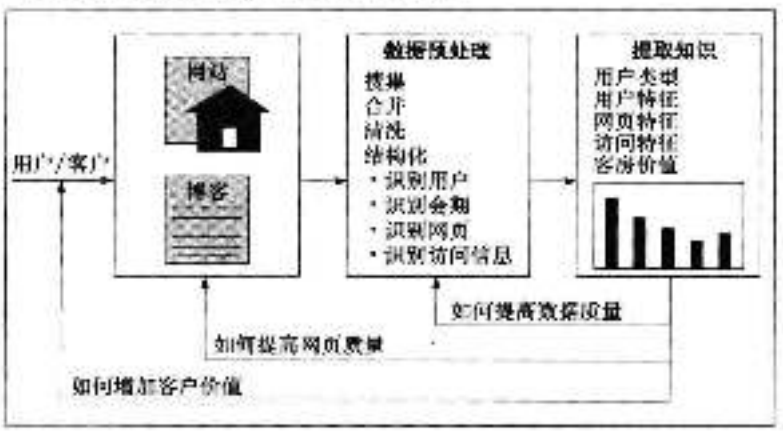


图5-9 从网页使用记录中提取知识的过程

Amazon.com提供了一个动态改变网页使用记录的例子。一个已注册的用户再次访问Amazon网页时，会显示欢迎用户的信息。这是一个通过客户使用记录（如计算机网站写在用户计算

机上一个简单的文件）来识别用户特征的一个简单应用。Amazon.com 还提供了一个根据以前的购买记录，或与该用户购买记录相近的其他用户的关联分析，为用户提供个性化的可选择的产品列表。而且在短时间内提供"优惠产品"。这些推荐都是通过对客户的分析，以及用户以前的购买行为进行聚类、序列模型挖掘、关联和其他数据、文本挖掘技术分析而得出的结论。

表5-3列出了一些重要的Web挖掘产品。

表5-3 Web使用挖掘软件

产品名称	产品介绍	URL
Angoss Knowledge WebMiner	包括ANGOSS Knowledge STUDIO和点击流量分析	angoss.com
ClickTrack	可以在网站上显示用户类型	clicktracks.com
LiveStats from DeepMetrix	实时日志分析，提供在线Demo	deepmetrix.com
Megaputer WebAnalyst	数据和文本挖掘应用	megaputer.com/products/wm.php3
MicroStrategy Web Traffic Analysis Module	网址流量、内容分析和Web访问分析报告	Applications/WTAM
SAS Web Analytic	网站流量分析	sas.com/solutions/webanalytics/
SPSS Web Mining for Clementine	Web事件挖掘	spss.com/web_mining_for_clementine
WebTrends	网站流量信息的数据挖掘	webtrends.com
XML Miner	利用模糊逻辑专家系统机制对XML文件中的数据和文本信息进行挖掘的系统和软件。	scientio.com

5.8 节复习题

1. 解释Web使用挖掘。
2. 在电子商务环境下，Web使用挖掘的应用有哪些？
3. 什么是点击流量数据？为什么点击流量数据在Web使用挖掘中非常重要？
4. 当用户访问一个网页并进行交互性的活动时，Web服务器收集了哪些信息类型？
5. 了解电子商务网站应用Web使用挖掘在提升客户价值中所发挥的作用。

5.9 Web挖掘的成功实例

Ask.com（ask.com）是一个知名的搜索引擎网站。Ask.com认为它能成功的基本因素在于它能提供更好的搜索结果的能力。然而，严格用数量指标（如点击率、报价和搜索频率）对搜索的结果进行评估几乎是不可能的，要有更好的数量和质量指标。Ask.com以定期地向它的观众调查问卷的方式，用数量和质量混合的指标作为关键绩效指标，如"用户声称找到了要找的东西的百分比"，"用户想再次使用该网站的百分比"和"无效搜索结果的百分比"，以及开放式客户对用户体验评估的质询。利用收集到的数量和质量两个方面的数据，Ask.com进行了"Ask 3D"设计，尽管在测试中，新设计和旧设计在纯粹的数量分析结果上没有区别。

Scholastic.com（scholastic.com）是一个专门提供儿童教育书籍的网上书店，该书店发现很多用户很难决定是否购买书籍。书店想知道的关键问题是"我们哪儿做错了？""为什么这些顾客不购买书籍？"以及最终"怎样才能留住这些顾客？"数据分析结果表明网站上可能不包含用户查找的书籍的书名。如，客户想从Scholastic的网站上查找出他们几十年前读过的书，前提是以为Scholastic仍然还有这些书。书店通过分析用户的这种需求，找出这些对不再出版书籍进

行搜索的用户的行为，确定未被满足的客户的数量和对以后客户购买行为所造成的影响。Scholastic 开始在网页上加上这些较老的书籍目录，如果用户需求的这些书籍不再出版，就会发邮件通知客户并对这些客户进行登记。最终，收到邮件的客户中有35%的客户买到了他们想要的书籍。

St. Jone Health System 是一个拥有8个医院和125个医疗站点和3 000名医生的保健系统。它的 CRM 系统中有110万客户。St. Jone 网站对业务处理的满意数据进行追踪，如用户通过在线注册的方式了解自己健康状况并访问医生，网站对这些新注册用户进行登记，从而了解有多少用户是新注册的。尽管保健行业的现状是市场竞争压力大，客户总数量呈下降趋势，但 St. Jone 却发现新用户的增长率是15%，并将四分之一的投资回报资金用于改善网站的满意度。这一成功事例表明作为组织的领导人员，应将在线客户的满意度作为包括全方位价值的一个关键绩效指标。但 St. Jone 通过对网站上的数据进行分析，制定了促使用户进入网站的广告策略，优先资助提高客户满意度的跨部门合作项目，并将客户的呼声作为企业决策的核心问题。

像 Ask. com、Scholastic 和 St. Jone Health System 这些有远见的公司，利用 Web 挖掘工具回答了以下几个关键问题："谁"、"为什么"和"怎么"。综上所述，有效地集成这些系统是很重要的，既能增加财务增长，又能提高顾客的忠诚度和满意度。

如果要持续增加自己的广告资金投入、资源，或许最重要的是顾客访问网站的渠道，企业管理者相信，利用观察顾客历史行为的 Web 挖掘技术比靠自己的直觉、预感和猜测更重要。应用案例5.7讲述了一个网络最优化的案例。

应用案例5.7 网络最优化生态系统

似乎网络上的任何东西都会被测量——每一次点击可以被记录，每一个场景都可以被捕获，每一次访问都被用于分析——这些记录都可以帮助企业实现网站最优化。不幸的是，用在线的方法实现"无限可测性"和"自动最优化"是非常复杂的。假设每次应用 Web 挖掘技术提供重要的范围见解，需要理解网站访问者行为可能是欺诈的并具有潜在风险。理想情况下，对用户访问行为的整体观察是需要的，这样可以捕获到定量和定性的数据。本节中介绍的这些有远见的企业（如，Ask. com、Scholastic. com 和 St. Jone Health System）已经致力于捕获和分析网站访问者的整个访问过程的记录，这就增加了企业的收益也提高了客户的忠诚度和满意度。

据 Peterson（2008）介绍，可以用两个坐标轴描述数据属性及如何使用数据，从而达到优化网站的目的。一个轴表示数据和信息，数据被定量化，信息被定性化。另一个坐标轴是度量值和行动；产生行动的度量值报表，分析和推荐措施；网站流程的实际改变，市场营销的最优化。这些维度产生的每一个象限利用不同的技术产生不同的输出结果，就像一个生态系统，每一个技术都和其他技术相互作用，从而对整个网络环境产生影响（见图5-10）。

多数人认为网站最优生态系统取决于日志、语法分析和对网站访问者点击流量行为的记录能力。决定这种能力的基本技术是 Web 分析。尽管 Web 分析工具提供了宝贵的见解，但是对访问者的行为分析只是一个定性地确定用户兴趣和意图，定量地确定网页的点击量的简单功能。幸运的是，还有另外两个根据用户的在线行为，对用户的行为提供定性分析并得到用户和访问者反馈的应用：客户体验管理（Customer Experience Management，CEM）和客户呼声（Voice of Customer，VOC）；

- Web分析的重要作用在于利用集成、挖掘和可视化数据，在线销售报告和能得访问者的努力，包括网页上访问者的交流信息，以及通过定义多步流程概括访问者流量等信息，解决“什么地点和什么时间”的问题。

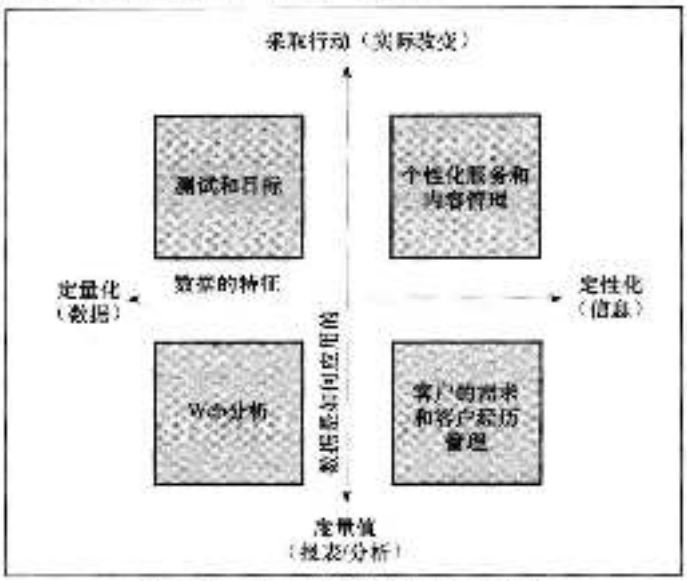


图5-10 用于网站优化的三维输入图

- 客户呼声的重要作用在于通过对网站访问者的反馈信息、网站监测信息和离线渠道进行收集和分析，以及对未来访问者行为的预测建模的支持，解决“什么人和什么方式”的问题。
- 客户体验管理的重要作用在于通过发现网络的应用问题，追踪和解决商业过程和使用上的障碍，报告现场绩效和实用性，实现实时变更和监控以及对访问者的行为进行深入诊断，解决“是什么和为什么”的问题。

这3个应用需要对用户的行为进行全面的分析，每个应用具有不同的价值并扮演着不同的角色。Web分析、客户体验管理和客户呼声构成了网站最优化生态系统的基础，支持在线企业影响产出的商业能力（图5-11以可视化的形式展现了网站最优化系统的过程）。

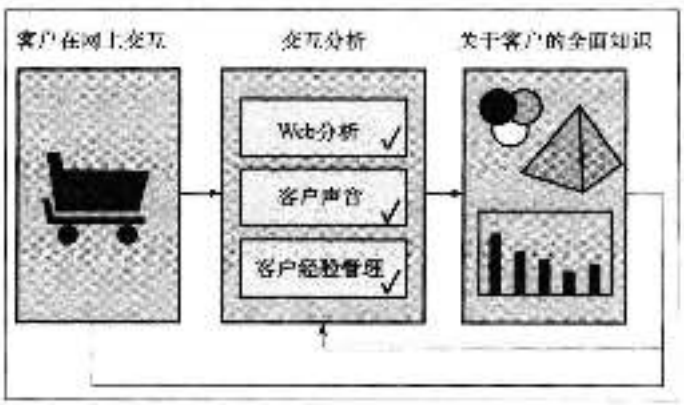


图5-11 从流程视角看网站最优化生态系统

这些相似但又有区别的应用，帮助网站运营者认识到绝大多数网站管理者所面临的挑战，并对这些挑战做出反应和回应。最优化流程的根本是度量，通过利用Web挖掘工具和技术将收集到的数据和信息转化为可以提升企业的有形分析和建议。如果应用合理的话，可以将这些应用收敛验证，将收集到的同一个访问者的不同数据集中处理，对访问者的行为进行更丰富更深入的理解。这种收敛验证模型——将描述同一个访问者的不同资料信息聚集在一起增加了分析结果的深度和丰富性——形成了网站最优化系统的架构。一方面，VOC应用提供了定性的输入信息；另一方面，对CEM提供的定量数据进行关键数据发现及减小数据的差异性。当正确实施时，3个系统提供的信息来自于同一用户，这些数据进行合并（通过数据集成项目或者执行好的简单分析处理过程）支持比任何单个生态系统成员更可操作的洞察力。

来源：Based on E. T. Peterson, "The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization," 2008 foresseeresults.com/Form_Epeterson_WebAnalytics.html(accessed on May 22,2009).

5.9 节复习题

1. 为什么我们需要Web挖掘？
2. 用自己的语言说明Web挖掘的优缺点。
3. Web挖掘成功事例的共同点有哪些？

本章重点

- 文本挖掘是指从非结构化（大部分是文本型）的数据资源中挖掘出有用的知识。假设大量的信息是以文本的形式存储的，文本挖掘是商务智能领域一个发展最快的分支之一。
- 企业通过对顾客在网页、博客和维基百科网站（Wiki）上留下的反馈信息进行文本挖掘和Web挖掘，来更好地理解顾客的需求。
- 文本挖掘应用实际上覆盖了商业和政府的每个方面，包括：市场营销、财务、保健、医学和国土安全。
- 文本挖掘利用自然语言程序将结构转化为文本，然后再利用数据挖掘算法，如分类、聚集、关联和序列等从文本中提取出知识。
- 成功的文本挖掘需要利用一个类似于数据挖掘中的CRISP-DM的结构化方法。
- 文本挖掘和信息提取、自然语言处理、文献总结紧密相关。
- 文本挖掘需要从非结构化的信息中产生数字型的指标，然后利用数据挖掘算法对这些数字型的指标进行分析。
- Web挖掘是指对网络上的、关于网络的以及基于网络工具的人们感兴趣的和有用的信息进行挖掘和分析。
- Web挖掘可以认为由以下3个部分组成：Web内容挖掘、Web结构挖掘和Web使用挖掘。
- Web内容挖掘是指自动地从网页中提取出有用的信息，这可以优化搜索引擎的搜索结果。
- Web结构挖掘是指从网页包含的链接中提取出人们感兴趣的信息，如谷歌的网页排名算法对显示的网页进行排序。
- Web结构挖掘可以识别出一个特殊组织中的成员，并有可能识别出每个成员所扮演的角色。
- Web使用挖掘是指从Web服务器日志、用户特征和交易信息中发现有用的知识。
- Web使用挖掘对CRM、个性化制定、站点导航和优化商业模型起到辅助作用。
- 文本和Web挖掘是下一代商务智能工具中的关键组件，帮助组织取得成功。

关键术语

analytical model 分析模型
analytical techniques 分析技术
association 关联
authoritative pages 权威网页
classification 分类
clickstream analysis 点击流量分析
clickstream data 点击流量数据
clustering 聚类
corpus 语料库
Customer Experience Management（CEM，客户体验管理）
deception detection 欺诈行为侦查
hub 网页中心
Hyperlink-Induced Topic Search（HITS，超链接搜索）
inverse document frequency 逆文献频率法
Natural Language Processing（NLP，自然语言处理）
part-of-speech tagging 词性标注
polysemes 多义词
search engine 搜索引擎
sentiment analysis 语义分析
sequence discovery 序列挖掘
Singular Value Decomposition（SVD，奇异价值分解）
speech synthesis 语音合成
stemming 词根
stop words 无用词
Term-Document Matrix（TDM，文献术语矩阵）
text mining 文本挖掘
tokenizing 标记处理
trend analysis 趋势分析
unstructured data 非结构化数据
Voice Of Customer（VOC，客户呼声）
Web analytic　Web 分析
Web content mining　Web 内容挖掘
Web crawler 网络爬虫
Web mining　Web 挖掘
Web structure mining　Web 结构挖掘
Web usage mining　Web 使用挖掘
Wiki 维基百科

讨论题

1. 说明数据挖掘、文本挖掘和 Web 挖掘之间的关系。
2. 企业在购买文本挖掘或 Web 挖掘之前要考虑哪些问题？
3. 讨论文本挖掘和 Web 挖掘的区别和联系。
4. 用自己的语言定义文本挖掘，并说明文本挖掘和 Web 挖掘的共同点。
5. 讨论数据挖掘处理过程（如 CRISP－DM）和本章提到的文本挖掘的 3 个步骤之间的相同点和不同点。
6. 什么是将信息转化为文本型数据？阐述转化为文本型数据的方法。
7. 文本挖掘中自然语言处理的作用是什么？分析文本挖掘中文本处理过程的作用和局限。
8. 指出并讨论文本挖掘的 3 个主要的应用领域。这 3 个应用领域中的共同主题是什么？
9. 分析 Web 挖掘和 Web 分析的关系。
10. Web 挖掘的有哪 3 个主要的应用领域？讨论它们的不同点和相同点。
11. 什么是 Web 内容挖掘？和文本挖掘有什么区别？用实际例子做出解释。
12. 什么是 Web 结构挖掘？什么是权威网页？它们和 Web 结构挖掘有什么关联？
13. 讨论 Web 结构挖掘的作用，从现实生活的应用中举出你熟悉的例子。
14. 什么是 Web 使用记录挖掘？用图形表示 Web 使用记录挖掘的过程，并解释该过程的主要步骤。
15. 举出两个典型的 Web 使用记录挖掘的商业例子，分析它们的作用和价值。

练习

Teradata 和其他的动手练习

1. 进入 teradatastudentnetwork.com，找到文本挖掘和 Web 挖掘的示例，阐述该领域最近的发展状况。如果在该网站上不能找到足够的资料，可以扩展到其他网站资源。

2. 进入teradatastudentnetwork.com，或者查看白皮书、网络研讨会或者其他的与文本挖掘和Web挖掘相关的资料，总结自己的发现，写一份总结报告。
3. 查找网络资源或图书馆数据库找出文本/Web挖掘与当今商业智能相关联的文献。

小组作业和角色扮演

1. 如何利用Web技术自动地获得文本型数据，一旦获取了这些数据，你可以从这些非结构化数据资料中提取出几种类型？
2. 对你所在学校的行政人员或所在组织的管理者进行访问，确定文本挖掘和Web挖掘在他们的工作中起到了什么作用，根据自己的发现写一份报告，报告中包含初步的成本/收益分析。
3. 进入自己的在线图书馆资源，知道如何对一个专业领域的文献（期刊文献）进行下载，下载并利用应用案例5.5中的方法对这些文献进行处理。

网上练习

1. 进入ClearForest.com和Megaputer.com以及dmreview.com和一些文本挖掘产品和服务提供者的网站，找出本章中没有涉及的一些文本挖掘工具和供应商。
2. 找出最近的Web挖掘应用的成功案例，找出文本挖掘供应商和咨询公司的例子或成功示例，写一份包含5个案例的报告。
3. 进入statsoft.com，选择"Downloads（下载）"，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
4. 进入sas.com，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
5. 进入spss.com，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
6. 进入terdeta.com，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
7. 进入fairisaac.com，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
8. 进入salfordsystem.com，下载至少3篇应用白皮书，这些应用中哪些用到了本章中提到的数据/文本/Web挖掘技术？
9. 进入kdnuggets.com，进入应用和软件部分，至少找出3个数据挖掘和文本挖掘组件。

本章结尾应用案例

HP和文本挖掘

惠普公司（Hewlett-Packard Company，HP）由William R. Hewlett和David Packard建立于1939年，总部设立在美国加利福尼亚州的帕罗奥图市。HP为全球的个人、中小型商业和大型企业提供产品、技术、解决方案和服务。同时HP也提供管理软件方案，使企业客户能够管理他们的IT基础、运作、应用、IT服务、业务流程以及各种应用平台。著名的HP产品种类包括：商务和消费者个人计算机、工作站、掌上电脑设备、喷墨打印机、数码娱乐系统、计算器以及和这些相关的配件、软件和服务，还包括了数码摄影和娱乐设备，图表，成像和打印硬件中的打印机耗材、打印设备、扫描仪和网络架构产品，如以太网交换机。零售商组成了公司分销渠道，公司还通过产品总分销、生产厂商和系统集成商来销售产品。

文本挖掘

HP的顾客通过电子邮件的形式和企业进行联系。结构化数据分析在发现特性方面是有效的，如消息是什么人、什么时间、什么地点和如何产生的。如果挖掘技术能够发现这些邮件发送的原因，就能够获取有价值的信息。电话服务中心是顾客和企业之间交互的一个普通方式。HP通过电话服务从与顾客交流的信息（如问卷文献、电子邮件和其他资源）中看到了商机。将结构化和非结构化的数据结合起来具有巨大的潜在价值，HP从中发现了商业价值。

系统

HP之前使用的标准工具不能够从顾客的交流资料中获取有用的信息。如今，HP用SAS的Institute's Text Miner能够从客服中心与顾客的交流资料中找出分析指标，然后再将这些指标进行标准化处理。HP利用文本挖掘工具将结构化数据和文本数据混合为结构化/非结构化的数据集，该数据集存储在Microsoft的SQL Server数据库中并提供在线分析处理引擎。如今，该系统包含了300 000个文本文献，几乎有500亿字节，覆盖了3个客服中心18个月的所有记录。

HP实现了管理者视图，该视图是一个由Temtec（termtec.com）开发的网络工具，该工具帮助HP利用SAS的企业挖掘工具中的预测模型，顾客忠诚度打分和客户差异性来扩展OLAP多维数据集。

流程

各种概念，如使用的产品、顾客打电话的频率和顾客存在的一些常见问题被应用于文本挖掘中，结果增加了文本聚集的程度。将这些聚集的文本和第三方提供的结构化数据联系起来，HP实现了结构化数据的组合和分析，如客户的心愿、态度和需求的收入。

由于文本资料的广度和分散性，文本分析成为一个富有挑战性的工作。不同的顾客数据库中包含不同的结构化信息，这些信息很容易集成，问题在于文本中除了包含了结构化信息之外还包含了非结构化信息。SAS的Text Miner采用了单一价值分解技术，该文本挖掘软件需要预先制定一个词典和同义词列表；然而，组织制定一个符合自身情况的商业环境信息集是一项非常复杂的工作。除了传统的数据仓库，文本数据可以应用于各种环境中。SAS的Text Miner所面临的最大挑战是找出顾客在HP网站上的活动，以及从这些顾客活动中发掘出商机。

除了文本挖掘的主要应用之外，SAS的Text Miner还可以对顾客网站行为进行预测，从而为HP提供用户的潜在需求。同时，该工具还能利用文本中各种数据和信息对供应商/厂商进行多层次的分析。

结论

SAS的Text Miner能够实现标准的数据定义，而且保证了产品分类模型的准确度达到80%以上。该系统通过改进的交叉销售、目标市场营销、顾客持有量和更好地预测顾客需求使HP成为领先企业。结构化/非结构化数据中产生的信息支持企业不同部门的各种业务。

本章结尾案例的问题

1. 文本挖掘典型的应用有哪些？
2. 文本挖掘技术是如何应用到其他商业中的？
3. HP的文本挖掘的挑战有哪些？是怎么克服的？
4. 你认为在其他领域HP能够利用文本挖掘吗？

来源：Based on M. Hammond, "BI Case Study: What's in a Word? For Hewlett-Packard, It's Customer Insight," *Business Intelligence Journal*, Vol. 9, No. 3, Summer 2004, pp. 48 – 51; and B. Deal, "Text Mining: A Golden Opportunity for HP," *SearchCRM.com*, June 6, 2005, searchdatamanagement.techtarget.com/originalContent/0,289142,sid91_gci1136611,00.html (accessed November 2008).

