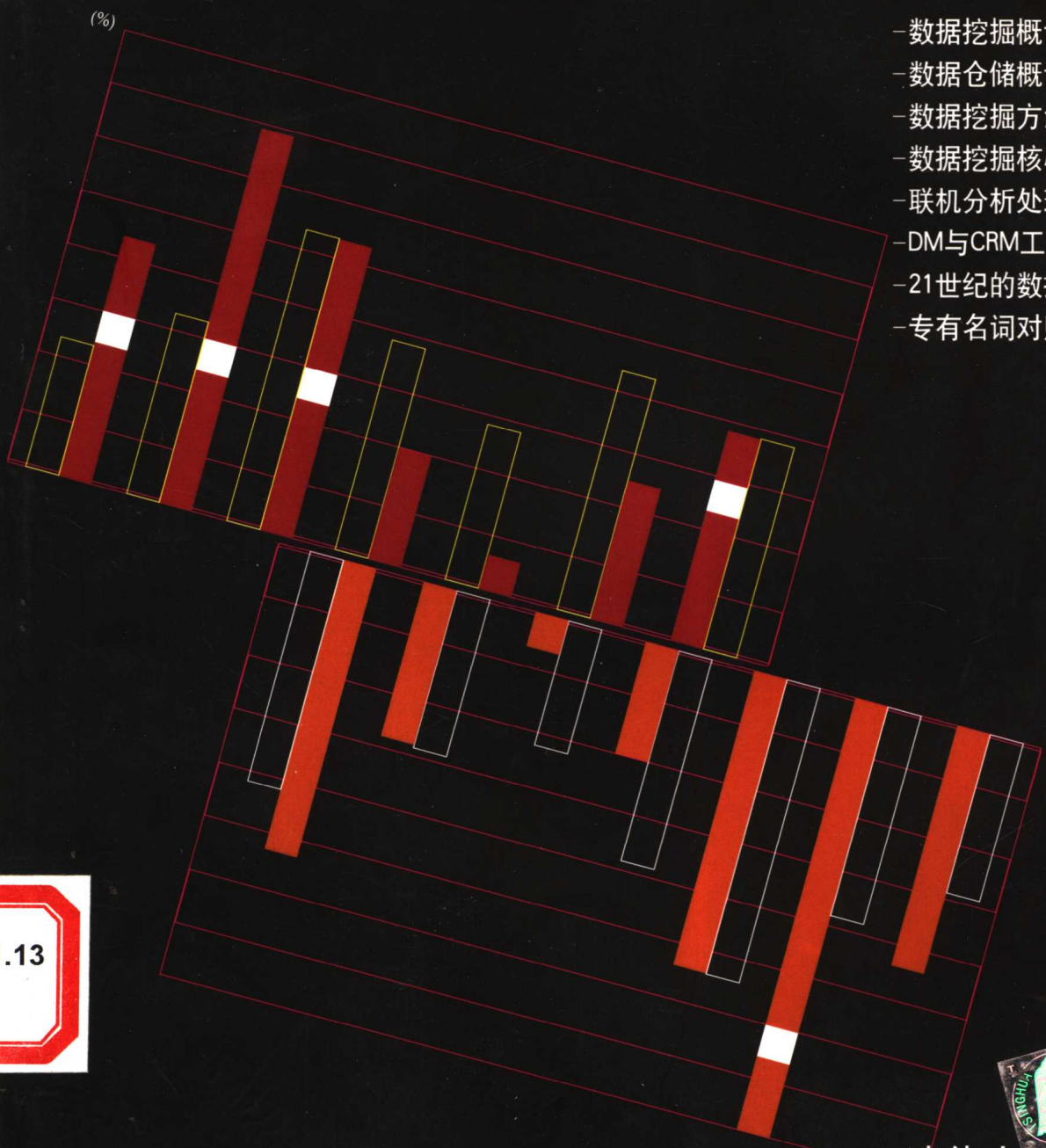
数据挖掘与OLAP
理论与实务

林杰斌　刘明德　陈湘　编著

本书主要内容:

- 数据挖掘概论
- 数据仓储概论
- 数据挖掘方法论
- 数据挖掘核心技术
- 联机分析处理(OLAP)
- DM与CRM工具软件
- 21世纪的数据挖掘
- 专有名词对照表



清华大学出版社

数据挖掘与 OLAP 理论与实务

林杰斌　刘明德　陈　湘　编著

清 华 大 学 出 版 社

（京）新登字 158 号

内 容 简 介

数据挖掘是近年来伴随着人工智能和数据库技术的发展而出现的一门新兴技术。本书讨论数据挖掘理论与应用专题，包括数据挖掘和数据仓库简介、数据挖掘方法论、数据挖掘核心技术、联机分析处理、DM 与 CRM 工具软件、21 世纪的数据挖掘等内容。

