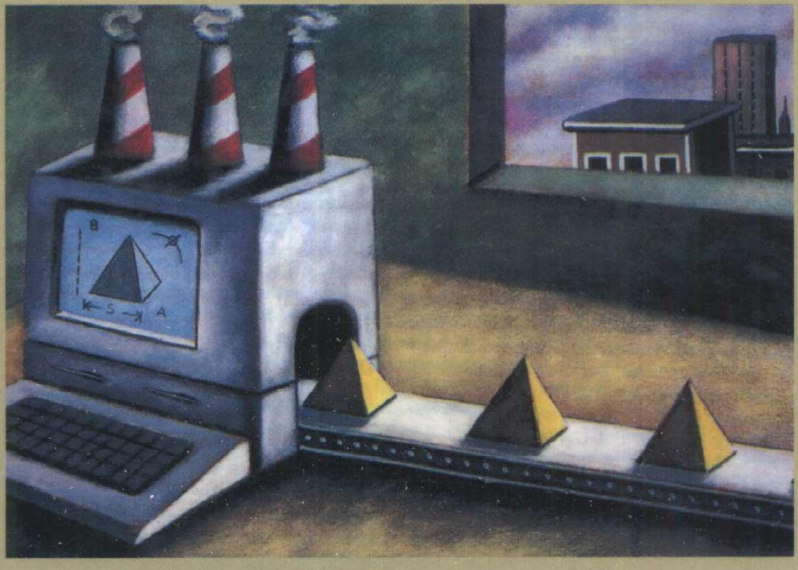
数据挖掘
——构筑企业竞争优势

DATA MINING

Building Competitive Advantage

〔美〕R·格罗思

侯 迪 宋擒豹 译



- Hands-on, "Learn-doing" approach
- Focuses on business solutions and results
- "Test drive" trial versions of leading mining tools

西安交通大学出版社

数据挖掘
——构筑企业竞争优势

〔美〕R·格罗思
侯迪　宋擒豹　译

西安交通大学出版社

内容提要

本书以面向实际的风格介绍了数据挖掘这一数据库领域中的最新技术。全书包括3个部分。第一部分主要介绍数据挖掘的基本概念、术语、方法、过程及常用的几种挖掘算法，如决策树、聚类分析、遗传算法、神经网络、贝叶斯置信网络、统计分析、关联规则分析等。最后对数据挖掘工具市场的状况、主要厂商及信息源作了详细介绍。第二部分介绍了两个目前较为流行的数据挖掘工具产品：Knowledge SEEKER 和 Right Point DataCruncher。书后附带的 CD-ROM 还为读者提供了这两个产品的演示版程序。第三部分以大量实际例子介绍了数据挖掘技术在银行、金融、零售、医疗保健和电信等行业中的应用，并以实际例子介绍了如何在数据仓库基础上进行数据挖掘的技术和方法。

本书主要以企业中从事经营管理、市场营销和计算机信息系统开发等方面的实际工作人员为对象。因此，本书对于广大的企业实际工作人员快速了解和掌握数据挖掘技术极具参考价值。

本书既可以作为数据挖掘技术的培训教材也可作为计算机信息系统相关专业的高年级大学生、研究生和教师的教学参考书使用。

图书在版编目(CIP)数据

数据挖掘：构筑企业竞争优势/(美)格罗思(R·Groth)著；侯迪，宋擒豹译，—西安：西安交通大学出版社，2001.8
书名原文：Data Mining: Building Competitive Advantage
ISBN 7-5605-1420-0

Ⅰ.数… Ⅱ.①格… ②侯… ③宋… Ⅲ.①市场营销学②信息-分析 Ⅳ.F713.50

中国版本图书馆 CIP 数据核字(2001)第 050705 号

*

西安交通大学出版社出版发行
(西安市兴庆南路 25 号 邮政编码：710049 电话：(029)2668315)
西安交通大学印刷厂印装
各地新华书店经销

*

开本：890 mm×1 240 mm 1/32 印张：8.75 字数：244 千字
2001 年 8 月第 1 版 2001 年 8 月第 1 次印刷
印数：0001～5 000 定价：20.00元(含光盘)
陕版出图字：25-1999-088 号