参考文献

Deal, B. (2005, June 6). "Text Mining: A Golden Opportunity for HP." *SearchCRM.com*. **searchdatamanagement.techtarget.com/originalContent/0,289142,sid91_gci1136611,00.html** (accessed November 2009).

Bond, C. F., and B. M. DePaulo. (2006). "Accuracy of Deception Judgments." *Personality and Social Psychology Reports*, Vol. 10, No. 3, pp. 214–234.

Chaem, P. K. (2001). "Kodak Turns Knowledge Gained About Patents into Competitive Intelligence." *Knowledge Management*, pp. 11–12.

Chun, H. W., Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, and T. Hishiki. (2005). "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning." *Proceedings of the 11th Pacific Symposium on Biocomputing*, pp. 4–15.

Coussement, K., and D. Van Den Poel. (2009). "Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers." *Expert Systems with Applications*, Vol. 36, No. 5, pp. 6127–6134.

Coussement, K., and D. Van Den Poel. (2008). "Improving Customer Complaint Management by Automatic Email Classification Using Linguistic Style Features as Predictors." *Decision Support Systems*, Vol. 44, No. 4, pp. 870–882.

Delen, D., and M. Crossland. (2008). "Seeding the Survey and Analysis of Research Literature with Text Mining." *Expert Systems with Applications*, Vol. 34, No. 3, pp. 1707–1720.

Etzioni, O. (1996). "The World Wide Web: Quagmire or Gold Mine?" *Communications of the ACM*, Vol. 39, No. 11, pp. 65–68.

EUROPOL. (2007). "EU SOCOL Work Program for the 2007." **statewatch.org/news/2006/apr/europol-work-programme-2007.pdf** (accessed October 2008).

Feldman, R., and J. Sanger. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Boston, MA: ABS Ventures.

Froelich, J., S. Ananyan, and D. L. Olson. (2005). "Business Intelligence Through Text Mining." *Business Intelligence Journal*, Vol. 10, No. 1, pp. 48–50.

Fuller, C. M., D. Biros, and D. Delen. (2008). "Exploration of

Feature Selection and Advanced Classification Models for High-Stakes Deception Detection." *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*. Big Island, HI: IEEE Press, pp. 80–99.

Ghani, R., K. Probst, Y. Liu, M. Krema, and A. Fano. (2006). "Text Mining for Product Attribute Extraction." *SIGKDD Explorations*, Vol. 8, No. 1, pp. 41–48.

Hammond, M. (Summer 2004). "BI Case Study: What's in a Word? For Hewlett-Packard, It's Customer Insight." *Business Intelligence Journal*, Vol. 9, No. 3, pp. 49–51.

Han, J., and M. Kamber. (2006). *Data Mining: Concepts and Techniques*. 2nd ed. San Francisco, CA: Morgan Kaufmann.

Kanayama, H., and T. Nasukawa. (2006). "Fully Automatic Lexicon Expanding for Domain-Oriented Sentiment Analysis. (EMNLP: Empirical Methods in Natural Language Processing." **trl.ibm.com/projects/textmining/takmi/sentiment_analysis_e.htm.**

Kleinberg, J. (1999). "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632.

Lin, J., and D. Demner-Fushman. (2005). "Bag of Words Is Not Enough for Strength of Evidence Classification." *AMIA Annual Symposium Proceedings*, pp. 1031–1032. **pubmedcentral.nih.gov/articlerender.fcgi?artid=1560897.**

Mahgoub, H., D. Rösner, N. Ismail, and F. Torkey. (2008). "A Text Mining Technique Using Association Rules Extraction." *International Journal of Computational Intelligence*, Vol. 4, No. 1, pp. 21–28.

Manning, C. D., and H. Schutze. (2009). *Foundations of Statistical Natural Language Processing (Second Edition)*. Cambridge, MA: MIT Press.

Mobasher, B. M., M. Spiliopoulou, J. Srivastava, and O. R. Zaïane. (2002). "Web Mining for Usage Patterns and Profiles." *SIGKDD Explorations*, Vol. 4, No. 2, pp. 125–132.

McKnight, W. (2005, January 1). "Text Data Mining in Business Intelligence." *Information Management Magazine*. **information-management.com/issues/20050101/1016487-1.html** (accessed May 22, 2009).

Menn, J. (2003). *Investigative Data Mining for Security and Criminal Detection*. Burlington, MA: Elsevier Science.

Miller, T. W. (2005). *Data and Text Mining: A Business Applications Approach*. Upper Saddle River, NJ: Prentice Hall.

MITRE Corporation. **mitre.org** (accessed on May 20, 2009).

Nasraoui, O., M. Spiliopoulou, J. Srivastava, B. Mobasher, and B. Masand. (2006). "WebKDD 2006: Web Mining and Web Usage Analysis Post-Workshop Report." *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 2, pp. 84–89.

Nakov, P., A. Schwartz, B. Wolf, and M. A. Hearst. (2003). "Supporting Annotation Layers for Natural Language Processing." *Proceedings of the ACL*, interactive poster and demonstration sessions. Ann Arbor, MI: Association for Computational Linguistics, pp. 65–68.

Peterson, E. T. (2008). "The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization." **foreseeresults.com/Form_Epeterson_WebAnalytics.html** (accessed May 22, 2009).

Shatkay, H., A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher. (2007). "SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by Integrating Text and Protein Sequence Data." *Bioinformatics*, Vol. 23, No. 11, pp. 1410–1417.

SPSS. "Merck Sharp & Dohme." **storieshttp://www.spss.com/success/template_view.cfm?Story_ID=185** (accessed May 15, 2009).

StatSoft. (2006). *STATISTICA Data and Text Miner User Manual*. Tulsa, OK: StatSoft, Inc.

Tseng, Y-H., C-J. Lin, and Y-I. Lin. (2007). "Text Mining Techniques for Patent Analysis." *Information Processing & Management*, Vol. 43, No. 5, pp. 1216–1245.

Turetken, O., and R. Sharda. (2004). "Development of a Fisheye-based Information Search Processing Aid (FISPA) for Managing Information Overload in the Web Environment." *Decision Support Systems*, Vol. 37, No. 3, pp. 415–434.

Wang, S. S., and C. K. Liu. (2004). "Using Text Classification and Multiple Concepts to Answer E-mails." *Expert Systems with Applications*, Vol. 26, No. 4, pp. 529–543.

Zhou, Y., E. Reid, J. Qin, H. Chen, and G. Lai. (2005). "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis." *IEEE Intelligent Systems*, Vol. 20, No. 5, pp. 44–51.

商务智能实施：整合和新兴趋势

学习目标

- 描述 BI 实施的主要问题
- 列出 BI 实施的关键成功因素
- 描述整合 BI 技术和应用的重要性和问题
- 理解将 BI 系统和其他信息系统连接的需要以及如何实施
- 定义面向需求的 BI 及其优势与限制
- 列出和描述具有代表性的隐私、BI 实施的主要法律和道德问题
- 理解 Web 2.0 及其与 BI 和决策支持相关的特点
- 理解社交网络概念、选择的应用和它们与 BI 的关系
- 描述虚拟世界技术是如何改变 BI 应用的使用的
- 描述 BI 应用中社交软件的整合
- 理解 RFID 数据分析是如何改善供应链管理和其他操作的
- 描述海量数据获得技术是如何使得现实挖掘实现的

介绍

本章涉及 BI 实施的主要问题，也介绍了一些可能会影响 BI 应用的新兴技术。多个其他有趣的技术也正在兴起，但是我们主要介绍了已经被实现的和一些将要影响到 BI 的技术。我们介绍了这些新兴技术，探讨了它们的应用，总结了它们和 BI 之间的关系。我们讨论了 4 个主要的实施问题：整合，与数据库和其他信息系统连接，基于需求的 BI，可能影响到 BI 实施的法律、隐私、道德问题。我们以一个案例结束这一章，这个案例描述了一种创新使用无线射频识别设备、BI 和决策支持的方法。

开篇场景：BI Eastern Mountain Sports 增加合作和生产力

Eastern Mountain Sports 是个中等大小的特产零售商（2009 年销售额为 20 000 万美元），它通过订单目录和在线形式，利用全国 80 个实体商店销售货物。Sports 的业务是在一个竞争很激烈的环境中进行的。公司需要做的决策包括：持续的产品开发、市场营销、生产、销售。好的决策需要来自员工、顾客、供应商的输入和合作。在过去的几年中，公司实施了 BI 系统，这个系统包括业务绩效管理和仪表盘。BI 系统从多种渠道收集原始数据，将它们处理成数据，实施对比绩效与操作标准等分析，从而评估商业健康程度（见图 6-1）。

接下来介绍系统是如何工作的。在 IBM 大型计算机中能够获得的销售点信息和其他相关数据，被载入到 Microsoft SQL 和数据集市中。数据随后被 Information Builders 的 WebFOCUS 7.12 平台进行分析。结果通过一系列用户能够通过网络浏览器查看的仪表盘展示出来。这就使得用户能够看见统一、高水平的关键绩效指标，例如，销售额、库存、边际利润情况，之后将指标分解成更小的粒度使其能够分析特定的业务。

尽管采用了尖端技术，但系统由于缺乏所有参与者之间的数据、沟通和合作，这个系统一直运行得不是很好。

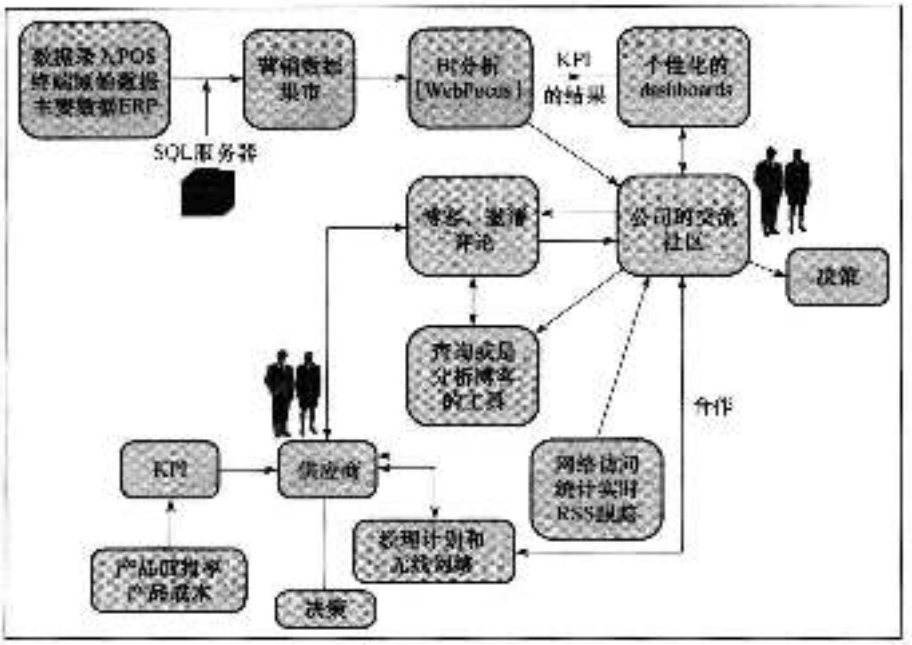


图 6-1 Eastern Mountain Sports 的协同决策概定

解决方案：将 BI 与社交软件整合

公司搭建了一个叫做 E-Basecamp 的多功能员工工作平台。E-Basecamp 包括与公司目标相关的信息，这些公司目标与生产率工具（如电子表格），以及以角色为基础为每个用户定制的内容相整合。该系统使得内部和外部相关利益者们的合作更加方便。EMS 正在使用的操作标准有 20 条（例如库存水平和转换）。这些标准也包括电子零售、网络营销，营销经理在这里可以监测每小时内的网络流量和转换率。仪表盘通过颜色码来显示与目标的偏离情况。

- **RSS Feed** RSS Feed（见 6.7 节）被嵌入到仪表盘中来驱动更多被关注的查询。这些 Feed 是信息共享和在线转换的基础。例如，通过显示哪个产品比其他的销售得更好，用户能够综合分析交易的特性并产生高销售额的销售行为。获得的知识在组织机构内传递。例如，一个经理观察到在商店 X 内鞋类销售量飙升。调查显示商店 X 的员工已经完善了一个多步骤的销售技巧，这种技巧包括（在网上和商店中）推荐特定的袜子，还同鞋底内部为特殊用途进行设计。信息通过 RSS Feed 进行传播。结果，鞋类的销售额在那一年增加了 57%。
- **Wiki（维基百科）** Wiki 用来鼓励公司中合作性的交互。仪表盘用户被鼓励去做出假设和要求来寻求帮助，然后发起评论和建议，就像是仪表盘边上的记事本。
- **Blog（博客）** Blog 应用在特定的数据或是关键标准中，用来发布信息和发起评论。工具常被用来存档、查询并且为了更容易查询可将 Blog 进行分类。例如，商店经理发布一项调查或者是对于销售偏离目标的解释。在 Blog 上进行评论使得读者能够观察到他们单独使用数据分析模式可能忽视的类型。
- **Twitter（推特）** 在 2009 年，微博变得非常流行。这项技术方便了沟通和合作，加速了业务处理。
- **社交网络服务** 鼓励员工和商业伙伴加入到 LinkedIn 中。主要的应用就是加入论坛和使用回答功能。这种工具鼓励外部的沟通和合作。

与商业伙伴外部联系

供应商通过 Blog、Wiki、RSS Feed 与 BI 系统相联系。例如，供应商可以在仪表盘上监测退货率，然后通过 Blog 或者 Wiki 邀请商店经理提供如何降低退货率的建议。假设安装了恰当的保护措施，那么供应商能够得到他们产品销售情况的实时数据，能够准备更好的产品计划。同样 Twitter 和 LinkedIn 在商业伙伴中也被广泛地应用。

目标就是与商业伙伴建立更加紧密的联系。例如，通过在供应商的仪表盘中安装 Blog，供应商就能够查看实时的销售信息并且在 Blog 上发布评论。产品经理利用 Wiki 来发布下一季的挑战（例如建议的销售增长百分比），然后让供应商提供创新的方法来实现目标。许多顾客和其他的商业伙伴订阅了 RSS Feed。

Blog 也能够嵌入 EMS 公司的产品生命周期管理（Product Lifecycle Management，PLM）工具中。这允许供应商进行产品开发管理的虚拟对话。

结果

BI 和社交软件结合的主要影响就是在 Blog、Wiki、RSS Feed 和论坛等这些双方都有兴趣参与的地方进行对话。这加速了信息的流动，增加了人们的参与度。销售和边际利润都呈持续增长趋势。

开篇场景的问题

1. BI 系统的原始动能是什么？
2. 为什么将 BI 和社交软件整合是有益的？
3. 区别案例中的内部和外部整合以及它们对于 EMS 的贡献。
4. 整合对于供应商的益处是什么？

我们从开篇场景中能够学到什么

通过将 BI 和社交软件整合，EMS 成功地加强了自己的经理们和供应商之间的沟通和合作。这样的整合叫做协同决策（Collaborative Decision Making，CDM）（详见 6.10 节）。社交软件是基于新的计算范例的 Web 2.0（见 6.7 节～6.10 节）。社交软件工具使得内部和外部的沟通和合作更加便利。这种整合就是一个将 BI 与其他的信息系统相整合能做什么的例子（6.2 节和 6.3 节）。在 BI 实施中，整合是遇到的主要问题（6.2 节）。6.5 节讨论了 BI 实施中遇到的法律、隐私和道德问题。6.6 节列举了一些趋势和新兴的技术，这些技术会在本章后面的节中给予描述。Web 2.0（6.7 节）、社交网络（6.8 节，6.10 节）和虚拟世界（6.9 节）是和 BI 相关的尖端技术。本书以其他的新兴问题，包括面向需求的 BI（6.4 节）、RFID（6.11 节）的使用和现实挖掘（6.12 节）等结束全书。

资料来源：Based on Neville, J., "EMS: Adventures in X-treme Web 2.0," *Optimize*, Vol. 6, No. 1, January 2007, p. 33, (accessed Jan. 2010) and from ems. com (accessed January 2010).

6.1 BI 实施：概述

实施 BI 系统非常复杂。除了一般的信息系统实施中的典型问题外，例如用太形变量证明系统的方法来进行合适的成本－利润分析和处理对于变化的抵抗，还有整合、安全、系统可扩展性、数据仓库的建设、分析和仪表盘等许多问题。

6.1.1 BI 实施因素

许多因素能够影响到 BI 的实施。这些因素是技术、行政管理和行为等。许多因素都是信息系统所普遍具有的，并且在信息系统文献中进行了广泛的研究。根据 Asif（2009）的报道，影响 BI 实施决策过程的主要因素有以下这些：

1. 报告和分析工具
a. 特征和功能
b. 可扩展性和可部署性
c. 可用性和可管理性
d. 定制应用的能力
2. 数据库
a. 可扩展性和性能
b. 可管理性和可用性
c. 安全和可定制性
d. 可回写的能力
3. 提取、转换和加载（ETL）工具
a. 读取任何资源的能力
b. 高效和生产率
c. 跨平台支持
4. 涉及的成本
a. 硬件成本（实际或是机会成本）
b. 软件成本（ETL、数据库、应用、前端）
c. 内部开发成本
d. 外部开发成本
e. 内部培训
f. 持续维护成本
5. 利润
a. 节约时间和操作便利
b. 更低的操作成本
c. 改善的客户服务和满意度
d. 改善的业务和战略决策水平
e. 改善的员工交流情况和满意度
f. 改善的知识分享情况

这些因素需要定性和定量的分析。

BI实施的关键成功因素 虽然可能有许多因素影响到BI的实施，但Wikipedia（http://en.wikipedia.org/wiki/Business_intelligence，2010年1月）显示的一份Vodapalli（2009）报告表明，以下是影响BI成功实施的因素：

a. 业务驱动的方法和项目管理
b. 清晰的愿景和计划
c. 管理支持和赞助
d. 数据管理和质量问题
e. 将解决方法与用户需求相匹配
f. BI系统运行情况考虑
g. 强大的可扩展性的框架

6.1.2 BI实施中的管理问题

有许多与BI实施相关的管理问题，如下面所述：

1. 系统开发和整合的需求 开发一个有效的系统是非常复杂的。为此，许多BI供应商提供高度整合的应用的选择，这些选择包括与ERP和CRM等系统的连接（见6.3节）。著名的公司有Oracle、Business Objects、MicroStrategy、IBM和Microsoft。多数BI供应商提供应用整合，通常利用网络是可行的。

2. 成本－利润问题和理由 本书中讨论的BI解决方案是非常昂贵的，并且只是在大的公司才有理由去做。小的组织机构如果能够利用现存的数据库而不是新建数据库的话，是可以使得解决方案符合成本要求的。一个解决方案就是采用面向需求的BI。然而，在BI实施之前都要进行仔细的成本－利润分析。

3. 法律问题和隐私 BI分析专家可能建议一个公司将电子或者打印的目录或者促销发送给一个年龄段或是一种性别的顾客。一个男性顾客起诉过Victor的Secret这个品牌（Limitedbrands的一个品牌），原因就是他的女性邻居收到的邮寄订单目录包含有折扣的商品，但是他收到的仅仅是常规的产品目录（折扣通常是在大量购买的时候才会有）。处理歧视诉讼会很昂贵。有些数据挖掘会侵犯个人隐私。

公司会采取什么措施来保护消费者？消费者应该如何保护自己的隐私？这些问题应该在BI解决方案实施时给予充分的考虑。更多的问题在6.5节中讨论。

4. BI和BPM的现状和未来 一个组织机构商业信息的质量和及时性有时不仅仅是获利和亏损的问题，有时会关系到一个组织的生死存亡。没有一个公司否认BI和BPM的好处。最近的行业分析报告显示在未来的几年里，成千上万的人们将会每天使用BPM仪表盘和业务分析（BA）。企业正在通过将信息发送给不同类型的员工和最大化地利用现存数据资产获得更多的价值。可视化工具包括正在被生产商、零售商、政府和专门代理机构使用的仪表盘。特定行业的分析工具会大量地出现在市场的支持分析和从高层到用户层的已有的决策中。BI利用现存的IT技术帮助公司利用他们的IT投资，使用他们的遗留和实时数据。这样有计划的、仔细的、积极主动的BI实施方法对于竞争来说是很有必要的。

5. 成本理由，无形利润 虽然企业提供有形利润，但是将他们的无形利润数量化是非常困难的。在一个高能量成本、抵押危机、政治动荡、经济不断下滑的环境中，IT投资必须要经过经济性的证明。

6. 文档化和安全支持系统 许多员工开发自己的决策支持系统或是BI模块，以提高他们的生产率和工作质量。将这些特别的系统保存起来，并确信实现文档化和具有安全措施，可以使保这些员工不在或者离开公司之后，生产率工具仍然可以使用。采取合适的保护措施是必需的。最终用户不是开发他们自己BI应用的专业系统开发者。由于这个原因，这里可能存在数据完整性和系统开发安全性问题。

7. 道德问题 BI和预测分析可能导致产生的道德问题，例如隐私和问责制。另外，错误能够对他人和公司造成伤害。例如，一家公司开发了一个决策支持系统来帮助人们计算提早退休的财务影响。然而，DSS开发者没有包括纳税影响，这就导致了不正确的退休决策。另外一个重要的道德问题就是决策中的关键因素——人类判断。人类判断有可能是主观的或是不正确的，所以它有可能导致不道德的决策。公司应该为系统开发者提供道德准则。同样，将经理工作自动化可能会导致大量的失业。专家系统和其他智能系统的实施同样会与道德有关。专家系统建议的行为可能是不道德甚至是违法的。例如，专家系统可能建议你做一些会伤害他人的事情，或者侵犯某些人的隐私。一个例子就是机器人的行为和机器人不按照程序设定的方式来行动的可能性。已经有很多行业的意外是由于机器人发生的，从而导致了许多伤亡。问题就是：组织是否应该使用一个不是100%安全的节约生产力的设备？另外一个道德问题就是从人类获得的信息的使用。这里的问题是：当员工的知识被他人使用时，公司是否应该补偿这名员工？与之相关的问题

还有动机问题。它还涉及了隐私。是否应该告知人们谁提供了某些知识？最后一个需要强调的道德问题就是：非人性化和感觉机器能够比人更加聪明。人们可能对于聪明的机器有不同的态度，这些态度会在他们的工作方式中体现出来。

8. BI项目失败的例子 所有类型的BI项目都有很多失败的例子。这样的失败有很多的原因，从人的因素到软件错误。下面是一些具体例子：

a. 没有认识到BI项目是企业范围内的商业活动，没有认识到它和独立的解决方案不同。
b. 缺少能够确保资金的商业赞助者。
c. 缺少与来自功能区域的商业代表的合作。
d. 缺少有能力的可用员工。
e. 没有认识到对商业利润有负面影响的"脏数据"的重要性。
f. 太过依赖供应商。

6.1 节复习题

1. 影响BI实施的因素主要有几种类型？
2. 列出工具和数据库方面的因素。
3. 列出管理方面的问题。
4. BI项目成功的关键因素是什么？

6.2 BI和整合实施

为了提高系统支持任务的有效性和效率，整合信息系统在企业中是广泛实施的。BI的实施几乎总是需要一个或是多个整合步骤。然而，整合就像接下来描述的一样并不简单。

6.2.1 整合的类型

计算机系统能够进行整合以使得系统的构成部分作为一个整体运行，而不是各自分散工作。整合可以在开发阶段也可以在应用系统阶段（也就是我们的主要兴趣领域——应用整合）。整合被认为是最为重要的问题已经好多年了（Spangler，2005）。有以下几种类型的整合：数据整合、应用整合、方法整合、流程整合。整合能够从其他两种特征来观察：功能和物理。

功能整合是指一个系统提供不同的应用。例如，在同一个系统中能够完成用电子邮件工作、使用电子表格、与外部数据库进行交流、产生图形表示、存储或是操作数据。相似地，在同一个交互界面能够同时使用商业分析工具和仪表盘，使用一个菜单和产生一个输出。

物理整合就是将软件、硬件和通信功能进行打包以实现功能整合。本章中的讨论主要是功能应用整合，这种整合能够以两种方式进行。

- 两个或者多个决策支持应用的整合，实现统一应用。
- 一个或者多个BI工具与其他信息系统（例如博客、知识管理、数据库、财务系统）的整合。

整合能够在一个公司中（内部整合）或是在两个公司之间（外部整合）进行。

6.2.2 为什么进行整合

BI软件整合有以下几个主要目标：

- **实施BI** 为了BI系统的运行，BI通常需要与数据源、实用程序和其他应用连接。这样的连接必须有效和高效地完成。
- **提高BI应用的能力** 许多BI开发工具可以相互补充。每个工具在它最擅长的分任务运

行中表现得最好。例如，BA能够用来推荐最优资源分配计划，仪表盘能够提供偏离计划管理预警的控制系统。开始的引例证明社交软件是如何使得BI运行得更好的。

- **实现实时决策支持** 通过紧密的整合，在实时环境中支持决策的制定是有可能的。例如一个运输系统应用无线通信和网络服务产生数据流。
- **实现更强大的应用** 例如，利用智能系统提供实时能力。
- **方便系统开发** 紧密的整合实现了更快的开发和系统组件之间的沟通。
- **平衡支持活动** 多个支持活动能够改善BI应用的运行。例如，博客、Twitter、Wiki、RSS Feed的提供，像在引例中展示的沟通和合作那样。

BI整合的结果是可能会提高不通过整合达不到的能力。关于成功整合的战略，参见Muegenthal（2005）。

6.2.3 BI整合的水平

前面提到的功能整合，能够在以下两个不同的层次进行：不同的BI之间和BI系统内部。这些类型的BI整合对于解决重复性和顺序性决策问题的系统是适合的。BI通过帮助将一个系统的输出转化成另外一个系统的输入为整合提供方便。结合多个分析，每次访问复杂决策问题的特定部分，是BI之间整合的一个例子。例如，一个支持营销活动决策的BA模型能够与一个支持改善生产计划的供应链的模型相结合，在此过程中将第一个系统的某些输出转化成第二个系统的输入。

第二种水平的整合是指在建立一个复杂的BI系统过程中将多个合适的BI技术进行整合，特别是利用某些技术长优势。

6.2.4 嵌入式智能系统

在过去的几年中，我们发现很多为了实施分析而嵌入了智能模块的系统。在这样的系统中，智能部分（例如智能代理）对于用户是不透明的，可能在实时的环境中工作。自动决策系统（Automated Decision System，ADS）就属于这种类型。

在大型或者复杂的BI系统中嵌入智能组件越来越成为一种趋势，以下就是一些例子：

- 电脑电话与智能电话中心整合，用来选择和分配能够实时处理特定顾客的人工代理。
- 在OLTP系统中建立的实时决策判定，例如在协作计划、预测、供应链管理（Supply Chain Management，SCM）中的增强，实时计划决策支持。
- 使用内置的智能代理支持战略管理计划和分析。
- 流程实施和协同决策判定管理的智能代理。

6.2 节复习题

1. 列出几种整合的类型。
2. 描述BI整合的需要。
3. 列出整合的不同层次。
4. 描述BI与非BI系统之间的整合。
5. 定义嵌入式智能系统并描述它们的好处。

6.3 BI系统与数据库和其他企业系统的连接

BI应用，特别是大型的应用，需要和其他的信息系统进行连接。本部分讨论的主要的整合领域是与数据库和后端系统的连接。

6.3.1 与数据库连接

几乎每个BI系统都需要与数据库和数据仓库（或是数据集市）相连接。例如，当BI分析顾客订单的时候，需要在数据仓库中找到产品描述、库存数量、订单信息。BI应用能够通过多种方式与数据库连接。今天，许多这样的连接是通过图6-2中描述的多层应用架构来实现的。这个架构包括4层：

1. Web浏览器，在这层，将数据和信息提交给用户，同时收集来自用户的数据。

2. Web服务器，这层主要是传输网页，收集最终用户的信息，同时传递和接收来自应用层的数据。

3. 应用服务器执行商业规则（例如用户授权），定制从Web服务器传输过来的基于数据的数据库查询，将这些查询传送给终端数据库（或者数据仓库或者数据集市），操作和格式化查询产生的数据，并将格式化响应传给Web服务器。

4. 数据库（数据仓库或者集市）服务器。数据存储和管理在这一层，同时此层对用户的请求给予处理。

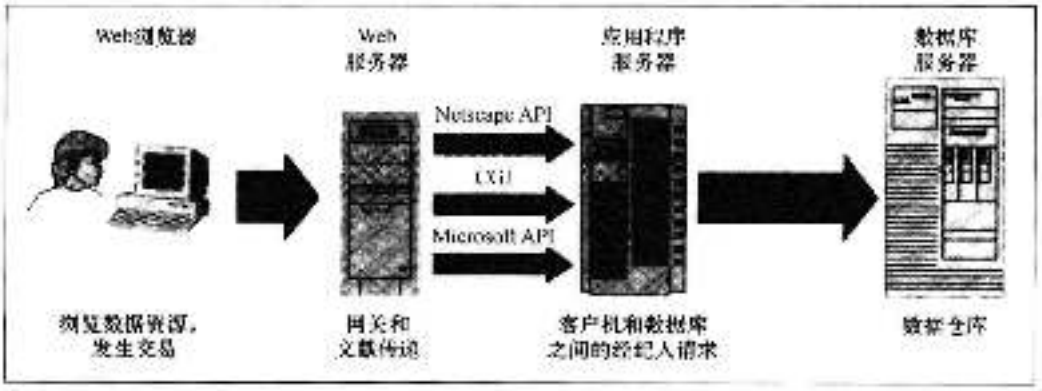


图6-2 多层应用架构

6.3.2 BI应用和后端系统的整合

许多技术能够用来将BI系统直接与后端应用进行整合。例如，只有一个数据集市（或者市场营销），但是需要与库存或者其他后端应用或者数据库进行连接。许多商业BI套件有内置整合能力。如果一个公司想要开发自己的数据库接口，那么可以实施许多选择。首先，所有的Web脚本语言（例如PHP、JSP、ASP）都有简化连接过程的命令。特别是这些脚本语言使得一个开发者能够开发向数据库发出请求的网页，同时网页能够处理数据库对于请求的响应。第二，许多专业的应用服务器能够简化BI应用和一个或者多个后端数据库的连接。在这些专业的服务器中，BEA Inc.的WebLogic（bea.com）服务器是市场中的佼佼者（现在是Oracle的一部分）。

除了与后端数据库连接外，许多BI应用也需要与其他的系统进行整合——ERP、CRM、知识管理、供应链管理、电子数据交换系统和其他的在企业内外的应用。这样的整合能够由企业应用集成（Enterprise Application Integration，EAI）软件进行处理。这个软件主要解决大型应用的整合。TIBCO（tibco.com）、WebMethods（softwareag.com）、WebSphere InterChange Server（来自IBM）都是一些提供EAI平台的公司。

有时，整合需要流程的重新设计。例如，Sterngold Corp（sterngold.com）需要将它的订单系统和后台（例如订单完成、库存、财务、支付）整合。整合需要对现有流程的研究和对修改

流程的重新设计。

一个重要的整合就是大型BI与企业ERP之间的整合。

整合BI和ERP以改善决策支持 在过去的许多年中，ERP平台仅具有交易处理能力和诸如简单的报告、分析和按照利润、营业额、顾客满意度将产品进行分类的能力。复杂的报告和分析都来自独立的BI系统。然而，公司认识到，如果能在员工工作的应用环境中实施分析或者与BI的连接，那么员工会处理得更好。所以，ERP供应商们开始在他们的平台中开发商务分析，以至于顾客不用在不同系统中转换。这可能导致决策质量的巨大改善。例如，当一个销售人员在接收订单的时候，BI类供所有需要的信息来决定是否给顾客信用额度，给多少信用额度。

ERP/BI结合通常应用在财务系统。但是，公司正在将它们应用在市场营销、人力资源和生产制造的各种领域中。

高水平分析需要将来自不同系统的信息放到一起，包括供应链管理、生产执行系统、客户关系管理和产品生命周期管理。通过计划、预测、情景模拟可以实现更好的决策。

然而，在各种系统之间的整合仍然需要花费时间。全部的整合可能会花费数月或者数年的时间，这取决于组织的大小、业务复杂程度和它的数据质量可能有欺骗性。系统开发者需要建立仪表盘来确定在一起工作的数据的语法和语义，检查在不同业务之间的数据是清洁和连续的。

注意，BI/ERP整合可能不是一个有限时间的项目，因为它在不停地发展。用户可能想要优化系统或者需要更多数据类型给更多的用户。

6.3.3 中间件

为了连接数据仓库，用户实施分析、发现信息或者数据可视化运行而使用的软件称为中间件。

通过将先前独立的应用和新系统整合，公司和组织正在开发企业范围内的BI。BI系统也必须连接到例如合作伙伴的系统或者进行公共的交换。BI用户需要通过多种工具与内部和外部应用进行交互，这些工具的特性和运行特点相差越来越大。在所有这些情况下，使用沟通协议和处于操作系统顶部的中间软件来达到下面的应用功能：

- 隐藏区域分布（例如一个应用通常是由许多分布在不同地点的内部相互联系的部分组成的）。
- 隐藏各种各样的硬件组件、操作系统、通信协议。
- 为系统开发者和集成商提供统一、标准、高水平的接口，使得应用能够轻松地组装、重用、移植和相互操作。
- 提供一套通用服务来运行不同目的的功能，这些功能能够避免重复性工作和为各种应用之间的合作提供便利。

中间的软件层就叫做中间件。中间件本质上提供给一个独立的程序作为不同客户和服务系统之间的接口。它的主要功能就是调节一个应用程序的不同部分或是不同应用程序之间的交互（更多的信息参见middleware.objectWeb.org和en.wikipedia.org/wiki/middleware）。

IBM是中间件软件提供商中的佼佼者。它为通信、政府、零售、银行、金融市场和其他行业提供了很多个性化的解决方案。IBM的中间件（ibm.com/middleware）能够帮助自动化系统、整合操作、人之间联系、软件开发。Oracle是另一个提供中间件和服务的公司。

Oracle Fusion Middleware（OFM，也称为融合中间件），由Oracle公司的软件产品组合组成。OFM包含多种服务：J2EE和开发工具、整合服务、商务智能、协同和内容管理。OFM依靠BPEL、SOAP、XML、JMS这些开放标准。

OFM提供开发软件、配置、面向服务架构的管理。它所被Oracle叫做“热插拔”的架构，

这种架构允许用户能够更好地利用现有的应用和其他软件供应商（IBM、Microsoft、SAP AG）提供的系统。更多的细节参见 en. wikipedia/oracle-fusion-middleware。

6.3 节复习题

1. 描述多层应用架构的基本元素。
2. 列举能够连接后端数据库和其他交易处理系统的管理支持系统应用的方法。
3. BI/ERP 整合的好处是什么？
4. 定义中间件，并描述它的特性。

6.4 面向需求的 BI

BI 实施在某种程度上说是耗费资金的，这可以在以上的讨论中看出。现在 BI 变得更加负担得起，甚至是对于小型和中性企业。在这部分我们介绍有关面向需求的 BI 的关键概念。

6.4.1 传统 BI 的限制

BI 解决方案最初的时候可能有负的投资回报率。原因包括高额的实施费用、每个用户的许可证费、维护费和咨询费用，在 BI 的生命周期中积累的很大的隐藏成本、不能实现最初的项目目标。传统的 BI 供应商很少能够提供统一的报告和分析方案从而允许管理人员实时地应对变化。除此之外，BI 解决方案还是很昂贵，它们有很长的实施周期，一般要 18 个月或者更长，这就要求在项目周期中投入很多有价值的 IT 资源。最后，无期合同使得客户不知道实施将在什么时候完成。

6.4.2 面向需求的选择

因此，公司特别是中小公司（Small to Medium-Sized Enterprise，SEM）正在转向面向需求的 BI 模型，它是比全面复杂的、昂贵的分析报告解决方案节约成本的替代方案。面向需求的计算，也被称做效用计算或软件即服务（Software as a Service，SaaS），将在下面进行描述。

软件作为服务（SaaS）是一个在需要的时候，可以使用的软件或者计算机资源的配置模型。它就像电或水一样。在你需要它的时候使用它，并且只为你使用的部分付费。因此，这个概念也被称做是效用计算。所以，用户不需要有自己的硬件、软件和其他设备，不需要维护它们。分发是由供应商完成的，供应商会给客户发放使用许可权，作为面向需求的服务来使用。SaaS 软件供应商可以在他们的 Web 服务器上拥有他们的应用或者将应用下载到客户的设备上，在使用结束或者需求合同失效的时候让软件失效。面向需求功能可能在一个公司内部分享许可权或者在许多的公司中通过第三方的应用服务提供商（Application Service Provider，ASP）分享许可权。所有的用户需要一个因特网接入和浏览器。付费是在实际使用的基础上进行的，或者通过固定的订阅费用（为给定数量的用户）。

6.4.3 关键特性和好处

SaaS 的特性和好处包括：

- 即使是在高峰时候（任何一个企业的需求），具有处理波动的能力。
- 减少服务器硬件和服务器使用的转换费用。
- 通过网络接入，可管理可使用的商用软件。
- 在一个集中地点而不是在每个顾客所在地管理活动，这就使得用户能够通过网络进行远程接入使用。

- 应用的分发通常都是一对多的模型（单实例多用户结构），而不是一对一的模型，包括架构、定价、合作和特性的管理。
- 集中的特性升级，这就使最终的用户避免了下载补丁和升级。
- 经常与更大的网络通信软件融合，或者作为 Mashup 的一部分或者嵌入一个平台。
- 起初的费用会比传统软件的许可费用低，但是会重复发生。所以从长远来看，作为一个服务，更像是许可软件的维护费用。
- 从长期看，总成本可能会高于或者低于、甚至是与购买软件或者付许可费相同。但是，在短期内使用 SaaS 费用要低。
- 顾客更多的功能要求，因为经常要求新的功能没有边际成本。
- 更快的功能发布，因为个体用户群均能在新功能中受益。
- 最被认可的实践体现，因为客户群会迫使软件出版者提供最好的应用。
- SaaS 应用的发展可能应用多种软件组件和架构。这些工具能够减少产品到达市场的时间并降低转换一个传统的软件产品的成本或者开发或者部署一个新的 SaaS 应用的费用。
- 就像其他的软件一样，软件即服务也能利用面向对象架构（Service Oriented Architecture，SOA）来使软件应用程序能够彼此通信。每个软件服务可能也是服务要求者，从其他的系统找入数据和功能。在开发 SaaS 时，企业资源规划（Enterprise Resource Planning，ERP）软件供应商使用 SOA。一个例子就是来自 SAP AG 公司的 SAP Business ByDesign。

面向需求的 BI 给了中小企业在今天快节奏的竞争市场中所需要的：一个简单好用、快速部署、合理价格的解决方案。面向需求模型为企业提供了低风险机会来使用 BI，不用投入巨大、昂贵的管理费用和有风险的项目。随着更多的中小企业获得了正的投资回报率和成功使用面向需求的 BI 的模型，我们能够看到即使是大公司也会采用这种模型。应用案例 6.1 中提供了一个这样的应用案例。

应用案例 6.1　零售商使用基于需求的 BI

Casual Male Retail Group 是一个专门经营体型高大的男士服装的供应商，它拥有 520 家零售直销店铺和电子商务运营，2009 年的销售额有 50 000 万美元。公司以前使用遗留应急报告应用来解决它的日常操作。但是，系统的报告功能非常弱，对于业务只有很少的可见性。例如，信息经理不知道他们正在销售了什么商品、每种型号的利润等实时信息。

公司使用提供不受欢迎报告的传统 BI（被称为例外情况报告功能）。用户到打印机前取得数百页的打印输出，有趣的是，因系统包括所有需要的信息。然而，用户不能以一种简单直觉的简单方式获得实时的业务销售和库存趋势的目录。当 Casual Male Retail Group 使用 Oco 公司提供的基于需求的 BI 之后，这种情况得到了改善。这种应用收集 Casual Male 的所有数据，为它的异地网点建立和维护一个数据仓库，生成实时响应的能够使用户一点鼠标就能获得所有信息的报告仪表盘。应用基于需求的 BI 系统，商品计划者和购买者能够轻松地通过仪表盘获得全部的目录数据。这就使得用户能够准确地知道任何时候在每个店铺哪种类型的服装正在销售。同样他们知道还有多少库存，哪里出现了短缺。

来源：Compiled from Wailgum, T., "Business Intelligence and On-Demand: The Perfect Marriage?" *CIO Magazine*, 2008, at www.cio.com/article/206551/Business_Intelligence_and_On_Demand_The_Perfect_Marriage (accessed 2010); advice.cio.com/thomas_wailgum/sorry_make_business_intelligence_suck_for_users (accessed 2010); and Casualmale.com (accessed 2010).

基于需求的 BI 的限制 以下是一些基于需求的 BI 的限制：

1. 将供应商的软件和公司软件整合可能是困难的。
2. 供应商可能破产，使得公司不能获得服务。
3. 为了更好满足用户需要，去改变主要软件是非常困难或是不可能的。
4. 升级可能成为一个问题。
5. 可能泄露战略信息给陌生人。

6.4 节复习题

1. 什么是基于需求的 BI？
2. 基于需求的 BI 的主要优点是什么？
3. 基于需求的 BI 的主要缺点是什么？

6.5 法律、隐私和道德问题

在 BI 实施中会遇到几个重要的法律、隐私和道德问题。我们提供具有代表性例子和一些资源。

6.5.1 法律问题

BI 的引入，特别是自动化推荐的使用可能产生与计算机系统相关的法律问题。例如，智能系统提供的建议的责任问题只是刚刚开始被人们关注。另外一个例子是将计算机分析的使用作为一种不公平的竞争手段（20 世纪 90 年代，曾经有个一个对航空订票系统使用计算机定价的知名的争论）。

除了解决一些 BI 系统没有预料到的可能造成危害的争论外，其他复杂的问题也会出现。例如，如果一个公司在使用了智能 BI 分析之后破产了，谁应该负责？没有充分地进行测试就将敏感数据委托给 BI，公司应该负责吗？审计和会计公司应该对没有提供合适的审计测试而负责吗？软件开发商是否应该负有连带责任？请考虑下面具体的问题：

- 当专业知识被编写进 BI 分析系统的时候，法庭中一个专家建议的价值是什么？
- 自动化 BI 提供的错误决策信息谁应该负责任？例如，经理接受电脑做出了错误诊断，并做出了对员工有负面影响的决策时将会发生什么？
- 当一个经理输入了错误的信息到 BI 系统中后，对于公司或者人员造成了很大的伤害时，将会发生什么？
- 谁拥有 BI 知识库中的知识？
- 管理者能够强迫经理使用 BI 系统吗？

以下是其他需要考虑的问题：

6.5.2 隐私

对于不同的人，隐私意味着不同的事情。通常，隐私是自己独处的权利，不受不合理人身攻击的权利。在许多国家，隐私已经是一个法律、道德和社会问题。隐私权在美国的各个州和其他国家已经有法规和法律来保护。隐私的定义可以被解释的非常宽泛。然而，下面的两条规则在过去的法庭决策中被遵守（1）隐私权不是绝对的。隐私在面对社会需要的时候是可以平衡的；（2）公众知情权高于个人的隐私权。这两条规则表明为什么在许多情况下，决定和实施隐私法规是非常困难的。隐私问题有自己的特点和政策。数据仓库环境中隐私和安全问题，参见 Elson and LeClerc（2005）。存在危险性的隐私区域将在下面进行讨论。

收集私人信息 实施 BI 可能需要员工个人的数据。在许多案例中，对数据进行收集、分类、

备案和手工连接不同数据源（公众或公司）的信息的复杂程度就是一种内置的保护，用来防止对于个人信息的滥用。侵犯个人隐私太昂贵、沉重和复杂了。因特网与大型数据库、数据仓库、社交网络，已经开创了连接和使用个人信息的全新维度。系统能够连接巨大数量数据的内在力量能够对社会和公司带来好处。例如，通过计算机将记录进行匹配，就能够减少或者消除诈骗、犯罪和公司的管理不善。然而，为了公司能够防止诈骗，个人在隐私损失方面应该付给一个什么价格？员工隐私信息可能有助于做出更好的决策，但是员工的隐私可能受到影响。对于顾客信息同样会产生相似的问题。

网络和信息收集 因特网提供了许多收集个人隐私信息的机会。以下是一些能够使用的方式：

- 通过阅读个人社交网络简介和帖子
- 在网络目录中查看个人姓名和身份
- 通过阅读个人邮件、博客和帖子中的讨论
- 通过窃听员工的有线和无线通信
- 通过监视员工
- 通过要求个人填写网络注册
- 当使用一个浏览器导航时，通过使用间谍软件记录个人行动