译者序

这是一本介绍数据挖掘应用技术的书籍。

数据挖掘是从大型数据源中抽取人们感兴趣的知识，这些知识是隐含的、事先未知的和潜在有用的重要信息。它不仅被数据库、人工智能和机器学习界的研究人员看作是一个越来越重要的研究课题，而且也被工商界视为能带来巨大回报的领域。为此，研究人员进行了大量的研究和探索，产生了许多的方法、技术和工具；工商界也以极大的热情采用了这些成果，出现了众多的应用系统，成为继 Internet 之后的又一技术热点。数据挖掘技术已广泛而成功地应用于诸如制造、零售、运输、电信、银行、医疗保险、资产评估、科学研究和政府决策等领域，取得了可观的经济效益和社会效益。

正如作者所言，早在越战时期人们就已经在进行数据挖掘的研究工作了。那么，为什么它突然会在一夜之间就充斥了报刊杂志并引起了人们的广泛关注呢？

本书对数据挖掘的基本理论和定义作了概要性的描述，对数据挖掘领域的现状进行了分析，并结合实际案例对数据挖掘在不同行业中的应用情况进行了讨论，也对各种数据挖掘产品进行了详细的介绍，相信读完本书读者自有结论。

与目前常见的学术专著比较，本书以应用为聚焦点，在内容上追求实用，是商务专家、商务顾问、大学生和有关科技人员了解数据挖掘技术的不可多得的参考资料。

本书的序、前言、第 2,4,8 章和附录 A,B 由侯迪翻译，第 1,3,5,6,7 章和索引由宋擒豹翻译。

由于译校者水平所限，加之数据挖掘是一个新兴的研究领域，有些专业术语商业界远未达成共识，给翻译工作带来了很大困难。译文中差错之处，诚望读者指正。

译者

2001 年 1 月于

西安交通大学

第二版　序言

从本书第一版出版到现在短短两年中数据挖掘领域已得到了飞速的发展。仅从下面这些较为引人注目的事件就可以知道这种发展的速度有多么惊人：

- SAS 研究所发布了 Enterprise　Miner。
- SPSS 收购了 ISL 公司及其 Clementine 数据挖掘软件。
- Yahoo!　收购了 HyperParall 公司。
- Aspen 技术公司收购了 NeuralWorks 公司。
- Oracle 公司收购了思维机器公司。
- 许多数据挖掘产品的供应商(如 DataMind 公司)对自身进行了重新定位，将重点转移到数据挖掘在行业领域中的应用。

本书第一版曾着重强调了数据挖掘技术已引起人们普遍的关注，并援引了一些例子，如 1996 年 1 月期的《银行系统与技术》上的一篇文章认为“数据挖掘将是 1996 年各项金融业务中最重要的一项应用”。1996 年，IBM 在 SuperBowl 期间举办的一场商业活动中出现了时装模特们参与讨论数据挖掘应用优势的场面。此外，MATE 集团发表的一份统计图表表明到 2000 年数据挖掘技术市场将达到 8 亿美元。

目前，数据挖掘这一领域仍然有着巨大的发展潜力，参与这一领域的竞争者也在发生着快速的变化。因此，本书第二版在有关当前这一领域的竞争者及行业发展趋势方面的内容也需要作出相应的修改。例如，目前推动数据挖掘技术发展的主要动力已从面向工具为主转变为以面向解决方案为主。本书第二版在有关如何应用数据挖掘技术解决业务问题方面做了大量扩充。作为本书第二版的读者，你不但能够从中了解到这一领域当前的发展趋势，而且可以了解什么是数据挖掘以及如何应用这一技术为你获得竞争优势。META 集团得出了这样的结论：“到公元 2000 年，全球 2 000 家主要机构将发现数据挖掘技术已成为其业务成功的关键因素。”尽管这已引起了许多人对数据挖掘的兴趣，但总有一些特殊的理由使我们能够相信这一结论的真实性。由此产生的问题是：为什么全球 2 000 家主要机构将发现数据挖掘技术已成为“关键因素”？数据挖掘

将给我及我的企业带来什么收益?