能够允许用户使用一个供应商的产品连接到不同服务的单点登录设备正在引起和Cookies一样的担忧。因特网服务（例如Yahoo、MSN）让用户永久地输入一个信息简介和密码，在不同的站点重复来使用服务。批评家说这样的服务创造了和Cookies同样的侵犯个人隐私的机会。

在BI分析、公司管理和法律和决策实施中使用数据仓库和挖掘技术，可能会引起人们对于隐私的担忧。这些由数据挖掘和商业分析的感知能力产生的担忧将必须在BI开发最开始之时进行解决。

移动用户隐私 许多用户不是很清楚私人信息正在通过移动个人数字移动助理（Personal Digital Assistant，PDA）或者手机被跟踪。例如，感知网络模型的建立，是利用移动电话公司从一个到另外一个电话塔跟踪用户手机得来的数据，或者利用GPS工具传输用户地点信息，或者利用PADS在Wi-Fi无线热点处传送的信息。这样的信息能够应用在BI分析中。感知网络从为公司在用户隐私方面要非常小心。

6.5.3 决策和支持中的道德问题

BI和计算机决策支持涉及多个道德问题。Chae et al.（2005）提供了道德问题形成和决策的全面概述，它提出了道德问题形成模型（如图6-3所示）。

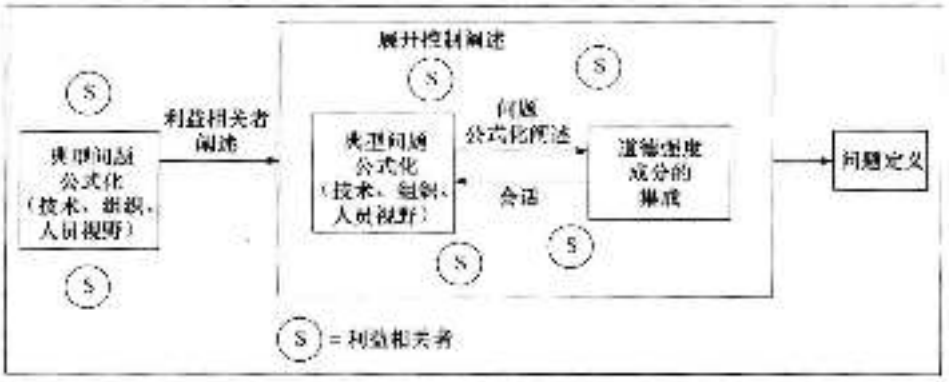


图6-3 道德问题形成模型

在BI实施中，比较有趣的、具有代表性的道德问题包括以下内容：

- 电子监视
- BI设计中的道德问题
- 个人隐私的侵犯
- 数据库所有权的使用
- 诸如知识和专业知识等知识产权的使用
- 数据、信息和知识的准确性
- 信息的可得性
- 公司计算机非工作目的的使用问题
- 多少决策委托给计算机

个人价值在BI和决策的道德问题中是主要的组成因素。因为它的多维性，所以BI中的道德问题的研究很复杂（Chae et al.，2005）。所以，开发框架描述道德进程和系统是非常有意义的。Mason et al.（1995）解释了技术和改革扩大了道德领域的范围，并讨论了一个道德模型，道德论证包括了4个基本关键问题：谁是代理？实际或预期采取什么行动？行动的结果是什么？结果是否公平？还是只对相关利益者公平？它们同样描述了道德等级论证，哪些道德判断或者行动是基于道德准则的，哪些是基于原则的，哪些是围绕道德理论的。更多决策中的道德问题参见Murali（2004）。

使用网络做与工作无关的事 员工试图使用电子邮件和电子商务网站等从事与工作无关的事。在一些公司中，这种使用大大超过了基于工作目的使用的比例（Anandarajan，2002）。问题有多个方面。例如，电子邮件能够被用来打扰其他的员工，这就给公司带来了威胁。它能够被用来实施非法赌博活动（例如，对于一场球赛的结果下注）。一些员工利用公司电子邮件做广告或是进行他们自己的业务。最后但并不是最不重要的就是在工作时间，员工花费在与工作无关的网点和在社交网站交流浪费的时间。

6.5 节复习题

1. 列举BI的一些合法问题。
2. 描述BI中对于隐私担忧问题。
3. 解释对于网络隐私的担忧。
4. 列举BI中的伦理问题。
5. 将BI与隐私相联系。

6.6 BI中的新兴话题：概述

BI正在变成应用的主要领域，在这个领域中，公司正在投入大量的资源并希望获得包括竞争优势在内的主要收益。所以，BI供应商和实施公司都在不断地努力提供顶尖技术。虽然预测未来哪个领域将会影响BI，或者BI将会影响哪些领域是个很大的挑战，但本章将讨论以下问题：

- 与BI相关的Web 2.0革命（6.7节）
- 与BI相关的在线社交网络（6.8节）
- 与BI相关的虚拟世界（6.9节）
- 社交网络和BI的结合（6.10节）
- RFID和BI（6.11节）
- 现实挖掘（6.12节）

商务智能的未来趋势

Gartner公司预测[1]（Gartner，2009）BI市场中的以下发展情况：

- 到2012年，业务单位会将40%的总预算投入BI中。
- 到2010年，20%的组织会将一个特定行业的分析应用作为他们BI组合的标准部分，这些应用是通过软件即服务提供的。
- 在2009年，协同决策将会作为新的产品种类出现，它将社交软件和BI平台特性进行整合。
- 到2012年，在商业过程分析上，1/3的应用将通过细粒应用聚合提供。
- 由于缺乏信息、流程和工具，到2012年超过35%的前5 000家跨国公司将会逐渐失去他们对市场和业务做出深入决策的能力。

6.7 Web 2.0 创新

Web 2.0是一个描述高级Web技术和应用的流行术语，包括博客、Wiki、RSS、mashup、用户产生内容和社交网络。Web 2.0的主要目标就是提高创造力、信息共享和合作。Web 2.0和传统Web的最重大区别就是因特网用户和其他用户之间、内容提供者和企业之间更多的合作。作为新兴技术、趋势和原则的总称，Web 2.0不仅仅改变了网络的内容，而且它也改变了它的工作方式。Web 2.0概念已经导致了基于Web的虚拟社区及其主要服务的革命，例如社交网络站点、视频分享网站等。许多人相信公司理解这些新的技术和应用，并尽早应用这些性能，代表其内部业务流程和市场的极大改进。最大的优势就是更好地与顾客、合作者、供应商和内部用户之间进行合作。

6.7.1 Web 2.0 的典型特征

以下是Web 2.0环境的典型特征：

- 对用户的知识进行收集的能力。用户贡献得越多，Web 2.0站点就变得越流行和有价值。
- 数据以一种新的或者从未使用过的方式使用。Web 2.0数据能够进行混合或者"混聚"，通常是一种Web 2.0界面以一种舞蹈俱乐部DJ混音的方式进行。
- Web 2.0依赖于用户产生和控制的数据。
- 轻量级编程技术和工具让每个人能够作为一个网站开发者。
- 软件升级周期的虚拟消失使得每种东西都是永久的测试版或者工作进度，并且允许将Web当做应用平台来快速地生成原型。
- 用户能够通过浏览器使用整个应用。
- 参与架构和数字民主鼓励用户在他们使用的时候为应用添加价值。
- 重点就是社交网络、社交计算和社交软件。
- 为信息分享和合作提供革命性的支持。Web 2.0中使用快速、持续的、新的商务模型。

Web 2.0其他重要的特征是它的动态性内容、丰富的用户经验、元数据、可扩展性、开源基础和自由（网络中立）。多数的Web 2.0应用拥有基于Ajax的丰富、交互、面向用户或者相近的框架。Ajax是一个用来创建交互Web应用程序的有效的网络开发技术。它的目的是通过与界面之后的服务器进行小数量的数据交换使得网页反应更加快捷，使用户在每次做出改变时，整个网页不需要重新载入。这就意味着增加网页的可交互性、下载速度和有用性。

[1] 这是指2009年的预测。——编辑注

6.7.2 Web 2.0 公司和新的商业模型

Web 2.0 主要的特征是创新网站的全球化传播和公司启动。当一个成功的理念在一个国家当使网站来部署时，其他的网站就会在全球出现。本书这部分将呈现这些网站。例如，将近 120 家公司在很多国家致力于提供Twitter-like服务。关于 Web 2.0 的一些非常好的资源可以参考 CIO 的《Executive Guide：Web 2.0》（参见 searchcio. techtarget. com/general/0，295582，sid19_gci1244339，00. html#glossary）。

从 Web 2.0 中新兴的一个新的商务模型是"群众的力量"的积累，这一商务模型的潜力是无限的。例如，Wikia（wikia. com）是一个专门从事基于社区开发的网络搜索。如果它们能够成功，谷歌将会有一个挑战者。

许多公司为 Web 2.0 提供技术，许多公司为社交网络提供基础设备和服务。从 2005 年到 2008 年间出现了许多初创公司。关于 25 个最热的 Web 2.0 公司和驱动它们最有力的趋势，参见 money. cnn. com/magazines/business2/business2_archive/2007/03/01/8401042/index. htm。

6.7 节复习题

1. 定义 Web 2.0。
2. 列举 Web 2.0 的主要特征。
3. 从 Web 2.0 出现的新的商业模型有哪些？

6.8 在线社交网络：基础和示例

社交网络建立在这样一种思想上：有一种人们如何相互认识和交流的结构。最基本的前提是社交网络给予人们分享、使世界变得更加开放和联系的能力。虽然社交网络通常是在诸如 MySpace、Facebook 这类社交网络上进行的，但它的一些特征也能在 Wikipedia、YouTube 中找到。

6.8.1 定义和基本信息

社交网络是一个人们能够建立自己的空间或者网页，能够在空间中写博客，传照片、视频或者音乐，分享思想，链接到他们认为有趣的网站。另外，社交网络的成员能够标记它们创作的内容，用他们自己选择的关键词来发布，这就使得这些内容是可以找到的。社交网络站点的大量使用是人类社会交往方面的一场革命。

社交网络的大小 社交网络正在快速地成长，有些已经拥有了超过 10 000 万的用户。一个典型的成功网站第一年用户的增长率是 40% ~ 50%，以后是 15% ~ 25%。包括用户数量信息的一些主要网站参见 en. wikipedia. org/wiki/List_of_social_networking_Websites。

社交网络分析软件 社交分析软件是用来识别、呈现、分析和可视化网络结点，或者利用各种类型的输入数据（相关的或是不相关的）和社交网络的数学模型模拟网络结点（例如，代理，组织或是知识）和边缘（关系）。存在许多输入输出文件格式。

网络分析工具使研究人员能够研究不同形式和大小的网络，从小型（家庭、项目团队）到大型。社交网络的可视化表现对于理解网络数据的传输和分析结果是非常重要和流行的。

一些能够实现这些显示的有代表性的表达工具是：

- 面向商务的社交网络工具，例如 InFlow 和 Netminer。
- 社交网络可视化或 SocNetV，它是基于 Linux 开源的包。

关于细节，参见 en. wikipedia. org/wiki/List_of_social_network_analysis_software。

社交网络与移动设备和网络密切相关。

6.8.2 移动社交网络

移动社交网络指的是成员相互之间使用手机或者其他移动设备进行联络的社交网络。像MySpace和Facebook这样的社交网络站点现在的趋势是提供移动服务。有些社交网络站点仅提供移动服务（如Brightkite和Fon11）。

有两种类型的社交移动网络。第一种是与无线供应商合作通过在手机浏览器上的默认起始网页来分布它们的社区。例如，用户能够通过美国电话电报公司（AT&T）的无线网络连接MySpace。第二种类型是没有这样的供应商关系（被称为"远程传送"），依靠它们自己的方法来吸引客户。第二种类型的代表包括Mocospace（mocospace.com）和Mobikade（mkade.com）。

Windows Live Spaces mobile能够在移动设备上使用有限的屏幕和缓慢的数据链接观看。它允许用户在它们的移动设备上浏览和添加照片，进入博客，直接发表评论。然而，它也引进了其他的特征来发展用户使用手持设备的体验。

关于更多Windows Live Spaces mobile的信息，参见mobile.spaces.live.com和en.wikipedia.org/wiki/Windows_Live_Spsces_Mobiles。

移动社交网络在日本、韩国、中国比在西方流行，主要是由于更好的移动网络和数据价格（在日本统一费率是非常普遍的）。Web 2.0服务和公司数量的激增，意味着很多基于移动电话和手持设备的社交网络的出现，将对这种网络的普及扩展到数百万不能经常和轻易接触到电脑的人们。

随着现行软件功能的实现，在移动社交网络内的交流不再仅仅是一对一的。交换纯文本信息。在许多情况下，它们正在朝向网络虚拟社区的复杂交流发展。

移动公司网络 许多公司已经开发（或完全资助）移动社交网络。例如，在2007年，为了吸引年轻人购买它的苏打水和其他产品，Coca-Cola公司开发了一个仅能通过手机访问的社交网络。

移动社区活动 在许多移动社交网络中，用户能够使用移动设备来创作它们的组合，交友，加入聊天室，创建聊天室，进行私人对话和分享照片、视频和博客。有些公司提供无线服务使它们的顾客能够建立自己的移动社区并给它们命名（如Sonopia公司的sonopia.com）。

通常与照片分享结合的移动视频分享是一个新的技术和社交网络方向。移动视频分享门户网站变得非常流行（参见myubo.com和myzenplanet.com）。许多社交网络站点提供移动服务。例如，MySpace与美国无线供应商有合作协议来支持它的MySpace移动服务。同样的，Facebook通过某些无线供应商实现在美国和加拿大都可访问。Bebo在英国和爱尔兰与O2无线合作。这些现象正是在建立多媒体网络社交站点竞争中的下一步措施。有些人认为这些合作与其说是推动社交网络站点，还不如说是在销售手机；然而社交网络很高兴能够获得剩余的关注。

6.8.3 主要的社交网络服务：Facebook和Orkut

既然我们已经熟悉了一些社交网络服务，下面在我们更仔细地研究一些非常流行的服务。

Facebook：网络效应 由哈佛学生Mark Zuckerberg于2004年创建的Facebook是全球第二大社交网络服务网站，在2009年3月拥有超过20 000万活跃用户。当Zuckerberg开始创建Facebook时，他有很强烈的社交抱负并想帮助人们通过网络相互联系。

Facebook快速扩展的主要原因是网络效应——更多的用户意味着更多的价值。当更多的用户参与社交空间时，更多的人就会被联系上。最初，Facebook是一个针对大学和高中学生的在线社交空间，能够自动连接在同一所学校的学生。然而，Facebook意识到它仅仅能够保有大学用户4年。2006年，Facebook对于年龄在13岁以上的拥有一个有效电子邮件地址的用户敞开了大门。

扩展到全球用户，使得 Facebook 与 MySpace 成为直接的竞争对手。

今天，Facebook 拥有许多支持照片、群组、事件、市场、发布主题和注释。Facebook 也拥有一个叫做“你可能认识的人”的应用，这一应用帮助用户与他们可能认识的人进行联系。更多的应用在持续地增加。Facebook 的一个独特的特点是新鲜事，能够使用户跟踪它们社交圈内朋友的活动。例如，当用户改变他的个人资料时这种更新会通知订阅这种功能的用户。用户还能够开发自己的应用或者使用其他用户开发的 Facebook 应用。

Orkut：开发社交网络站点的本质特征 Orkut 是土耳其 Google 程序员的心血结晶。Orkut 是谷歌针对 MySpace 和 Facebook 实施的本土策略。Orkut 使用了与其他网络社交站点相似的格式：使用各种多媒体应用显示他们期望的生活的各个方面的一个网页。

Orkut 的一个主要亮点是个人能力被提供给创建自己的群组和论坛（被称为社区）的人。谁能够加入和帖子如何编辑和控制仅仅由社区创建者管理。管理一个 Orkut 社区与管理一个自己的网站相似，给予创建者设计和控制内容的权利。Orkut 的用户应用 Web 2.0 工具获得体验，创造在线精通的浪潮，这无疑对于在线环境的发展是有益的。

Orkut 认识到是用户决定了它们所选社交网络站点内容。鉴于此，Orkut 采用了很多有趣的方法。首先增加了更多的语言，扩大到印地语、孟加拉语、泰米尔人语等，这增加了网站的知名度并改善了用户对网站的控制。第二，Orkut 为它们的用户在国家和宗教节日时提供有趣的应用。例如，它通过允许用户使用排灯节主题颜色和装饰重新设计他们的网站来祝贺印度用户排灯节（en. wikipedia. org/wiki/Diwali）快乐。

6.8.4 商业和企业社交网络的意义

虽然在公共社交网络中，广告和销售是电子商务活动的主要活动，但出现了应用于商务活动的商务方向网站，例如 LinkedIn 和企业内社交网络。

意识到了机会，许多软件供应商正在开发网络工具和应用来支持企业社交网络。例如，IBM Lotus 正在鼓励它的 5 000 多个从事 Notes/Domino、Sametime 和其他 Lotus 软件的方案提供商，以其他 LOTUS 软件提供者，添加 Lotus Connections 到它们的产品中，建立基于社交网络技术的应用。

下面是企业社交网络的代表区域和示例。

发现和招募员工 许多公共社交网络，特别是商务方向的网络能够使招聘和应聘更加的便利（Hoover，2007）。例如，招聘是 LinkedIn 的主要活动并且是站点发展的驱动力。为了获得竞争优势，公司必须在全球市场寻找人才，它们能够使用全球社交网络站点找到他。大的公司正在使用它们的内部社交网络来为空缺职位发现内部人才。应用案例 6.2 讲述了一个将 BI 和社交网络结合的应用。

应用案例 6.2 应用智能软件和社交网络来改善招聘流程

网络使广告和在线申请工作变成了一个很简单的过程。然而，有时简单化也会导致复杂化。现在对于一些大公司的挑战是如何以最优的成本管理在线招聘流程，因为在线广告正在吸引大量的应聘者。例如，Infosys 现在每年收到超过一百万个工作申请者来应聘 9 000 个职位。拥有如此多的应聘者听起来可能是好事，但是公司发现在它们需要的技能和特性与成千上万的应聘者之间的匹配度很低。这样，除了吸引了很多的应聘者外他们正在遭受缺乏好的应用程序的痛苦。另外，公司如何确定它们吸引了在某个领域的最好的人才？一些有趣的新发展正在改变公司面临的问题。

Trovix（a Monster.com 公司）给公司提供了一个基于它的获奖 HR 软件的服务，这个智能的服务帮助管理整个招聘流程。Trovix 说它的工具 Trovix 招募和 Trovix 智能搜索能够模仿人类决策和评估一个申请者的数量、深度、相关性、工作经验的近似程度和教育程度。软件以一定的顺序排列，满足广告职位最好的申请者。其他功能能够跟踪申请者、报告和通信。有些研究机构也正在使用这项服务，包括需要雇佣年有数千个招聘职位的牛津大学。Trend Micro 采用 Trovix 并且能够在 20 分钟之内屏蔽 700 个申请者和列出前 10 名申请者。精确度可能没有手工处理好，但是软件能够在更短的时间内屏蔽一些申请者。

通过一些社交网站，一些人性化的方法正在发挥作用，这些方法能够为公司的某个特定职位找到最好的人才。这类站点有 Jobster（jobster.com）和更加依赖社交网络方法的 LinkedIn（linkedin.com）。例如，Jobster 上的工作帖子能与其他工作网站、博客、用户群组、大学校友网站链接。鼓励社交网络的人去推荐适合某个特殊工作的人才，不论他们是否在积极地寻找新的工作。通过这种方法希望找到最好人才的公司使它的工作信息在更广的范围内发布并且能够从口头推荐和推举获得好处。例如，LinkedIn 提供给有期望的雇主一个超过 800 万人跨越 130 个行业的网络，这就意味着对于空缺职位更大的曝光和在更广的范围内寻找人才。例如，Jobster 网站也能够跟踪应聘者来自哪里，帮助公司实施更好的招聘策略和从他们寻找最好的员工的投资中获得更好的收益。

来源：Based on J. McKay, "Where Did Jobs Go? Look in Bangalore," Gazette.com, March 21, 2004, post-gazette.com/pg/04081/288539.stm (accessed July 2009) and "Trovix Makes Good at Stanford University; Premier Educational Institution Turns to Intelligent Search Provider for Recruiting Top Talent," March 8, 2006, trovix.com/about/press/050806.jsp (accessed July 2009).

管理活动和支持 这部分应用与支持从社交网络中收集信息进行分析的管理决策有关。一些典型的例子，包括识别关键执行者、定位专家并找到能够联系到他们的路径、征求复杂问题的想法和解决方案，寻找和分析可能的管理继承计划的候选者。例如，Deloitte Touche Tohmatsu 建立了一个社交网络来帮助人力资源经理裁员和重新组建小组。Hoover 已经建立了社交网站，该网站使用可视化路径技术来识别目标商业用户，建立关系和接触特定的用户。关于社交网站中使用数据挖掘的社交网络分析和挖掘的优势会议（2009 年 7 月希腊举行）也在讨论这个问题。

培训 多个公司使用企业社交网络和特殊的虚拟世界来进行培训。例如，Cisco 在产品培训和执行情况简介的第二个阶段正在使用它的虚拟社区。IBM 也第二个生命周期里运行管理和顾客交互培训会议。

知识管理和专家定位 这部分的应用包括知识发现、创造、维护、分享、转移和传播。Wagner 和 Bolloju（2005）曾详细地讨论过论坛、博客和对话知识管理的维基百科的角色的讨论。

考虑下面的关于知识管理和专家定位的社交网络范例了：

- Innocentive（innocentive.com），一个拥有超过 150 000 名致力于解决与科学相关问题（为了现金奖励）的研究人员的社交网站。
- Northwestern Mutual Life 创建了一个拥有超过 7 000 名金融代表来分享获得知识的内部社交网站（使用 Awareness.com 博客软件）
- Caterpillar 为员工建立了一个知识网络系统，它甚至将它的软件卖给了别的公司

公司也在建设退休员工合作社交网络来使退休员工之间以及和公司之间保持联系。这些人拥有大量的能够增加生产力和解决问题的知识。（如 SelectMinds 的校友联系）。接下来的几年（每个会议委员会）将有 6 400 万名退休人员，获得他们的知识是非常关键的。

加强合作 社交网站中的合作会发生在企业内部和外部。例如，来自不同部门的员工在一个虚拟团队中工作。在外部，与供应商、顾客和其他商务伙伴进行合作。合作通常是在论坛和其他形式的群组中利用维基百科（Wiki）和博客进行的。关于社交网站合作的细节，参见 Coleman and Levine（2008）。

在企业中使用博客和维基百科（Wiki） 这些工具的使用正在快速地传播。Jeffries（2008）报告了一项研究：在以下的应用中，71%的一流公司使用博客，64%使用 Wiki。

- 项目合作和交流（63%）
- 处理和程序文件（63%）
- 常见问题回答（FAQ）（61%）
- 电子学习和培训（46%）
- 新想法的论坛（41%）
- 公司专有的动态的词汇和术语（38%）
- 与顾客的合作（24%）

术语 Web 2.0 是 2004 年由 OReilly Media 用来指假定的第二代互联网服务产品，它能够让人们使用维基百科、博客、社交网络和 folksonomies 来创造和控制内容（OReilly，2005）。认识到 Web 2.0 的潜质，麻省理工大学数字商务中心（Brynjolfsson and McAffee，2007）和哈佛商学院（McAfee，2006 and Cross，2005）的研究者将 Web 2.0 的概念延伸到 Enterprise 2.0（在企业内部使用 Web 2.0），声称 Web 2.0 工具创造了一个平台，反映了知识自然真实的工作方式。这些工具有增进沟通和合作并帮助虚拟团队决策流程的潜质。

6.8 节复习题

1. 定义社交网络。
2. 列举社交网络站点的主要特征。
3. 描述社交网络的全球化特性。
4. 描述移动社交网络。
5. 识别 Facebook 的主要战略问题（参见 insidefacebook.com 和 facebook.com 的营销努力）。
6. Facebook 的早期成功归功于它与其成员的网络之间的密切联系。Facebook 是如何在不失去最初使网站流行的特点，不牺牲现在用户的前提下扩大市场的？

6.9 虚拟世界

虚拟世界已经以多种形式存在了很长时间，包括立体镜、电影院、模拟器、网络游戏、头盔展示。对于我们而言，虚拟世界就是电脑系统建立的虚拟世界，在这里用户有一种沉浸在其中的感觉。目的就是活动临场感和远距离参与感。现在流行的虚拟世界包括 Second Life（second life.com）、Google Lively（lively.com）和 EverQuest（everquest.com）。关于虚拟世界的技术、应用、社交和组织问题更好的综述能够在 Wikipedia（en.wikipedia.org/wiki/Virtual_world）找到。在这些虚拟世界中，树随风动，水流成溪，鸟在树上唧唧叫，卡车在街上飞驰。用户创造叫做头像的数字人物，能够交友、走路、在电脑产生的场景中与其他电脑产生的个体谈话。有些甚至经营全球业务。

真实世界的研究机构，从大学到商业，到政府组织都在越来越多地将虚拟世界融入到自己战略营销活动中。虚拟世界正在变成一个接触更广泛用户的重要渠道，以及“看着”顾客并以一种几年前还不可能的方式进行交流。诸如虚拟货币这样的概念允许参与者买卖像服装和培训这样的货物或者服务。虚拟世界提供丰富的广告形式，这种形式可以是身临其境的、主动的或者

被动的。除了文本外，广告可以是音频和视频，这些形式都是为了增加产品知识和顾客购买欲望。虽然，关于在线头像在市场营销中使用的研究还很少，但一些证据表明头像和虚拟形象能够积极地影响信任和在线购买意图，因为它们模拟顾客在真实世界中的购物体验（Stuart，2007）。然而，并不是所有的真实世界的特性都能够在虚拟世界中体验到，因为不是所有的人类感觉（例如味觉）能够数字化并能在计算机显示器上实现。

Second Life 可以作为一个有效的商务工具。经理们可以将 Second Life 应用在现实世界的决策中。John Brandon 在 Second Life 中的顶尖商务网站——计算机世界（2007）中的一篇文章写道：

> 使 IBM 有点甚至是非常有趣的事情就是在紧闭门后发生的事情。与顾客进行的常规的头脑风暴会议能够产生有趣的想法，例如，杂货商能够在 Second Life 中销售他们的商品，并将它们送到家中；燃气公司能够举行正规的不向公众开放的员工培训会议。

将 Second Life 应用在决策支持时需要精心计划。Dan Power 教授写了一个关于虚拟世界应用在决策方面的优势和劣势的专栏。参见技术前沿 6.1 中关于专栏的摘录。

虽然虚拟世界正在变成商务和顾客有趣的工具，但一些短期的技术和实际的考虑正在阻止它们被广泛地接受。例如，参与到这些虚拟世界中需要下载插件。然而，许多商业和政府组织阻止员工下载任何种类的软件到他们电脑上。这就限制了这些服务被员工使用，特别是 IT 员工。

尽管有一些限制，但虚拟世界顾客应用正在高速增长。本书的合著者 Sharda 研究虚拟世界在贸易展示中的应用。贸易展示是描述临时市场事件众多术语中的一个，它通常隔一段时期进行，在贸易展示中许多潜在买方和卖方为了更多地了解新产品和服务进行交流。贸易展示，例如书展、技术展和人力资源展览（招聘会）每年在世界范围内举行。

技术前沿 6.1 将 Second Life 作为决策支持工具

Second Life 作为决策支持工具具有以下优势：

1. **容易访问和低成本** 用户可以免费下载，人们不需要付会费就可以参与。客户仍然在发展，新的语音客户端被社区检验，所以软件需要每隔几周就重新下载更新的版本。

2. **有经验和忠诚的设计者/构建者** 访问 Second Life 展示了可能和奇迹仍在发生。Second Life 有很少的限制并为开发者提供了广泛和灵活的内容授权经验，目标、文档和可再次使用的脚本的数量巨大，设计者能够创造顾客头像、建筑物和产品。如果你能够制作简略的框架，那么一个好的建筑者能够快速建造一个原型。如果提供一个平面图和尺寸，那么一个建造者能够复制你的工厂，或者给予足够的时间，能够复制整个城市。

3. **交流驱动的决策支持工具和场地** 工具包括视频流、音频流、幻灯片和日志、会议管理工具、聊天记录甚至是头像名字标签。

4. **庞大忠诚的用户群** 在 Second Life 中，雇人为你的 Second Life 工作成本很低。支付用 Linden Dollars，你能够轻易地在超过 50 个国家中雇佣到员工。在 Second Life 中，像 Manpower 这样的公司很够帮助雇佣员工问题。Second Life 是公司走向全球化的一个简单方法。同样，许多用户有很高的计算机技能。

5. **印象管理和创造力增强** 头像可以是用户想要的任何形象。对于某种类型的决策系统来名是非常有优势的。Second Life 打破了创造性思想的壁垒并释放了想象力。一些人不愿意使用视频会议，因为对于他们如何出现有所担心。使用 Second Life，用户可以有意识地管理他们在会议、时间、活动中创造的想法。

6. **时间压缩** Second Life 中的一天就是 4 个小时。人们快速地交流并从一个场地传递到另外一个场地。Second Life 围绕时间进行操作。Second Life 的 7/24/365 性质能够加速行为和改变用户对于时间的看法。

7. **使用 RSS Feed，轻松地将真实生活数据整合** 将网络资源和 Second Life 的数据整合的可能性正在大大的增加。

8. **鼓励积极参与和经验学习** 人们体验 Second Life，这些体验影响真实生活。Second Life 的会议可以是享受和美好的。对于虚拟工厂的参观能够帮助人们理解在工厂建造时是什么样子。

Second Life 作为决策支持工具具有以下的劣势：

1. **学习时间和培训成本** 公司管理人员通常对 Second Life 不熟悉，学习曲线通常是 8 小时才会获得基本舒服水平。一个好的教练能够使得学习过程对于一个新接触的经理更加轻松

2. **分散注意力** Second Life 是一个拥有许多正在进行活动的虚拟空间，从购物到性，从海滩阳光到滑冰，从在浪漫的星光球室下的舞蹈到在 Second Life 犹太教堂中直播音乐会。一些分散注意力的事情是非常有趣的，但是员工可能在工作时间使用它们。同时，公司需要免责声明，HR 需要审新看待性骚扰的政策。

3. **恶作剧和垃圾邮件很普遍** 一个人会浪费太多的时间在 Second Life 中闲逛。许多人玩恶作剧、参与讨厌的活动，从极坏建筑到在教堂或者会议骚扰同事。存在许多类型的安全问题。

4. **技术问题存在** 一些技术问题包括反应慢、调整目标的滞后，需要在崩溃之后清理缓存和经常的软件升级。

5. **聊天是一个非常缓慢的通信工具** 新的语音客户端会加速 Second Life 中人之间的通信，但是聊天仍然需要，特别是在多语言通信中的自动翻译器。语音交流在 Second Life 的会议中是没有价值的。

6. **对于使用的抵抗** Second Life 不像其他管理人员所经历的其他事情，他们会抵触使用这项技术。非常容易的就将 Second Life 看做是游戏并忽视真实世界决策的可能性。

7. **沉迷** 一些人对于使用 Second Life 很着迷，并花费许多时间在系统中，忽略了真实生活的活动。公司的 HR 部门需要监督经常使用 Second Life 的员工的行为和态度。

来源：D. Power, "What Are the Advantages and Disadvantages of Using Second Life for Decision Support?" *DSS News*, Vol. 8, No. 15, July 29, 2007, dssresources. com/newsletters/195. php(accessed July 2009).

实体贸易展览允许最常用的交流形式，面对面地交易。传统贸易展览的不足之处包括地理限制，运营时间限制，较高的参与成本，需要通过获取展台战略位置获得最大的展示，从贸易展示中得到最大的利润。为了让更多的人看见自己的产品，现在许多参与者使用虚拟世界等新技术。一些信息技术工具能够模仿贸易展示的特殊活动。例如，如今非常普遍地使用通过网络传播的在线会议、陈述、演讲、研讨会。这些工具提供了从陈述者到观众的单向的沟通方式，但是这种方式能够使陈述者和观众之间能够交互地发出、接收和讨论信息。然而，网络研讨会不能像传统贸易展览那样传递给参展者相关的内容、相关利益者信息和数据。

虚拟世界技术对于通过组织虚拟事件来复制传统贸易展览的参与经验是有用的，这些虚拟事件能够通过扩大事件的影响力来吸引更多的参与者和参展者。虚拟的贸易展览在虚拟空间中举行，被看做是实体展览的延伸或者实体事件的场所。它复制实体事件的许多信息交换、通信和群体聚合方面。它的结构通常包括一个虚拟的展厅，具有特殊能力的用户通过许可进入展厅来观看虚拟贸易展览展示，或者建立虚拟展台来展示信息，就像它们在一个会议中心中举行的贸易展览会。虚拟贸易展览可能包括其他部分，例如虚拟网络会议，网络研讨会集合，或者其他的教育展览。参观者在进入展厅参观各种展台之前，填写一个在线注册表格来创建一个在线标志。虚拟展台一般很像真实世界贸易展览的展台，有桌子和用户能够轻松得到的展示。虚拟贸易展览能够成为国际贸易展览、业务媒人、采购洽谈会和产品发布会。这种经验同样适用于其他的应用，例如虚拟招聘会、虚拟福利博览会、员工在线网络、分销商展览会和风险投资展览会。虚拟世界和贸易展览之间协同效应的认知已经被许多虚拟贸易展览公司使用。其中一个就是 iTrade-Fair. com。图 6-4 是一个虚拟展台的例子。

贸易洽谈会的参加者来到一个特定的虚拟贸易展览网页。参加者首先访问一个虚拟的展览场地。在虚拟展览场地参加者能够选择一个虚拟展台，收集信息或者参加生动的交流和信息传播。通过聊天、Web 同步、传真和电子邮件等技术特点来实现通信。特殊的发言者或者客人能够通过视频专题网络直播来进行通信。参会者能够通过聊天室进行通信。虽然，这使事件参与者能够在同一时间、不同的地点交换信息，但是它没有像 Second Life 中的头像可视经验那样丰富的媒体体验。

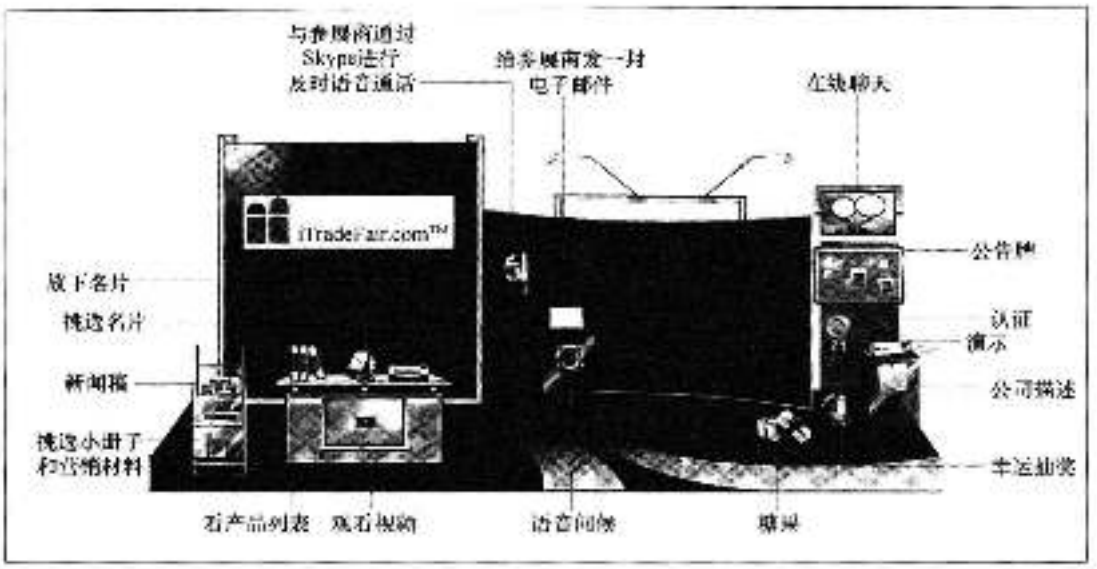


图6-4 虚拟展台范例

展出者参加贸易洽谈会的主要原因是获得新的指引和合同。在虚拟展示中，展览者能够获得实时的参加者的指引。一个包含每个注册参展者信息的参展者报告（与传统参展者名单相似）通过事件组织者提供给所有的展览者。展览者也能够获得访问它们虚拟展台的参观者详细的贸易报告。访问展台的参观者能够留下商务名片。所有留下数字商务名片的参加者的记录是可以得到的，这一报告包括所有参观者的名字、职位和相关的合同信息，对于某一产品和服务参观者是否需要更多的信息、公司的一般信息、工作机会。一个全面的“展台脚印报告”能够提供给所有参观过展览者虚拟展台的注册的参观者。这份报告提供了每一个特殊访问者对什么感兴趣的洞察。出于隐私和安全的考虑，所有的报告控制访问。但是，这样的报告为贸易展览的组织者和展示者提供了丰富的信息，并且能通过商务智能技术进行分析。

就像这节描述的那样，虚拟世界提供了一个以新方法提供决策支持的机会。在接下来的几年中，我们将看见决策支持能力更广泛的使用。另外，这样的环境（例如 iTradeFair. com 的虚拟贸易展览）产生大量关于用户活动和参与者在线活动的数据。这些大量的数据集能用 BI 技术进行分析，这样可以更好地理解顾客行为、定制产品/服务或者技术环境。

6.9 节复习题

1. 什么是虚拟世界？
2. 通过虚拟世界提供决策支持的优点和缺点。
3. 现实贸易展览中的哪些活动能够在虚拟世界中体验，哪些活动能够复制？
4. 针对用户在特定虚拟世界中的数据你会实施哪种类型的数据分析？

6.10 社交网络和 BI：协同决策

开头的引例表明了博客、Wiki 和 RSS 是如何辅助 BI 的。描述的系统展示了使用社交软件和 BI 的潜在好处。事实上，这种结合是非常有用的。的确，作为主要的 IT 咨询公司之一，Gartner 有限公司预测这种通过输入信息到企业决策中的整合为 BI 活动提供了机会。它们将这样的结合称为协同决策。

6.10.1 协同决策的崛起

在一份报告中，Gartner有限公司的研究者们（Schlege et al.，2009）描述：**协同决策（Collaborative Decision Making，CDM）**作为一种新的决策支持的类型，将社交软件和BI相结合。它能够通过直接将BI系统中的信息与使用社交网站产生的信息联系起来，大大地改善决策的质量。

这份报告的重要发现是：

- CDM是一类为非常规的、复杂的、需要人类反复交互提供支持的决策支持系统。
- 涉及价值、相关性、信用度和决策内容的特别标记能够不断地丰富决策过程和对决策有帮助的内容。
- 将BI输入到决策和可衡量的结果中，能够使组织更好地证明BI的商业价值。

在过去的10年中，尽管拥有空前的信息有效性，但公共和私有部门仍然遭受了多个不完善的决策。提供足量的信息并希望最终做出好的决定是不够的。大量的社会、文化和教育因素影响着个体和组织如何来提高他们的决策能力，这些因素需要在分析中进行考虑。CDM能够通过添加缺失的因素来改正低效的决策。

6.10.2 虚拟团队决策中的协同

由于经济的不景气造成的旅行限制迫使许多公司寻找工作、合作和决策的新方式。Gartner有限公司的研究者相信信息技术市场将通过创造一个使用社交软件培育CDM流程的系统来应对在虚拟团队中的合作需要。由顾客驱动的社交网络服务倡导的社交软件技术商务应用，例如Facebook、Myspace，运行良好。组织已经使用协同社交软件来了解同事在哪儿、他们正在做什么和想什么；发动他们召集紧急会议来解决问题。设计协同环境是上述趋势的自然进化，这种环境使决策者讨论问题、进行头脑风暴选择、评估它们的利弊、对一系列问题达成一致。添加社交软件因素（例如标签、推荐、评级、文件信息）丰富了协同环境并使它（和源于它的结果）更加有用。

CDM使BI系统将它模型化的信息和在合作环境中做出的决定紧密相关。BI系统过去明显地不能与商务过程相联系。结果了解BI的商业价值通常是困难的，即使是在最深刻的报告和分析中。另外，决策被认为是一种不能重复的非结构化过程，所以缺乏为决策者提供便利的工具。图6-5表明了CDM工具是如何支持决策过程的。

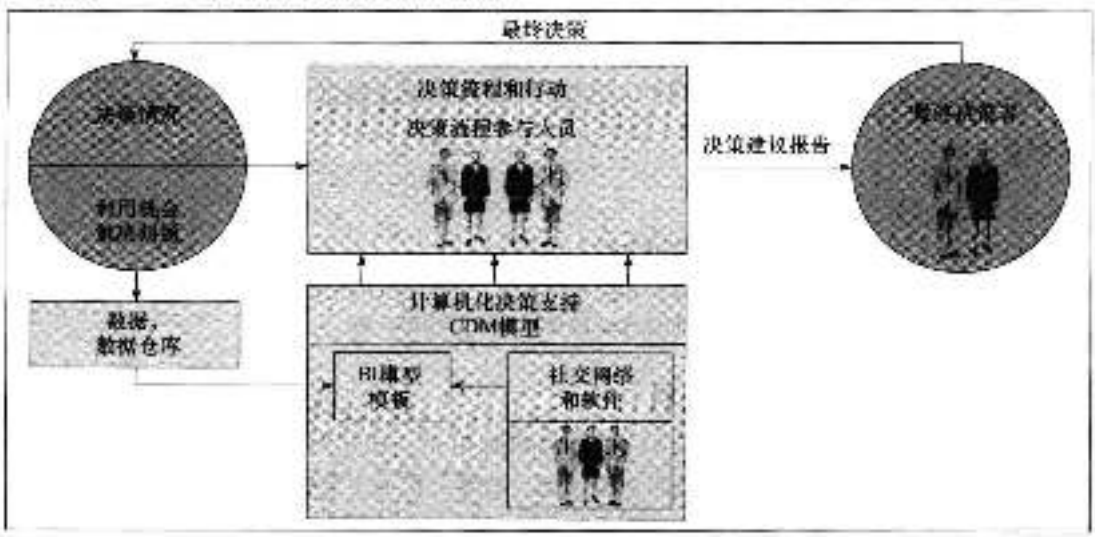


图6-5 协同决策框架

CDM是如何工作的 BI与社交软件结合提供了一种展示BI价值的方法，因为分析的洞察力和措施是与商业决定相联系的，并在社会环境下形成的框架。例如，在投资决策过程中，用户能够评价他们未来收益、花费者或者利率的假设，并能将这些预测的结果与衡量销售与利润的关键绩效指标进行对比。BI平台能够用合适的关键绩效指标的真实、临时结果来更新预测模型，帮助用户（关键绩效指标的参与者）超过关键的临界值，要求重新考虑决策。今天，决策中的协同更加具有战略性，决策涉及非常规的活动，例如头脑风暴发现、改革、创新和引导团队、学习和联系，能够改变商务活动。手工决策的结果会轻易地丢失，或者变成以奇闻的形式存在的公司民间传说的一部分。没有正规的决策审计、评估、闭环学习的过程。很明显，这是一个需要信息系统来为手工过程提供便利的领域，CDM能够成为一个理想的机制。

6.10 节复习题

1. 将BI和社交网络结合的理由是什么？
2. 为什么它被称做是协同决策？
3. 图6-5我们能够学到什么？
4. CDM的主要好处是什么？
5. 与社交软件的结合的特殊贡献是什么？
6. 解释CDM是如何工作的？

6.11 RFID和新的BI应用机会[8]

2003年6月，Wal-Mart要求它的前100个供应商在将运至得克萨斯地区的达拉斯商店的托盘和箱子上全部安装RFID标签，在这个命令前，Wal-Mart启动了一项50年的技术，这项技术以许多合适的区域为基础，被有限地使用（但是很成功）。自从此项声明发表后，RFID行业开始兴起，美国国防部不久就发布了自己的命令；Target、Albertson和Best Buy也迅速地跟进。起出的努力集中在零售供应链中的大型供应商（例如Procter & Gamble、Gillette、Kraft），但是现在扩散到小一些的零售供应商，Wal-Mart另外200个大型供应商在2006年1月开始运送标签产品。

RFID技术是指使用无线射频来识别物品。根本上，RFID是一系列自动识别技术的一种，它包括无处不在的条形码和磁条。从20世纪70年代中期开始，零售供应链（和其他领域）已经使用条形码作为自动识别的主要形式。RFID的潜在优势已经促使许多公司（由大型零售商如Wal-Mart、Target、Albertson带领）使用这项技术作为改善它们的供应链，以减少成本和增加销售。

RFID是如何工作的呢？在其最简单的形式中，一个RFID系统包括一个标签（贴在产品上以被识别）、一个阅读器、一个或多个与阅读器相连的天线和一个计算机（控制阅读器和储存数据）。现在，零售供应链主要兴趣是使用被动RFID标签。被动标签从电磁区域接收能量，电磁区域是由阅读器生成的，并在需要时反向散射信息。被动标签将能量保留在阅读器的电磁区域内。