本书的目标

我们编写《数据挖掘——构筑企业竞争优势》一书的主要原因是,虽然数据挖掘正在成为主流性的趋势,但目前还没有一本有关数据挖掘方面的书是专门针对商务专家编写的。本书为商务专家、大学生及商务顾问掌握数据挖掘技术提供了一种既新颖又简便的学习方法。本书后面所附带的 CD-ROM 将使读者的学习过程更加生动具体。读者可以对CD-ROM中包含的数据挖掘软件包进行尝试,并可学习到如何使用这些工具解决实际问题。

本书将着眼于如何在不同行业领域进行知识发现,并介绍了几种比较流行的数据挖掘软件产品。此外,本书还提供了几个针对不同行业领域的研究案例,其中涉及的行业包括零售、银行、保险及医疗保健等。

与目前市场上常见的学术性书籍相比,本书在对数据挖掘的介绍上采取了一种不同的方法。本书的重点是数据挖掘技术在行业领域的应用以及对特定问题的分析,而在内容上则以一种实用可行的风格向读者展示了如何运用数据挖掘工具来获得业务上的收益。

本书将回答下面几个基本问题:

- 什么是数据挖掘?
- 在当今行业领域中应当如何应用数据挖掘技术?
- 为什么要进行数据挖掘?
- 在数据挖掘市场中都有哪些厂商?
- 在什么地方可以找到有关数据挖掘方面的信息?
- 你应当如何进行数据挖掘?

行业焦点

数据挖掘正处在不断发展中,新的术语和方法层出不穷。为了帮助读者理解数据挖掘究竟是什么,本书广泛介绍了数据挖掘在目前各个行业领域中的实际应用。此外,本书还介绍了一些目前市场上较为流行的数据挖掘工具,并列出了可以使读者获得更多信息的资料提供商及 Web 站点。

本书全面概括了数据挖掘技术在各种行业领域中的的应用，其中包括：

- 银行和金融。
- 零售和营销。
- 长途电信。
- 医疗保健。

本书拓宽了学习数据挖掘技术的相关范围。即我们不仅应当掌握为使用数据挖掘所需要的方法和术语，而且还应当了解有关如何在企业环境中快速获得结果的实际例子。

本书为读者提供了大量以前难以接触到的有关数据挖掘的行业解决方案。出于竞争上的考虑，目前大部分公司都不大愿意谈论是什么原因使其获得极高投资回报。使用数据挖掘技术的行业应用促进了企业的竞争优势。人们可以使用数据挖掘技术进行下列预测：哪些客户最有可能对促销活动作出响应？哪些保险索赔具有欺诈性？客户最愿意购买哪些产品？企业获得的这种预测结果越成功，他们的嘴巴就会闭得越紧。

面向实际的教学风格

本书为学习数据挖掘提供了一种面向实际的方法。你只要花费 3 个小时左右的时间就可以使用本书附带的 CD-ROM 熟悉所有的主要过程。

在介绍了数据挖掘的概念之后，我们将直接进入实际应用，从而表明通过数据挖掘工具将数据转变为信息是如何的容易。本书附带的 CD-ROM 中包含了两个数据挖掘工具，分别是 Angoss 公司的 KnowledgsSEEKER 和 RightPoint Software 公司的 DataCruncher。

本书的读者

本书对数据挖掘作了一般概述，其内容可适应多种类型的读者。本书将对下列读者有所帮助：

商务专家

在商务领域中任何需要做大量数据处理工作的人都将对数据挖掘特别是本书产生兴趣。本书中除了提供许多实际例子外，还在帮助读者理

解数据挖掘产品的使用方面下了许多工夫。

数据库管理员(DBA)