相反，主动标签装有电池来为自己提供能量。因为主动标签有它们自己的能量源，所以它们不需要阅读器来给它们提供能量；它们能够触发数据传递过程。积极地来看，主动标签有更长的阅读范围，更好的准确率，更复杂的可重写信息存储，更强的处理能力（Moradpour Bhuptani, 2005）。消极地来看，由于电池，主动标签拥有有限的生命，比被动的标签尺寸更大，价格更

---

8 这部分改编自我们与阿肯色州大学及RFID研究中心负责人埃格里太·比尔博士一起合作的研究项目。

质。现在，多数零售商使用被动标签来设计和操作。主动标签经常在防御和军事系统中使用，然而它们也会出现在某些技术中，如 EZ Pass，在这些技术中标签与预付账户相连接，使得司机能够通过行驶过一个阅读器而不是停下来在一个交费岗来交费（U. S. Department of Commerce，2005）。

最常用 RFID 技术的数据表示方法是电子产品代码（Electronic Product Code，EPC），EPC 在许多行业被看做是通用产品代码（Universal Product Code，UPC）的下一代（通常由条形码表示）。与 UPC 相似，EPC 包含一系列能够识别产品类型和供应链上的生产商的数字。EPC 代码也包括额外的一组数字来识别商品。

现在多数的 RFID 标签包含 96 位数据，形式是系列化的全球贸易识别数字（Serialized Global Trade Identification Numbers，SGTIN），用来识别箱子或者系列货运包装箱代码（Serialized Shipping Container Code，SSCC）来识别托盘（虽然 SGTIN 能够被用来识别托盘）。标签数据标准的全部指导能在 EPCglobal 的网站上找到（epcglobalinc. org）。EPCglobal 是一个面向订购者的行业领导者的组织，致力于为 EPC 制定全球标准，用来支持 RFID 的使用。

图 6-6 描述的标签数据最简化的形式是一系列二进制位。这类二进制位能够被转化成 SGTIN 十进制。如图 6-6 所示，一个 SGTIN 是一个含有一系列数字的 UPC（UCC-14，应用于包装箱识别）。系列数字是当前使用的 14 位 UPC 和包含在一个 RFID 标签中的 SGTIN 两者的主要区别。应用 UPC，公司能够识别产品系列属于哪一个箱子，但是它们不能够区别一个箱子与另外一个箱子。使用 SGTIN，每个箱子是被唯一识别的。这就提供了箱子级别的可视化而不是产品系列识别的可视化。

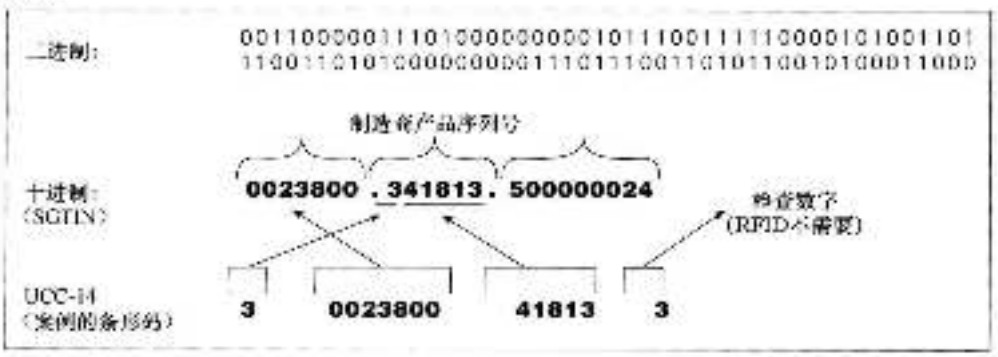


图 6-6 RFID 标签数据范例

RFID 产生的大量数据的应用之一是在供应链管理中（Delen et al.，2007）。多数的供应商在产品离开他们工厂的时候在产品上贴标签。一个产品由一个供应商流向零售分销中心（Distribution Center，DC），然后到零售货架，它可能经过许多 RFID 读取地点。当产品经过这些地点时，阅读器捕捉并记录箱子的标签数据。当产品被送到配送中心时，阅读器（由静态的阅读器和传送门户两边的天线产生）捕获托盘和箱子上的数据。作为一个展示示例，表 6-1 跟踪了实际的一箱产品（SGTIN：0023800. 341813. 500000024），从它到达配送中心到它在压碎机处结束它的生命的几个活动。这个产品的箱子在 8 月 4 号到达配送中心 123，在 8 月 9 号放在传送系统上，不久以后离开。（为了可读性，仅一个事件被显示）。它在离开 DC 之后的 12 个小时到达了 987 商店，然后立即去往销售大厅，在 5 小时之后由销售大厅返回，被放在商店辅助仓库直到第二天，它又一次去到销售大厅，45 分钟后返回，然后来到压碎机等待最终的处理。产品多是沿着描述的线路，但是在离开销售大厅和返回两个不同的场合时最终偏离了轨道。

我们能从表 6-1 中得到什么数据（RFID 数据的一个简单例子）？如果我们仔细地检查，数据

提供了很多可以深刻理解的地方。

表6-1 RFID数据

位置	EPC	日期/时间	读者
DC 123	0023800.341813.500000024	08-04-05 23：15	Inbound
DC 123	0023800.341813.500000024	08-09-05 7：54	Conveyor
DC 123	0023800.341813.500000024	08-09-05 8：23	Outbound
ST 987	0023800.341813.500000024	08-09-05 20：31	Inbound
ST 987	0023800.341813.500000024	08-09-05 20：54	Sales floor
ST 987	0023800.341813.500000024	08-10-05 1：10	Sales floor
ST 987	0023800.341813.500000024	08-10-05 1：12	Backroom
ST 987	0023800.341813.500000024	08-11-05 15：01	Sales floor
ST 987	0023800.341813.500000024	08-11-05 15：47	Sales floor
ST 987	0023800.341813.500000024	08-11-05 15：49	Box crusher

首先，知道移动的次数和时间对于确保一个产品的新鲜度，跟踪回访电话，并以一种及时的方式使产品到达商店（特别是对于时间敏感的产品）非常重要的。例如，考虑公司产品促销面临的情形。广告（国内、国外）通常是为了进行产品促销而发布的，产品的命运是在促销开始后的前几天中决定的。如果一个产品不是以一个及时的方式放在货架上，那么销售会受到损失。Gillette已经使用RFID来决定商店是否在它们的货架上存储有某项促销活动的某一商品。我们发现在一个促销开始之前，使用RFID将产品从商店辅助仓库移动到货架上的这些商店的销售额要比那些没有及时移动产品的商店的销售额高48%（Evans，2005）。RFID提供了所需要的数据和调查。

第二，数据提供了对于将货物从商店辅助仓库运至销售大厅这一过程的观察。在表6-1提供的例子中，我们可以发现产品移动到销售大厅2次。可能第一次它是被带出来的，它与货架不相符并被退回到商店辅助仓库。第二次它出来，它与货架相符。这一"不必要的箱子往返"产生了几个问题。将产品移到销售大厅并发生不必要的退回浪费了宝贵的人力资源，一个产品被处理的次数越多，它被损坏的概率就越高。为什么产品被2次送至销售大厅？如果产品直到8月11号（产品上架的那一天）才被需要，那么为什么它在8月10号被送到存储室？这就反映出了预测和更新系统的问题。或者可能一个员工在产品不需要的时候下达了一个手工订单。如果是这样，手工订单为什么会出现？它可能是产品在商店辅助仓库但是没有被发现或者没有被找到。员工不是去花时间找到它，而是手工发出了需要产品的订单。当产品在传送时，另一个员工在商店辅助仓库找到了产品并存放在货架上。当手工订单的产品到达时它可能不适合货架，（对于手工订单的产品）一个不需要的运送过程发生了。RFID是如何帮助解决这种情况的？当一个员工试图下手工订单时，系统能够检查现在是否有一个箱子存在商店辅助仓库中（由回收室中的阅读器决定）。如果箱子存在，就系统能够通过使用一个手持的RFID阅读器来帮助员工找到箱子。

第三，它提供了产品在供应链中移动的精确时间，以及一个箱子接一个箱子的每个关键阅读点的相隔时间。这种调查在以前是不可能的。交货时间通常是基于大量的产品系列在系统中的移动来估计的。存储水平的可视性在RFID之前是不可能的。这种可视性要求制定合适的指标来决定配送中心的行为。Delen et al.（2007）提出几个性能指标来捕获这种可视性。

公司使用RFID能够提高效率，并通过增加的流程变化来提高各种已存在过程的有效性。例如，早期证据表明RFID能够减少仓库接收产品的时间（Katz，2006）。RFID标签产品能够在一

个门户自动地读取而不是使用条形码一个一个地浏览每个产品箱。Gillette 发表报告说，使用 RFID 和源头标签战略能够使配送中心的托盘接收时间降低 5 秒到 20 秒（Katz，2006）。接收货物的流程没有明显的改变（也就是说，forklifts 像从前一样卸货）。唯一的改变是减少了人工浏览货物的需要。这样流程效率变得更高，流程也变得更加的有效。例如，Wal-Mart 发现通过使用 RFID 数据，产生更好的货物补给清单，能够使缺货率降低 26%（Hardgrave et al.，2006）。货物补给过程没有改变但是通过使用 RFID 有所改善。Wal-Mart 也减少了 10% 不必要的手工订单，这就使订货和预测系统更加有效（Sullivan，2005）。RFID 也能够用在收货时以减少错误的数量，这就改善了库存准确率，最终得到了更好的预测和补给。

RFID 数据还用在许多其他相关的应用中。例如，易腐商品为供应链管理带来了巨大的挑战，这是由于存在很多拥有不同易损性质的货物，在供应链中有不同的流通要求，大量的货物需要保管很长的距离。虽然食物是易腐品组合中出现频率最高的货物，但其他产品包括新摘的鲜花、药品、化妆品、汽车零件同样需要很严格的环境控制来保持它们的品质。由于需要处理的货物的数量非常大，所以可能出现的问题也在增加（Sahin et al.，2007）。例如，即使是非常小的腐败率的消除，也能够为供应链带来巨大的改善。所以，易腐品供应链管理优化是在市场竞争中最为重要的。

今天的易挥发易腐坏产品供应链的成功取决于产品可视化的水平和及时性。可视化能够为"我的货物在哪？"和"我的货物状况是什么？"等这些问题提供答案。许多公司已经开始应用 RFID 来管理易腐产品。考虑下面的例子：

- Samworth Brothers Distribution（英国；三明治、糕点等）已经在它的卡车里实施了实时温度监测（Swedberg，2006a）。
- Fresh Express 使用 RFID 来查看货物的流通情况和过期日期（Intel，2007）。
- Starbucks 使用温度跟踪流向零售网点的货物（Swedberg，2006b）。
- Sysco 使用 RFID 在不开门的情况检查装货物状况（Collins，2005）。
- 一个区域连锁餐饮（700 家餐厅）使用基于 RFID 的温度监测来决定牛肉饼、鸡蛋和洋葱等状况（Banker，2005）。
- TNT 使用 RFID 监测从新加坡到曼谷的货物温度情况（Bacheldor，2006）。

本章结尾的应用案例介绍了一个有趣的新兴的应用，这一应用包括了 BI 和 RFID 的创新使用。RFID 技术能够产生大量的数据，通过分析这些数据能够对公司的情况有更深入的了解，这是每种 BI 和决策支持存在的目的。下面部分说明了另外一种新兴的 BI 机会，是从大量收集的信息中产生的。

6.11 节复习题

1. 什么是 RFID？
2. RFID 读取或记录那种类型的数据？
3. 通过在配送中心使用 RFID，一个公司能够得到什么？
4. 上网收集一些 RFID 在健康护理、娱乐和运动方面的应用？

6.12 现实挖掘

就像 RFID 产生大量的数据通过商务智能进行更深的分析来帮助决策一样，另一个大量的数据源正在兴起，技术的发展使数据变得有意义。这种数据挖掘有个新名字——现实挖掘。Eagle 和 Pentland（2006）最早使用这个术语。MIT 的 Alex（Sandy）Pentland 和哥伦比亚大学的 Tony Jebara 拥有一个叫做感知网络（Sense Networks）的公司（sensenetworks.com），该公司致力于

开发现实挖掘应用。本节采用和包括的材料得到了感知网络公司许可。

许多由顾客和商务人士使用的设备都在不断地发出有关它们位置的信息。归功于连接网络的定位技术，如GPS、Wi-Fi和移动电话塔，小汽车、公交汽车、出租车、移动手机、照相机和个人导航设备都在传送它们的位置。许多用户和业务人员使用可知位置的设备来找到附近的服务，找到朋友和家人，导航，跟踪财产和宠物，问路，参加体育活动，游戏和娱乐。定位服务的增加导致产生了包含大量历史和实时的位置信息的数据库。当然，它是分散的并且单个是没有使用价值的。现实挖掘是在这样一个思想基础上实施的，这一思想是：数据集能够提供实时的聚集大量人群活动趋势的洞察。

通过分析和学习大规模行为方式，能够区分特定内容中的行为的分类，叫做"部落"（Eagle and Pentland，2006）。Macrosense是由SenseNetworks开发的应用平台，该平台使用由各种移动设备产生的数据，经过空间和基于时间的清洗之后，对这些大量数据源使用合适的聚类算法来对输入的信息进行分类，因为这些数据属于不同类型的顾客、客户等。这种方法使企业能够更好地理解它的客户的行为方式并且能够对于促销、定价等做出更好的决策。

感知网络公司现在采取这些技术来帮助用户找到有相同兴趣的人们。这一应用被叫做城市感官（Citysense）图6-7是旧金山某一区域的地图。在sensenetworks.com/citysense.php网站能够更好的看见，但是即使是基础相知的图形也能表明知道人们在某一特定时间将去哪里是有可能的。每一个点代表人们的出现，表示人们是如何分组和在城市中活动的方式。感知网络的核心分析平台Macrosense也能够分析大量的在Citysense中展示的信息，来对用户进行分组和识别部落。Macrosense能够通过提取某一地点和时间的部落分布的样本来识别这些部落在那，这就使得当一个用户在一个地点和时间时所处的位置来推断这意味着什么是可能的。例如，摇滚俱乐部和街舞俱乐部每一个都拥有一个截然不同的部落分布。当一个用户在夜间外出时，Macrosense从人们在这些地方所花费的时间中了解到他们喜欢的部落分布。感知网络公司说，在Citysense未来的版本中会包括部落，当用户访问其他城市时，他们将能够看到基于这种分布的推荐地点和这些地点活动的全面信息。

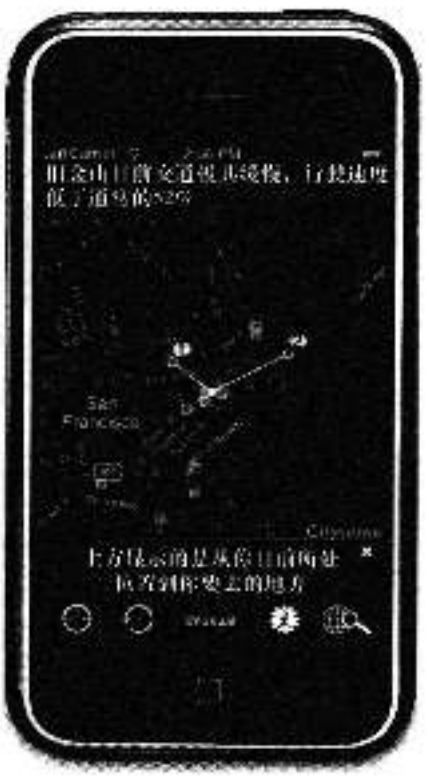


图6-7 旧金山城市感官范例

去摇滚俱乐部的用户会看到摇滚俱乐部的地点，经常去街舞俱乐部的用户将会看到街舞俱乐部的地点，两种都去的顾客能够看到所有的信息。这就回答了"像我一样的人现在都在那？"这样的问题，即使是在他们之前从没有去过的城市。通过使用部落来模仿真实世界，能够为每个顾客提供个性化的服务而不需要收集个人身份信息。

通过使用减少地点数据维度的算法，现实挖掘能够通过不同地点之间的活动来区分地点。从大量的高维度的地点信息中，这些算法能够显示趋势、含义和关系最终产生人类能够理解的表达方式。使用这样的信息可以自动地进行智能预测和找到重要的不同人和地点之间的匹配和相似性。Loecher et al.（2009）提供了它们算法的相关细节信息。基本上使用通过移动电话数据记录的活动信息来研究现实世界地点之间的行为联系。这也需要考虑时间，因为一组人可能在

早晨去一个地方上班，而另一个不同组的人可能在晚上经常去同一个地点是由于附近的俱乐部。由于数字背时的敏感性和人们经常去一个地点的类型（这些比在一个网络中的静态网页的数据更加的动态），描述现实世界的原始数据需要交错的维度。

通过由感知网络提供的资料，它给在一个城市中的每个地点赋予了487 500个维度。维度是建立在人们某一段时间进出某一场所的活动，和人们在这段时间前后访问的地点的基础上的。它们的"最低数量嵌入"算法将地点的维数和暂时的数据降到了二维，这里面保留超过90%的信息。这就使数据拥有了可视化，这允许人们更好地理解关键维度，提取一个城市中人们之间的关键关系，例如购物的人们、上下班的人们或者社交的人们。另外，它们也使用综合了人口信息、天气和其他变量的历史数据。一旦对于一个城市中的空间行为的理解是有效的，那么公司就能够利用持续更新的分类来更好地理解来自分散地点的顾客，发现大量顾客行为的趋势，预测服务和地点的需求。

使用这些技术一个关键的担忧是隐私的泄露。如果人能够跟踪手机的移动，那么顾客的隐私将是一个大的问题。但是，感知网络公司声称它仅需要收集大量流动的信息而不是个体能够识别身份的信息来将某一个人归入某一类。

访问感知网络公司的网站（sensenetworks.com）来查询这一领域的最新发展。技术正在飞速地发展。Baker（2009）和最近《Economist》（2009）中的一个故事强调了现实挖掘应用在商务管理中的一些可能的应用。例如一个叫做 Path Intelligence 的公司（pathintelligence.com）已经开发了一个叫做 FootPath 的系统，它能够查明人们在一个城市中是如何活动的，甚至是在一个商店中。所有的这些都是通过自动跟踪活动而不使用任何摄像记录移动的情况下实现的。这样的分析能够帮助产品或者公共交通站最好地布局。通过移动电话捕获和无线网络热点接入自动收集数据，为非侵入式市场研究数据的收集以及大量数据集微观分析提供了一种新的有趣的维度。

6.12 节复习题

1. 定义现实挖掘。
2. 在现实挖掘中使用哪种类型的数据？
3. 简明的描述数据是如何被用来产生用户简介的？
4. 如果你能够获得移动手机地点数据，你还能够想到哪些其他的应用？对能够提供地点的服务进行调查。

关键术语

Application Service Provider（ASP，应用服务提供商）
knowledge base 知识库
knowledge management 知识管理
problem solving 问题解决
utility computing 效用计算
reality mining 现实挖掘
brainstorming 头脑风暴
middleware 中间件
Radio Frequency IDentification（RFID，无线射频识别）
virtual community 虚拟社区
Collaborative Decision Making（CDM，协同决策）
mobile social networking 移动社交网络
robot 机器人
social network analysis software 社交网络分析软件
virtual team 虚拟团队
data integrity 数据整合
multidimensionality 多维性
virtual worlds 虚拟世界
functional integration 功能整合
physical integration 物理整合
Software as a Service（SaaS，软件即服务）
interactivity 交互性
privacy 隐私

讨论题

1. 一些人认为面向需求的 BI 将会成为 BI 的主要模型。请讨论该观点。
2. 讨论虚拟团队中的协同决策的利弊。
3. 区分在一项针对市场营销决策支持的 BI 项目实施中的实体和功能整合。
4. 讨论在 BI 应用中嵌入智能的价值。智能的 BI 是否会取代人类？
5. 讨论为什么将数据库和后端系统与 BI 连接是如此重要。
6. 在供应链管理中使用 RFID 的潜在好处和挑战。
7. 如果 RFID 数据是大量的侵犯权益基本的跟踪信息，那么你如何从这样的数据中挑选出有用的信息？你可以阅读 Delen et al.（2007）。
8. 使用虚拟世界来进行决策支持的好处与问题是什么？
9. 如果你有机会参加一场虚拟招聘会，哪些因素会激发和禁止你参加？
10. 基于位置追踪的简介（现实挖掘）是非常强大的但是同样存在隐私威胁，请评论。
11. Web 2.0 的主要特征是什么？Web 2.0 应用有哪些优势？
12. 讨论利用虚拟社区在网络上进行商务活动。
13. 维基百科（Wiki）是怎样用来为知识管理提供方便的？
14. 讨论移动设备和社交网络之间的关系？

练习

网络练习

1. BI 应用支持的实时决策被认为是主要的好处。上网查阅相关信息来识别例子和好处。访问 Teradata 大学网络和 Information-management. com 网站。
2. 查找一些关于 BI 实施的博客，并确定它们正在讨论的一些问题。
3. 访问 3 个 BI 供应商（例如，Cognos IBM 公司、Business objects、SAP 公司、Oracle）找到它们在面向需求的 BI 中的活动。写一份报告
4. 进入 Oracle. com。查找它为 BI 提供了哪些中间件，检查它的融合项目的状况
5. 进入 RFID 期刊网站（rfidjournal. com），列出至少 2 个与供应链管理有关的应用，和 2 个在下面领域的应用：健康护理、教育、娱乐和法律实施。
6. 进入 blog. itradefair. com。虚拟贸易展销会的一些有趣的应用是什么？
7. 进入 youtube. com，搜索关于云计算的视频，观看至少 2 个，总结你的发现。
8. 进入 sensenetworks. com 查看 Citysense 的应用和关于它的媒体报告。写一份关于你所学到的东西的报告。
9. 进入社交网络服务的网站（myspace. com-facebook. com）。建立一个网页，添加一个朋友录，使用鱼拨工具添加一个信息板。描写其他可有的能力。结交至少 5 个新朋友。
10. 进入 pandora. com。找到你如何才能创作和与朋友分享音乐。为什么它是一个 Web 2.0 应用。
11. 进入 smartmobs. com。访问博客链接，找到 3 个与 Web 2.0 相关的博客，并总结它们的主要特征。
12. 进入 mashable. com，查看最新的关于社交网络和网络战略的信息。写一份报告。
13. 进入 businessweek. com/print/magazine/content/06_18/b3982001. htm? chang = gl 阅读"My Virtual Life"，与幻灯片中的 7 个居民见面。准备一个表格来表述它获取利润的方式，所需的技能和能在 Second Life 中完成它的原因。
14. 识别 2 个虚拟世界（除了 Second Life 之外）。
15. 进入 secondlife. com，查找一个与 BI 和决策支持相关的内容。写一份关于它们提供什么的报告。
16. 进入 yedda. com，研究它的知识分享的方法。

小组作业

1. 通过拜访供应商（Cognus、IBM、Oracle）调查基于需求 BI 的状况，使用搜索引擎（bing、google、yahoo）、访问论坛（如，TSN、aspnews.com、utilitycomputing.ccom/forum/）。识别主要的担心和成就，准备一份报告和班级展示。
2. 准备一份关于社交软件能够支持 BI 应用的示例。提供参考文献和公司的名字。
3. 商品层次的 RFID 标签对于零售商店和用户都将是非常有用的。商店已经使用 RFID 来更精确的跟踪库存。顾客能够使用 RFID 来确定商品的位置甚至使用一个商店范围内的 GPS 来找到商品。然而，RFID 有一些潜在的隐私问题。分成 2 个小组进行辩论，一个小组支持使用 RFID，另外一个小组反对使用 RFID。
4. 上网搜索关于"虚拟贸易展览"。分成 2 个小组进行辩论，一组支持在商务应用中使用虚拟世界，另一组反对。
5. 基于跟踪定位的分组为个性化服务提供了可能但是也为隐私带来了挑战。将班级分成 2 个部分来辩论是否应该使用诸如 Citysense 这样的应用。
6. 每组进入以商务活动为特征的商务网站（例如 LinkedIn、Xing、Facebook、Second Life）。每个小组接下来将在 hellotxt.com 中注册找到与现在的商务活动相关的网站正在进行什么活动。写一份报告并做一次班级展示。
7. 使用 HelloTXT，登录网站将你的文本信息输入到网站中。然后，选择你想使用你的新身份信息更新的网站，HelloTXT 会做余下的工作，将你的新的身份信息传到你的各种各样的主页中。这是一种将你的各种各样的简介信息以一种尽可能实时更新的，非常集中化的保存方式，通过回答"你从事什么工作"这样的问题来更新你的 LinkedIn 身份。
8. 作为一个小组进入 secondlife.com 并新建一个头像。每一个组成员要访问一个特定的商业区域（虚拟房地产、教育活动、外交岛），确定头像与其他人的头像正在交流。写一份报告。
9. 进入 facebook.com 和 myspace.com 找出 10 个著名的公司是如何利用网站来进行商业活动。同样比较 2 个网站的功能性。
10. 多家医院正在引进或者考虑引进一个智能床位帮助系统。这一系统能够为医生和员工提供一个病人的药物记录数据库用来诊断和预测。这个系统从病人药物记录中提供任何需要的信息，在症状的基础上进行诊断，描述药品和其他治疗。这个系统包括一个专家系统和一个 DSS。这一系统是被用来减少一些人为错误和改善病人护理的。你是一个医院的管理人员并且对于系统为病人提供的好处非常的兴奋。然而，当你召开一个员工会议时，被要问到以下的问题：如果系统不能正常工作怎么办？当程序或规则中存在没有被察觉的错误时怎么办？系统一旦实施会担负起所有病人护理的责任因为医生依赖它。数据的丢失或一个程序的错误可能会导致灾害。例如，假如在数据库程序中存在错误，导致一个关键的信息从病人的记录中消失了。依赖系统的医生会在信息不完全的基础上开出药方。错误的结果会有生命的危险。另外一个可能性就是在知识数据库中的一些规则可能对于所有的患者来说并不是准确的。你是否愿意开一个系统吗？为什么会和为什么不会？
11. 阅读 Chae et al.（2005），总结文章中描述的所有道德问题并找出在每个领域中的例子。
12. 将全班分成两个部分：一部分相信 BI 将会取代商务分析专家，另一部分反对这一观点，进行讨论。
13. 识别有关管理决策的关键问题。上网查阅资料，加入聊天室，阅读网上的文章，根据你的发现准备一份报告。
14. 浏览美国律师协会的技术资源中心（abanet.org/tech/ltrc/techethics.html）和 nolo.com。在这里强调的主要的法律和社会担心与优势各是什么？它们是被如何处理的？
15. 访问一些关于健康护理的网站（例如 WebMD.com、who.int），找到与 MSS 和隐私相关的问题。写一份关于这些网站是如何改善健康护理的报告。
16. 访问 computerworld.com 并找到 5 个与 BI 和 MSS 相关的法律问题。

本章结尾应用案例

Coca-Cola 基于 RFID 的自动售货机作为商务智能的新类型

总部设在佐治亚州亚特兰大的 Coca-Cola 饮料公司，想要寻找一种增加销售的方法并找到一种更低成本的测试新产品的方法。在 2009 年的夏天，公司在加利福尼亚州、佐治亚州、宾夕法尼亚州的快餐厅中安装了新的装有 RFID 的自动售货机，并计划在全国推广自动售货机。新的自动售货机叫做 Freestyle，这种柜台装有 30 种口味的饮料，使顾客能够制造包括苏打水、果汁、茶等在内的 100 种不同的饮料。每种饮料仅仅需要几滴原料。顾客在自动售货机的 LCD 板上选择一个品牌和原料来使用自动售货机，这些是在 Windows CE 操作系统上进行的。

RFID 技术使 Coca-Cola 能够测试新的饮料口味和概念，观察顾客正在选择的口味和组合，识别区域喜好，跟踪他们正在饮用的数量。通过自动售货机能够使用多种组合的口味，公司能够看见哪种新的组合是最流行的，然后为其他的市场生产这种产品。这一过程为 Coca-Cola 节约了成本。之前公司会将新产品制作成瓶，并将它们送到各种各样的市场。有时在一两年后产品由于没有受到欢迎被取消。

RFID 技术帮助个体餐厅来预测什么时候购买新的原料，这样就增加了库存的准确率；并能帮助餐厅决定哪种口味最流行的，这样它们就知道存储哪种产品。个体的餐厅能够查看报告，报告的内容是关于饮料的消费量，饮料的消费量是由从 RFID 系统收集来的数据和通过使用 Coca-Cola 开发的电子商务客户端的产品记录生成的。这一技术甚至能够使餐厅看到一天中不同的时间哪种饮料是最受欢迎的。

RFID 技术是通过安放在每一种口味原料上的 RFID 的条码和安装在自动售货机内的 RFID 阅读器来工作的。每晚，记录的信息通过一个私有的 Verizon 无线网络被送至位于亚特兰大总部的 Coca-Cola 的 SAP 数据仓库系统。为移动设备服务的 Microsoft System Center Configuration Manager 在 Coca-Cola 总部运行，并通过无线网络控制自动售货机。另外 Coca-Cola 能够通过无线网络为自动售货机上新的组合发出指令，立刻在全国范围内关闭兴回的原料。

这一简短案例说明，当使用创新的思想发明出创新的应用时新技术有很大的潜能。在这章中所描述的大多数技术是新生的和没有被广泛使用的，所以有创造下一个"杀手"应用的机会。例如，RFID 的使用正在增加，每个使用的公司正在探索它在供应链、零售商店、生产，服务运营方面的用处。这个案例说明了将想法、网络、应用进行正确的结合，开发具有创造力的技术是可能的。这些技术能够以多种方式影响一个企业的运作。

本章结尾应用案例的问题

1. 在案例中 RFID 在减少库存方面发挥的作用？
2. 一个餐厅如何通过使用 RFID 进行探测原料管理获利的？
3. 一个顾客从新的自动售货机中会得到哪些好处？
4. 哪些问题会影响到这种自动售货机的广泛使用？

来源：Adapted from M. H. Weier, "Coke's RFID-Based Dispensers Redefine Business Intelligence," *Information Week*, June 6, 2009, informationweek.com/story/showArticle.jhtml?articleID=217701871 (accessed July 2009).

参考文献

Anandarajan, M. (2002, January). "Internet Abuse in the Workplace." *Communications of the ACM*.

Asif, S. (2009). "An Overview of Business Intelligence (**www.inforica.com/in/download/bipresentation.pdf**). Inforica Inc.,

Bacheldor, B. (2006, October). "TNT Uses RFID to Track Temperatures of Sensitive Goods." *RFID Journal*. Available at **www.rfidjournal.com/article/articleview/2726/1/1/**

Baker, S. (2009, February 26). "Mapping a New, Mobile Internet." *BusinessWeek* **businessweek.com/magazine/content/09_10/b4122042889229.htm** (accessed July 2009).

Banker, S. (2005, September). "Achieving Operational Excellence in the Cold Chain." *ARC Brief*. **sensitech.com/PDFs/coldchain_info/achieving_excellence_CC.pdf** (accessed September 2009).

Brandon, J. (2007, May 2). "The Top Eight Corporate Sites in Second Life." *Computerworld*, **computerworld.com/s/article/9018238/The_top_eight_corporate_sites_in_Second_Life** (accessed July 2009).

Brynjolfsson, E., and A. P. McAfee. (2007, Spring). "Beyond Enterprise." *MIT Sloan Management Review*, pp. 50–55.

Chae, B., D.D. Parchive, J.E. Courtney, and C.J. Cagle. (2005, August). "Incorporating An Ethical Perspective into Problem Formulation." *Decision Support Systems*, Vol. 40.

Coleman, D., and S. Levine. (2008). *Collaboration 2.0*. Cupertino, CA: Happy About Info.

Collins, J. (2005, June 13). "Sysco Gets Fresh with RFID." *RFID Journal*. Available at **www.rfidjournal.com/article/articleview/1652/**

Cross, R., J. Liedtka, and L. Weiss. (2005, March). "A Practical Guide to Social Networks." *Harvard Business Review*. Available at **ftp://ftp.cba.uri.edu/Classes/Beauvais/MBA540/Readings/Cross_et_al.pdf**

Delen, D., B. Hardgrave, and R. Sharda. (2007, September/October). "RFID for Better Supply-Chain Management Through Enhanced Information Visibility." *Production and Operations Management*, Vol. 16, No. 5, pp. 613–624.

Eagle, N., and A. Pentland. (2006). "Reality Mining: Sensing Complex Social Systems." *Personal and Ubiquitous Computing*, Vol. 10, No. 4, pp. 255–268.

Economist. (2009, June 6). "Sensors and Sensitivity." **economist.com/sciencetechnology/tq/displaystory.cfm?story_id=13725679** (accessed July 2009).

Eson, R.J. and R. LeClerc. (2005, Summer). "Security and Privacy Concerns in the Data Warehouse Environment." *Business Intelligence Journal*. Available at **http://findarticles.com/p/articles/mi_qa5525/is_200507/ai_n21377336/**

Evans, B. (2005). "Business Technology: Implementing RFID Is a Risk Worth Taking?" *InformationWeek's RFID Insights*. **informationweek.com/news/mobility/RFID/showArticle.jhtml?articleID=164302282** (accessed July 2009).

Gartner. (2009, January 15). "Gartner Reveals Five Business Intelligence Predictions for 2009 and Beyond." **gartner.com/it/page.jsp?id=850714** (accessed January 2010).

Good, R. (2005). "What is Web Conferencing?" **master-newmedia.org/reports/webconferencing/guide/what_is_web_conferencing.htm** (accessed July 2009).

Hardgrave, B. C., M. Waller, and R. Miller. (2006). "RFID Impact on Out of Stock: A Sales Velocity Analysis." White paper, RFID Research Center, Information Technology Research Institute, Sam M. Walton College of Business, University of Arkansas. **itrc.uark.edu** (accessed July 2009).

Intel. (2007). "Early RFID Adopters Seize the Initiative." White paper. **www.intel.com/cd/00/00/33/74/337466_337466.pdf** (accessed September 2009).

Jefferies, A. (2008, October 30). "Sales 2.0: Getting Social about Selling." *CRM Buyer*. Available at **www.crmbuyer.com/story/64968.html?wlc=1270580693**

Katz, J. (2006, February). "Reaching the ROI on RFID." *IndustryWeek*. **industryweek.com/ReadArticle.aspx?ArticleID=11346** (accessed July 2009).

Kisilevich, M., D. Rosenberg, and T. Jebara. (2009). "Citysense: Multiscale Space Time Clustering of GPS Points and Trajectories." *Joint Statistical Meeting 2009*.

Mason, R.O., F.M. Mason, and M.J. Culnan. (1995). *Ethics of Information Management*. Thousand Oaks, CA: Sage.

McAfee, A. P. (2006). "Enterprise 2.0: The Dawn of Emergent Collaboration." *MIT Sloan Management Review*, Vol. 47, No. 3, pp. 21–29.

McKay, J. (2004, March 21). "Where Did Jobs Go? Look in Bangalore." **Gazette.com, post-gazette.com/pg/04081/288539.stm** (accessed July 2009).

Mirchandani, S., and M. Bhupatiraju. (2005). "RFID Field Guide: Deploying Radio Frequency Identification Systems." New York: Sun Microsystems Press.

Morgenthal, J. P. (October, 2005). "Strategies for Successful Integration." *Optimize*, Vol. 48, pp. 87.

Morris, D. (2004, December 2). "Ethical Dilemmas in Decision Making." *Business Line*.

Neville J. (2007, January). "EMS: Adventures in X-treme Web 2.0." *Optimize*, Vol. 6, No. 1, p. 33.

O'Reilly, T. (2005, September 30). "What is Web 2.0?" **O'Reillynet.com: oreilly.com/web2/archive/what-is-web-20.html** (accessed July 2009).

Power, D. J. (2007, July 29). "What Are the Advantages and Disadvantages of Using Second Life for Decision Support?" *DSS News*, Vol. 8, No. 15, **dssresources.com/faq/index.php?action=article&id=138** (accessed Jan. 2010).

Sahin, F., M. Z. Rahal, Y. Dallery, and R. Vaillant. (2007). "Ensuring Supply Chain Safety Through Time Temperature Integrators." *The International Journal of Logistics Management*, Vol. 18, No. 1, pp. 102–124.

Schlegel, K., R. L. Sallam, T. Austin, and C. Howell. (2009). "The Rise of Collaborative Decision-making." *Special Research Note G00164718, April 9*. **mediaproducts.gartner.com/Microsoft/vol16/article8/article8.html** (accessed Jan 2010 – registration required).

Spangler, T. (2005, May). "Top Projects in 2005." *Baseline*, Issue 44, pp. 5–10.

Stern, B. (2007). "Virtual Worlds as a Medium for Advertising." *SIGMIS Database*, Vol. 38, No. 4, pp. 45–55.

Sullivan, L. (2005, October). "Wal-Mart RFID Trial Shows 18% Reduction in Product Stock-Outs." *Information Week*. **informationweek.com/story/showArticle.jhtml?articleID=172301246** (accessed Jan 2010).

Swedberg, C. (2006a, October). "Sainsbury Keeps Cool with RFID." **RFID Journal**. **rfidjournal.com/article/articleview/2733/** (accessed September 2009).

Swedberg, C. (2006b, December). "Starbucks Keeps Fresh with RFID." *RFID Journal*. **rfidjournal.com/article/articleview/2690/** (accessed September 2009).

Tse-Wei, L., P. Gabrielle, and I. Blake. (2007). "Marketing Strategies in Virtual Worlds." *SIGMIS Database*, Vol. 38, No. 4, pp. 77–90.

Turban, E., A. P. J. Wu, and T.-P. Liang. (2009). "The Potential Role of Web 2.0 Tools in Virtual Teams Decision Making and Negotiation: An Exploratory Study." Working paper. National Sun Yat-sen University, Kaohsiung, Taiwan.

U.S. Department of Commerce. (2005). "Radio Frequency Identification: Opportunities and Challenges in Implementation." Department of Commerce, pp. 1–38.

Vedapalli, N. K. (2009). "Critical Success Factors of BI Implementation". IT University of Copenhagen. **mit.itu.dk/ucs/pb/download/BI%20Thesis%20Report_New.pdf?file_id=871821.**

Wagner, C., and N. Bolloju. (2005). "Supporting Knowledge Management in Organizations with Conversational Technologies: Discussion Forums, Weblogs, and Wikis." *Journal of Database Management*, Vol. 16, No. 2.

Wailgum, T. (March 27, 2008). Business Intelligence and On-Demand: The Perfect Marriage? *CIO Magazine*. Available at **www.cio.com/article/206551/Business_Intelligence_and_On_Demand_The_Perfect_Marriage_**

Weier, M. H. (2009, June 6). "Coke's RFID Based Dispensers Redefine Business Intelligence," *Information Week*. **informationweek.com/story/showArticle.jhtml?articleID=217701971** (accessed July 2009).

术　语

active data warehousing（动态数据仓库）　参看 Real-time Data Warehousing（RDW）。

ad hoc query（特定查询）　在查询发出时刻没有察觉的查询。

adaptive resonance theory（自适应共振理论）　由 Stephen grossberg 创造的一个非监督的学习方法。自适应共振理论是一种神经网络架构，像大脑一样的一种无人监督状态。

algorithm（算法）　通过循序渐进的搜索来一步步提高，直到发现最好的解决方案。

arialytical models（分析模型）　加载数据用来分析的数学模型

analytical techniques（分析技术）　使用数学公式直接派生出优化方法或预测某个结果，主要用来解决结构化问题。

Application Service Provider（ASP，应用服务提供商）　为组织提供租赁软件的软件零售商。

Apriori algorithm（Apriori 算法）　通过使用递归的方法来识别频繁项集，发现关联规则的最通用算法。

area under the ROC curve（ROC 曲线下面积）　ROC 曲线下面积是一种在真阳性率为 Y 轴、假阳性率为 X 轴上绘制的二元分类模型图形评价技术。

artificial intelligence（人工智能）　计算机科学的分支，主要关注符号推理和解决问题。

Artificial Neural Network（ANN，人工神经网络）　尝试让计算机像人脑一样工作的计算机技术。机器拥有暂时记忆存储，处理模糊而可的信息。有时也简单称为神经网络。参看 neural computing。

association（关联）　一类数据挖掘算法，建立与给定的记录一起发生的项目之间的关系。

authoritative pages（权威网页）　由于被其他网页和指令链接被认为特别流行的网页。

automated decision support（自动决策支持）　对决策管理问题提供解决方法的一种基于规则的系统，也称为企业决策管理。

Automated Decision System（ADS，自动决策系统）　使用智能来对重复决策问题推荐解决方法的一种基于商业规则的系统。

axon（轴突）　一个生物神经元的突出延伸（也就是终端）。

backpropagation（反向传播）　神经计算中最知名的学习算法，通过对计算输出结果与期望的训练集输出结果进行比较来完成学习。

Balanced Scorecard（BSC，平衡记分卡）　一种绩效评价和管理方法，用来帮助将组织的财务，顾客，内部流程，学习与成长目标和任务转换成一套可行的措施。

best practices（最佳实践）　组织中解决问题的最好方法。这些通常储存在知识管理系统中的知识库中。

BI governance（商务智能治理）　优化商务智能的流程。

bootstrapping（拔靴法）　一种从原始数据中不断抽取固定数目的实例作为（替代）样本用于训练，数据集中剩余的数据用于测试的抽样方法。

brainstorming（头脑风暴）　人们产生想法的过程，通常有软件支持，（例如对问题形成可替代方案），也称为产生构思。

business analyst（业务分析员）　从事商业流程分析，并且得到信息技术支持的这类工作的人。

business analytics（业务分析）　对商业数据直接进行模型应用。业务分析包括使用决策分析工具，特别是模型，用来辅助决策者。最基本的是联机分析技术和决策支持系统，参看 Business Intelligence（BI）。

Business Intelligence（BI，商务智能）　决策支持的一个概念框架。它将架构、数据库（或数据仓库）、分析工具及应用程序结合起来。

Business Performance Management（BPM，业务绩效管理）　一种先进的绩效评价和分析方法，包含计划和战略。参看 Corporate Performance Management。

categorical data（分类数据） 代表多个类的标签，用于转换为特定的群体变量。

clickstream data（点击流量数据） 网络环境中产生的数据分析。

cloud computing（云计算） 可作为服务的信息技术架构（硬件、软件、应用、平台），通常作为虚拟资源。

clustering（聚类） 将数据库分到成员，每成员具有相似的性质。

Collaborative Decision Making（CDM，协同决策） 将商务智能和社交软件集成的一种新的决策方式。

complexity（复杂度） 经济优化模式，所需要的优化努力，或者随机属性，测试问题多么困难的一种度量方法。

confidence（置信度） 在关联规则里，发现规则的 RHS 出现在规则 LHS 的交易列表中的条件概率。

connection weight（连接权重） 神经网络模型中每个连接环节的权重，神经网络学习算法评价连接权重。

Corporate Performance Management（CPM，企业绩效管理） 一种先进的绩效评价和分析方法，包含计划和战略。参看 Business Performance Management。

corporate portal（公司门户网站） 进入公司网站的入口。公司门户网站可以通信、协调、接入公司信息。

corpus（语料库） 在语言学中，用于实施知识发现目的的大量结构化的文本（通常以电子形式存储和处理）。

CRISP-DM（跨行业数据挖掘过程标准） 一个实施数据挖掘的跨行业标准流程。按顺序一共六步，开始于对商务很好的理解及对数据挖掘的需求（也就是应用域），结束于满足特殊商务需求的解决方案部署。

Critical Success Factors（CSF，关键成功因素） 描述组织必须在它的市场空间中必须胜出的关键因素。

cube（立方体） 允许用户有组织地将一个立方体中任何属性的高度相互关联的数据子集（如商店、产品、顾客、供应商）与另一个立方体中的度量（如销售、利润、单位、年龄）相结合来制造各种二维的视角，这些可以显示在顾客的计算机屏幕上。