对于数据库管理员而言,本书将从终端用户如何能够从当前的关系数据库和数据仓库中抽取数据以满足挖掘要求这一方面提供帮助。本书对不同行业的样例数据结构以及在不同类型的数据挖掘研究中所使用的数据域进行了讨论。

市场分析人员

由于数据挖掘可以使企业对其客户的了解和分析达到前所未有的水平,因此它对于从事市场营销的组织将具有特殊的实用价值。有人称此为"一对一"的市场营销。目前大批量邮寄广告发行商都已普遍使用了数据挖掘工具。对于市场营销组织而言,几年之内,数据挖掘将成为一种必须,而不再是一种无关轻重的尝试。

在校大学生

在校大学生所需要的是对数据挖掘基础的实用性介绍,并且他们可以本书为起点进一步参与市场实践。

系统分析员和咨询员

系统分析员和咨询员将可以从涉及该市场的有关供应商的评价以及对特定行业实际例子的讨论中得到帮助。

本书的范围

本书的标题为:数据挖掘——构筑企业竞争优势。这本书没有对数据挖掘中使用的算法进行详细的解释。如果你希望对数据挖掘的算法有更多的了解,我们建议你可以进一步阅读由 Usama M. Fayyad, Gregory Piatestsky-Shapiro, Padhraic Smyth 及 Ramasamy Uthurusam 等编写的《知识发现与数据挖掘进展》(Advances in Knowledge Discovery and Data mining)一书。该书篇幅为 550 页,是目前有关数据挖掘理论方法方面最权威的一本著作。

本书是为商务专家了解数据挖掘而编写的，其目标定位于那些不具备统计学知识但希望了解数据挖掘的应用以及那些希望亲自尝试进行数据挖掘的读者。

本书的组织

本书分为以下 3 个部分：

第 1 部分　入门篇

本书开头部分的几章主要介绍数据挖掘、数据挖掘过程及参与这一市场的有关厂商。

第 1 章“数据挖掘引论”，介绍数据挖掘的基本概念并解释了其重要性。

第 2 章“数据挖掘入门”，介绍数据挖掘中可以采用的几种方法以及这些方法的优点。

第 3 章“数据挖掘过程”，概括介绍数据挖掘的过程并介绍了不同类型的挖掘以及数据准备方面的问题。

第 4 章“数据挖掘算法”，介绍目前数据挖掘领域中所使用的各种类型的算法和技术。

第 5 章“数据挖掘市场”，介绍数据挖掘市场中的有关厂商，并对 SAS 的企业挖掘器和 IBM 的智能挖掘器的应用状况进行了讨论。

第 2 部分　工具篇

本书第 6 章和第 7 章介绍了两个有代表性的数据挖掘软件产品。

第 6 章“决策树方法在数据挖掘中的应用：KnowledgeSEEKER 简介”，全面介绍了一个由 Angoss 公司开发的商用数据挖掘软件产品，KnowledgeSEEKER 采用决策树模型并以终端用户为主要应用对象。

第 7 章“RightPoint DataCruncher 简介”。

第 3 部分　应用篇

本书第 8 章和第 9 章集中介绍数据挖掘在特定行业中的应用。其中包括进行实际挖掘的例子，以及如何在企业数据库中进行数据挖掘的有

关技巧。

第8章“数据挖掘的行业应用”,介绍数据挖掘在银行和金融、零售、医疗保健以及长途电信等行业中的应用,并给出了若干企业开展数据挖掘的实际例子。

第9章“基于数据仓库的数据挖掘”介绍如何应用数据仓库辅助进行数据挖掘的方法。该章还以4个行业性数据仓库为实际例子讨论了需要集成的数据的类型并介绍了若干使用这些数据仓库进行数据挖掘的案例。

目 录

第 1 部分 入门篇

第 2 部分 工具篇

第 3 部分 应用篇