Customer Experience Management（CEM，客户经验管理） 通过检测网络应用事件及问题，跟踪和解决商务流程和应用障碍，报告现场性能和可用性，实现实时警报和控制，对被观察参观者行为深度处理的诊断，从而得出整体用户经验的报告。

dashboard（仪表盘） 让执行人员查看关键数据的可视化表示，可以让执行者看到每分钟的热点来探索情况。

data（数据） 自身没有意义的原始事实（例如名字、数字）。

data cube（数据立方体） 二维、三维，或者更高维对象，里面每个维度的数据代表一个列表的度量。

data integration（数据集成） 集成由三个主要过程组成：数据接入，数据联合，改变捕获。当这三个过程正确实施时，数据能够被访问，并且访问数据提取、转换和加载的数组，分析工具，以及数据仓库环境。

data integrity（数据完整性） 数据质量的一部分，在任何操作中（如转移、存储、检索）都保持数据总体的真实性。

data mart（数据集市） 数据仓库的一个部门的数据，只存储相关数据。

data mining（数据挖掘） 使用统计、数学、人工智能、机器学习技术从大型数据库中抽取，识别有用信息和后续知识。

data quality（数据质量） 数据的历史质量，包括真实性、精确性、完整性和相关性。

data visualization（数据可视化） 数据以及数据分析结果的图形化、动画或视频演示。

data warehouse（数据仓库） 相关数据按照标准形式特殊地组织起来，能够提供企业范围级别的干净数据的一个物理库。

Data Warehouse Administrator（DWA，数据仓库管理员） 负责数据仓库管理的人。

database（数据库） 被认为是简单存储概念的文件集，这样数据可以被更广泛的用户使用。

Database Management System（DBMS，数据库管理系统） 建立、升级、查询（例如管理）数据库的软件。

deception detection（欺诈行为侦查） 以声音、

文本以及人体语言形式识别欺诈行为（故意传播不正确的信息）的方法。

decision making（决策判定） 在可替换中选择的行为。

Decision Support System（DSS，决策支持系统） 支持管理决策流程的概念性框架，通常通过建模问题和定量模型来分析解决方案。

decision tree（决策树） 在假定风险下，作为关联的决策序列的图形表示。这一技术将特殊实体按照实体属性分为特殊类；根的后面是内部节点，每个节点（包括根）有一个问题标签，与每个节点相关的弧包括所有可能的反应。

dendrite（树突） 生物神经元部分，给细胞提供输入。

dependent data mart（非独立数据集市） 数据仓库直接生成的子集。

diagnosis control system（诊断控制系统） 一种控制论系统，它具有输入，将输入转化为输出的处理，与输出结果对比的一个标准或对照标准，一个反馈渠道，允许信息在输出和标准之间变化，实现通信和采取行动。

dimension table（维度表） 表示数据如何被分析的表。

dimensional modeling（维度建模） 支持高容错查询访问的检索系统。

discovery-driven data mining（发现驱动型数据挖掘） 一种数据挖掘的形式，用来发现数据的类型、关联和关系，目的是发现组织以前不知道，或者没有考虑的事实。

distance measure（距离测度） 在大多数聚类分析中，用来计算项目之间亲密程度的方法。流行的距离测量方法包括欧几米德距离（用一个规则刻度两点之间的通常距离）和曼哈顿距离（也称为两点之间直线距离，或出租车距离）。

DMAIC 一个闭环业务改进模型，包括以下几步：定义、测度、分析、改进和控制一个流程。

drill down（钻取） 信息的详细调查（例如，不仅发现整个销售量，还要发现按地区、产品、销售人员划分来销售量），发现详细的来源。

Enterprise Application Integration（EAI，企业应用集成） 提供将数据从源系统输入数据仓库工具的一种技术。

Enterprise Data Warehouse（EDW，企业数据仓库） 为了分析目标开发的企业级数据仓库。

enterprise decision management（企业决策管理） 请看 Automated Decision Support。

Enterprise Information Integration（EII，企业信息集成） 实现将来自关系数据库、网络服务、多维数据库等各种来源的实时数据集成的进化工具。

entropy（熵） 在数据集测量不确定性或随机性程度的一个度量标准。如果数据集中的所有数据只属于一类，那么这个数据集中就没有不确定性和随机性，因此熵就为零。

expert（专家） 在特殊或很窄的领域内做出熟练判断的人。

extraction（提取） 从多个来源中抽取数据，合成数据，提炼数据，决定哪些是相关的，然后以有效的集成方式将它们组织起来的流程。

Extraction，Transformation，and Load（ETL，提取、转换和加载） 由提取，转换和加载构成的数据仓库流程。提取就是从数据库读取数据，转换就是将提取的数据从以前的形式转换为需要的形式，这样就可以放入数据仓库或者另一个简单的数据库。加载就是将数据放入数据仓库。

functional integration（功能集成） 通过一个单一的，一致界面的简单系统提供不同的支持功能。

fuzzy logic（模糊逻辑） 一种逻辑上一致的推理方式，能够处理不确定性或者部分信息。模糊逻辑具有人类思考和专家系统的特性。

genetic algorithm（遗传算法） 以渐进方式学习的软件程序，与生物逻辑系统进化相近。

Geographical Information System（GIS，地理信息系统） 能够集成、编辑、分析、分享、显示地理相关信息的信息系统。

Gini index（基尼系数） 经济上用来度量人口多样性的度量标准。相同的概念可以用于作为一个分支，按特定属性/变量决定的结果确定一个特定的类的纯度。

Global Positioning System（GPS，全球定位系统） 使用卫星让用户能够相对精确地检测到连接设备上的（车或人）在地球上的位置的一种无线装置。

grain（粒度） 数据仓库中支持最高级别的详细定义。

Graphical User Interface（GUI，图形用户界面） 一个交互的、用户友好的界面，通过使用图标和类似的对象，用户能够控制和计算机的通信。

Heuristics（启发式） 应用领域里，判断规则组成的非正式和判断性的知识。启发式包含如何有效解决问题，如何制定解决复杂问题的计划步骤，如何提高性能的知识等。

hidden layer（隐含层） 具有三层或多层的人工神经网络的中间层。

hub（中转站） 一个或多个网页，提供链接到授权网页的集合。

Hyperlink-Induced Topic Search（HITS，超链接主题搜索） Web 挖掘中众所周知使用最广泛的引用算法，用于发现链接权威度和内容权威度。

hypothesis-driver data mining（假设驱动型数据挖掘） 一种数据挖掘的类型，从用户的一个命题开始，然后寻找命题的真实性。

independent data mart（独立数据集市） 为一个战略商务单位或部门设立的小数据仓库。

information overload（信息过载） 提供的数据太多，对个体来说处理和吸收任务很困难。

information（信息） 数据按照有意义的形式组织。

information gain（信息增度） ID3 算法中使用的分离机制。

intelligence（智能） 通常面向解决任务和问题的一定程度的推理和学习行为。

intelligent agent（智能代理） 一种嵌入计算机信息系统，使其更加聪明的基于知识或专家系统的系统。

interactivity（交互性） 软件代理的属性，允许不依赖于人的介入能够相互交互（通信和协同）。

interval data（区间数据） 在区间范围内可测量的变量。

inverse document frequency（逆文档频率法） 一种在术语文档矩阵中通用而且非常有用的目录转换，表明了词的特殊性（文档频率）以及它们整体发生的频率（术语频率）。

Key Performance Indicator（KPI，关键绩效指标） 面向战略和目标的绩效考核。

k-fold cross-validation（k 折交叉确认） 一种流行的用于预测模型的真实度评价技术。将完整的数据集大体按相同尺寸随机分成 k 份，分类模型要重复 k 次来训练和测试数据。每次选取 1 个作为测试数据，其余的作为训练数据。交叉检验评价模型的整体真实性是将每次实验的真实性测试进行平均的最后值。

knowledge（知识） 从教育或经验中理解，认知或精确获得的，任何可以学习、察觉、发现、推论、理解的；对信息使用的能力。在知识管理系统中，知识是采取行动的信息。

knowledge base（知识库） 事实、规则，具有模式的程序的集合。知识库是有关特殊兴趣领域的所有信息和知识的集合体。

Knowledge Discovery in Databases（KDD 数据库，知识发现） 执行规则归纳或相关程序，从大量数据库中建立知识的机器学习过程。

knowledge management（知识管理） 组织中专业知识的动态管理。包括收集、分类和传播知识。

Kohonen's self-organizing feature map（Kohonen 的自组织特征映射） 一种典型的机器学习神经网络模型。

Lean Manufacturing（精益生产） 消除流程中浪费或没有增值特性的生产方法。

learning（学习） 通过利用已经知道的来获得新知识的自我提高过程。

learning algorithm（学习算法） 用于人工神经网络的训练过程。

link analysis（链接分析） 许多有趣对象之间的链接被自动发现，例如学术刊物的作者群体之间的网页链接和引用关系。

machine learning（机器学习） 计算机从经验中学习的过程（例如，利用程序跑历史案例中学习）。

management science（管理科学） 应用科学方法和数学模型分析和解决管理决策情况（例如问题、机会），也称为运营研究。

metadata（元数据） 关于数据的数据。在数据仓库中，元数据描述数据仓库及其使用方式

的内容。

Microsoft enterprise consortium（微软企业财团） 从全世界范围接入 Microsoft SQL Server2008 软件套装，用于学术目的——教学与研究。

middleware（中间件） 连接不同计算机语言和平台应用模块的软件。

mobile social networking（移动社交网络） 成员之间使用手机或其他移动设备交谈和联系。

multidimensional analysis（多维分析） 包含多个维度数据分析的建模方法。

multidimensional database（多维数据库） 数据特殊地组织起来支持简单和快速多维分析的数据库。

Multidimensional OLAP（MOLAP，多维联机分析处理） 通过特殊的多维数据库（或数据存储）实施联机分析处理，集前将交易总结为多维视图。

multidimensionality（多维性） 从多个维度组织、呈现、分析数据，例如按地区、按产品、按销售员，按时间划分的销售额（四个维度）。

Multi-Layered Perceptron（MLP，多层感知） 人工神经网络分层结构中多个隐含层可以放置在输入层和输出层。

Natural Language Processing（NLP，自然语言处理） 使用自然语言处理器与计算机系统接口。

neural computing（神经计算） 一种实验计算机设计，目的是以一种人脑功能建模操作方式建立智能计算机。

neural network（神经网络） 请看 Artificial Neural Network。

neuron（神经元） 生物学习或人工神经网络的一个细胞（也就是处理元素）。

nominal data（名称数据） 一种包含为对象标签简单代码测量的数据类型，它是不用测量的。例如，婚姻状况变量能够通常分为：（1）单身；（2）已婚；（3）离婚。

numeric data（数值数据） 代表特殊变量数值的一类数据。大量的数字数值变量包括：年龄、孩子数目、家庭年休收入（以美元计算），旅行距离（英里），温度（华氏度）。

Online Analytical Processing（OLAP，在线分析处理） 一种信息系统，能够让用户查询系统，实施分析等，结果会在几分钟内产生。

Online Transaction Processing（OLTP，在线交易处理） 主要用于捕获和存储日常商务功能数据的交易系统。

oper mart（操作集市） 一种可操作数据集市；一个操作集市是组织中单一部门或功能区域使用的小规模数据集市。

Operational Data Store（ODS，业务数据存储） 一种数据库类型，通常用于数据仓库的一个过渡区，特别是客户信息文件。

optimization（优化） 辨别出可能是最好的问题解决方案的流程。

ordinal data（序数数据） 包含代表等级顺序，作为标签分配给对象和事件的代码。例如，信用分数变量可以被体分为几类：（1）低；（2）中；（3）高。

parallel processing（并行处理） 一种先进的计算机处理技术，允许计算机立刻并行地完成多种处理。

part-of-speech tagging（词性标注） 基于字在使用中的的定义、上下文、对文本中的词（如名词、动词、形容词、副词等）标记成为演讲中特殊的部分。

pattern recognition（模式识别） 一种将外部类型与计算机存储记忆中的类型匹配的技术（也就是说，将数据按已决定的类别进行分类）。类型识别用于推理机、图形处理、神经计算和语音识别。

performance measurement systems（绩效考核系统） 将业务指标与定期反馈报告结合起来显示围绕目标进步的系统方法。

physical integration（物理集成） 将多个系统无缝集成为一个功能系统。

polysemes（多义词） 也称为同音异义词，它们是语法相同（也就是拼写一样），但是意思不同的词，例如 bow 可以翻译为“弯腰”，或者是“船的前面”、“别箭的武器”、“一种系丝带”。

prediction（预测） 讲述未来的行为。

predictive analysis（预测分析） 利用工具帮助决定事件或状况发生的可能的未来输出。

predictive analytics（预测分析学） 一种用于预测的商业分析方法（例如需求、问题、机

会），而不是用于简单的报告数据。

privacy（隐私） 通常是指免处权和免受不合理的个人侵入。信息隐私是决定一个人的信息在什么时候，可以以什么程度传递给其他人的权利。

problem solving（问题解决） 一个人从最初的状态开始着手一个过程，通过问题空间的搜索，以确定一个预期的目标。

Processing Element（PE，处理单元） 神经网络中的一个神经元。

prototyping（原型） 在系统开发中，在很短的时间内构建按比例缩小的系统或者系统的一部分，经过多次迭代后进行改进的一种策略。

RapidMiner 一种流行的、开源的、免费的数据挖掘软件套件，采用了图形用户增强界面，具有大量的算法和一系列数据可视化特点。

ratio data（比率数据） 解释连续的数据差异和比率。比率规模的显著特点是拥有非任意零比率。

Real-time Data Warehousing（RDW，实时数据仓库） 加载数据，并通过数据仓库提供数据，使其可用的过程。

reality mining（现实挖掘） 基于本地数据的数据挖掘。

regression（回归） 一种用于真实世界预测问题的数据挖掘模型。它的预测值（也就是说，输出变量或者因变量）是数字（如，预测明天的天气是68华氏度）。

relational database（关系数据库） 数据库的记录有组织地形成表，这些表可以被关系代数或者关系演算进行处理。

Relational OLAP（ROLAP） 关系型联机分析处理。

result（outcome）variable（结果变量） 表达决策结果的变量（例如，关心的利润），通常是一个决策问题的目标。

RFID（无线射频识别技术） 利用射频波来识别物体的一种通用技术。

risk（风险） 概率或随机决定的情况。

robot（机器人） 具有没有人的干预可以完成手动功能能力的机器。

SAS Enterprise Miner 由SAS研究所开发的一款全面的商业数据挖掘软件。

scenario（场景） 关于一个特定系统在特定的时间运作环境的假设和配置的声明。

scorecard（记分卡） 一个可视化显示，通过图表显示战略、战术和任务目标。

search engine（搜索引擎） 发现并列出符合一些用户选择标准的网址或网页（通过统一资源定位符URLS设计）。

SEMMA SAS研究所提出的数据挖掘项目的替代流程，是抽样、探索、修正、建模和分析的缩写。

sensitivity analysis（敏感分析） 一个或多个输入变量对提出的解决方案影响的研究。

sentiment analysis（语义分析） 使用大量的文本数据来对特定商品或服务喜欢还是不喜欢观点进行探测的一种技术。

sequence discovery（序列发现） 随着时间推移进行关联辨别。

sequence mining（序列挖掘） 一种发现方法模式，事物之间的关系是根据它们出现的顺序来考察的，以此来辨别时间推移关联。

sigmoid function（S型函数） 从0到1的S型转换功能。

simple split（简单拆分） 数据被分割为两个相互排斥的子集，称为训练集和测试集。通常是将2/3的数据设计为训练集，1/3的数据设计为测试集。

Singular Value Decomposition（SVD，奇异值分解） 与主成分分析密切相关，它将输入矩阵（输入文档的数量和抽取术语的数量）的整体维度减少到低维，每个连续维度代表最大程度的（文字与文档的）可变性。

Six sigma（六西格玛） 一种绩效管理方法，目的是在业务流程中每百万缺陷机会减少到零。

snowflake schema（雪花架构） 雪花架构是多维数据库中表的逻辑关系，其实体关系图表现为雪花状。

Social Network Analysis（SNA，社交网络分析） 人、团体、计算机、其他信息和知识处理实体之间的关系和信息流的映射和测量。网络节点是人和团体，连接显示了节点之间的关系和流动。社交网络分析提供了关系的可视化和数学分析。

software agent（软件代理） 坚持完成（由所有

人）设计的任务的一款自主软件。

Software as a Service（SaaS，软件即服务） 软件是出租的而不是卖的。

speech（voice）recognition（语音识别） 人工智能研究的一个领域，尝试允许计算机识别人的语言字句。

SPSS PASW Modeler 由 SPSS（以前的 Clementine）开发的一款非常流行的商业化的，全面的数据、文本、Web 挖掘软件套件。

star schema（星形架构） 最常用的和最简单的三维数据风格。

stemming（词根） 为了在文本挖掘项目中更好地表现它们，减少单词直到它们的根形式。

stop words（无用词） 被过滤掉的自然语言数据处理之前或之后的词。

story（故事） 具有丰富信息和情节的实例。教训通常是从那些案例库中的案例中提炼出来的。

strategic goal（战略目标） 在指定时间内容易量化的目标。

strategic objective（战略目的） 描述组织目标方向的一个广泛的声明或行动。

strategic theme（战略主题） 与战略目标相关的一些集合体，用来建华战略地图的结构。

strategic vision（战略愿景） 关于组织在未来看起来会是怎样的图画或心理意像。

strategy map（战略地图） 从四个平衡记分卡的四个视角体现组织的关键目标之间关系的一个可视化显示。

Structured Query Lanaguage（SQL，结构化查询语言） 关系型数据库的数据定义和管理语言。SQL 前身是关系型数据库管理系统。

stummation function（求和函数） 添加到一个特定的神经元的输入机制。

supervised learning（监督学习） 一种人工神经网络的训练方法，样本案例作为网络的输入，为了减少输出的错误，权重被调整到最小。

support（支持） 测试产品或服务多久一起出现在相同的交易中，主关键词，数据集中在特殊规则下，包含所有产品和服务交易的百分比。

Support Vector Machine（SVM，支持向量机） 广义线性模型，从而实现输入功能的线性组合价值为基础的分类或回归的决定。

synapse（突触） 在神经网络处理单元之间的连接（权重）。

system architecture（系统架构） 系统的逻辑和物理设计。

Term-Document Matrix（TDM，文献术语矩阵） 创建数字化和组织化的文献（语料库）的频率矩阵，其中，列代表术语，行代表各个文档。

text ming（文本挖掘） 数据挖掘在非结构化和很少结构化文本文件中的应用。它可以从非结构化文本中产生有意义的数字指标，然后用各种数据挖掘算法处理那些指标。

tokenizing（标记处理） 根据它表现的功能对一块文本（表征）分类。

transformation（transfer）functions（转换函数） 在一个神经网络中，总结和转换在一个神经元出发之前的输入，显示了内部激活水平和神经元输出之间的关系。

trend analysis（趋势分析） 收集信息，并尝试发现信息的类型和趋势。

unsupervised learning（无监督学习） 这是自组织训练人工神经网络的一种方法，只有输入刺激能够在网络中显示。

user interface（用户界面） 一个计算家系统的组成部分，允许系统和用户之间进行双向沟通。

utility（on-demand）computing（（面向需求的）效用计算） 无限的计算能力和存储容量，就如电力、水和电话服务，账单基于每次使用的基础上付费，可按需获得，使用并按任何应用量新分配。

virtual（Internet）community（虚拟（因特网）社团） 具有相似兴趣的一组人，通过使用因特网进行相互交互。

virtual team（虚拟团队） 一个团队的成员在不同的地方一起开会。

virtual worlds（虚拟世界） 由计算机系统创造的人工世界，在这里用户具有沉浸在其中的感觉。

Voice of Customer（VOC，用户的声音） 通过网站访问者的直接反馈，对其他网站和线下渠道的标杆收集和报告，支持未来顾客行为的预测模型这些行为，将问题集中于“非和

如何做”的应用。

Web 2.0 高级因特网技术的流行术语，包括博客、维基百科、RSS、社会书签等。Web 2.0和传统万维网的一个最重要的区别是因特网用户和其他用户、内容提供商、企业之间更大的合作。

Web analytics（Web分析） 商务分析活动在基于Web流程，包括电子商务上的应用。

Web content mining（Web内容挖掘） 从网页上提炼有用的信息。

Web crawler（网络爬虫） 自动读取网站内容的一种应用。

Web mining（Web挖掘） 通过基于Web的工具从网页上发现和分析关于网页的有趣并且有用的信息。

Web service（Web服务） 使软件服务和联系它们的分布式应用程序组装在一起的一个架构。

Web structure mining（Web结构挖掘） 从包括网络文档等链接上开发有用信息。

Web usage mining（Web使用挖掘） 提炼通过登录网页、交易等方式产生的有用信息。

Weka 一种流行的、免费的、开源的机器学习软件套件，在怀卡托大学用JAVA编写而开发的。

Wiki（维基百科） 一个服务器软件，允许用户使用任何网络浏览器在网站上自由共创和编辑网页内容。

华章计算机科学丛书经典推荐

深入理解计算机系统（原书第2版）
[美] Randal E. Bryant 著
David R. O'Hallaron
ISBN 978-7-111-321334-4
定价：99.00元

深入理解计算机系统（英文版原书第2版）
[美] Randal E. Bryant 著
David R. O'Hallaron
ISBN 978-7-111-32551-1
定价：128.00元

软件工程：实践者的研究方法（原书第7版）
[美] Roger S. Pressman 著
ISBN 978-7-111-32631-7
定价：79.00元

软件工程：实践者的研究方法（英文版原书第8版）
[美] Roger S. Pressman 著
ISBN 978-7-111-24[illegible]
定价：[illegible]元

计算机组织与结构：性能设计（原书第8版）
[美] William Stallings 著
ISBN 978-7-111-[illegible]
定价：79.00元

操作系统：精髓与设计原理（原书第6版）
[美] William Stallings 著
ISBN 978-7-111-20425-5
定价：85.00元

- Stallings，大师级的人物，涉猎计算机安全、网络、体系结构等多方面，堪称计算机界的全才。
- 本书是其经典著作之一，得到全球计算机领域工程技术人员的好评！
- 以当代最流行的操作系统——Windows Vista、UNIX和Linux为例，全面清晰地展现了当代操作系统的本质和特点，具有先进性和适应性。

设计原本（ISBN 978-7-111-32567-4 定价：[illegible]元）

Java语言程序设计 基础篇（原书第8版）
[美] Y. Daniel Liang 著
ISBN 978-7-111-34301-2
定价：79.00元

Java语言程序设计 进阶篇（原书第9版）
[美] Y. Daniel Liang 著
ISBN 978-7-111-34238-8
定价：79.00元

教师服务登记表

尊敬的老师：

您好！感谢您购买我们出版的_____教材。

机械工业出版社华章公司为了进一步加强与高校教师的联系与沟通，更好地为高校教师服务，特制此表，请您填妥后发回给我们，我们将定期向您寄送华章公司最新的图书出版信息！感谢合作！

个人资料（请用正楷完整填写）

教师姓名		□先生 □女士	出生年月		职务		职称：□教授 □副教授 □讲师 □助教 □其他
学校			学院			系别	
联系电话	办公： 宅电： 移动：			联系地址及邮编			
				E-mail			
学历		毕业院校		国外进修及讲学经历			
研究领域							

主讲课程	现用教材名	作者及出版社	共同授课教师	教材满意度
课程： □专 □本 □研 人数： 学期：□春□秋				□满意 □一般 □不满意 □希望更换
课程： □专 □本 □研 人数： 学期：□春□秋				□满意 □一般 □不满意 □希望更换

样书申请			
已出版著作		已出版译作	
是否愿意从事翻译/著作工作 □是 □否		方向	
意见和建议			

填妥后请选择以下任何一种方式将此表返回：（如方便请赐名片）

地 址：北京市西城区百万庄南街1号 华章公司营销中心 邮编：100037

电 话：(010) 68353079 88378995 传真：(010)68995260

E-mail:hzedu@hzbook.com marketing@hzbook.com 图书详情可登录http://www.hzbook.com网站查询