本书可供信息技术、信息工程、信息管理、统计、电子商务、生物信息和计算分子生物学等相关科系及研究所学生作为教科书或参考书籍使用，也可作为统计信息软件公司、电子商务网络公司、设计/制造业、服务业(大型百货公司及超市)等相关行业的研发人员及客服中心人员的参考教材。

图书在版编目（CIP）数据

数据挖掘与 OLAP 理论与实务/林杰斌，刘明德，陈湘编著.—北京:清华大学出版社，2002
ISBN 7-302-06140-8

I. 数… II. ①林… ②刘… ③陈… III. ①数据采集-技术 ②数据库系统 IV. TP311.13

中国版本图书馆 CIP 数据核字（2002）第 097401 号

出 版 者：清华大学出版社(北京清华大学学研大厦，邮编 100084)
http://www.tup.com.cn
http://www.tup.tsinghua.edu.cn
责任编辑：赵乐静
印 刷 者：世界知识印刷厂
发 行 者：新华书店总店北京发行所
开　　本：787×1092 1/16　**印张**：16.5　**字数**：395 千字
版　　次：2003 年 1 月第 1 版　2003 年 1 月第 1 次印刷
书　　号：ISBN 7-302-06140-8/TP・3671
印　　数：0001～4000
定　　价：25.00 元

序

统计学是一门大约有一百年历史、经常推陈出新、与时代同步发展的崭新学科。尽管它充分运用了自动化的数据收集技术，采用了便宜而且快速的计算机软硬件，但在新的千禧年所面临的最大课题还是如何提高处理大量复杂数据集的能力。由于传统的统计推理理论局限于小型样本，建立在较有局限性的假设上，因此无法处理大量复杂的数据集，而从数据中搜索结构化与较粗糙化信息的重要性与日俱增。对大量的信息作数据挖掘(Data Mining，DM)将使以狭窄假设为基础的统计推理更具客观性。目前，科技综合研究将在生物信息学、制药业与高科技等新兴科技方面大放异彩，成为 21 世纪最热门的学科，未来将朝下列趋势发展：

- 大型、复杂的数据处理：运用人工神经网络模型及大型数据库的数据挖掘技术。
- 经验——面向物理的步骤：由数据与机械化知识所驱动的步骤。可由未知的状态推导出统计的形式，并从所观察的数据中归纳出未知的状态。
- 知识的表现：通过贝氏先验模型(在多维空间中)、计算算法及与认知科学的互动来呈现知识，而计算机、信息与统计科学的科技综合将涵盖下列方向：
  - 多重接口：数据挖掘、人工神经网络及人工智能。
  - 大胆假设：发展算法与程序设计，用来解决大型的实际问题。
  - 运用“数据挖掘”技术，而非统计推理与估计。

其中数据挖掘的方向将沿着可延展性算法、对数据的精简处理、数据压缩、结构化信息与具有重大效益的方向搜寻。

而科技综合研究将涵盖下列方向：

- 信息技术：数据挖掘、人工神经网元、影像与语音分析/识别、机器学习。
- 生物信息学：统计基因学、基因序列定序、芯片数据分析。
- 设计/制造业：计算机实验、大量生产数据的挖掘。
- 服务业：信用卡欺诈检验、购买模式分析、零售数据挖掘、销售与行销分析。

数据挖掘(又称数据库中的知识发现——Knowledge Discovery in DataBases，KDD)是近年来伴随着人工智能和数据库技术的发展而出现的一门新兴技术。它可从大量的数据中提取出隐含的、以前不为人所知、可信而有效的知识。它能够对数据进行再分析，以期获得更加深入的了解，并具有预测功能，可通过已有数据预测未来。

由于数据挖掘是尚未完全成熟的新兴学科，本书只能就较重要的理论与应用专题讨论，读者如有兴趣，可进一步参阅相关的参考数据。

本书可供信息技术、信息工程、信息管理、统计、电子商务、生物信息和计算分子生物学等相关科系及研究所学生作为教科书或参考书籍使用，也可作为统计信息软件公司、电子商务网络公司、设计/制造业、服务业(大型百货公司及超市)等相关行业的研发人员及客服中心人员的参考教材。

本书编写时间仓促，谬误或疏漏之处在所难免，恳请读者不吝指正。

内 容 摘 要

在数据仓库系统建立过程中，通过数据综合转换(ETL)将数据有计划地输入数据库中；接着要有前端用户接口以供用户作查询分析时查看。数据仓库可以应用的范围包括决策支持(DSS)、客户关系管理(Customer Relation Management，CRM)及一些固定报表查询功能等。

前端用户接口即根据使用方式(如一般报表查询、交叉分析、统计分析等)而建立不同的接口工具。以下介绍两种数据仓库前端用户接口的应用。

- 联机分析处理(OLAP)：该处理系统用来帮助用户有效且轻易地完成商业信息的维面结构分析工作，可将数据仓库的数据加以筛选、分类、汇总，产生物理数据，再通过各种数据模型呈现给用户，让用户可从不同的主题和角度根据专业的直觉，经由复杂的查询、数据对比、数据选取和报表来提供不同层次的分析。通过OLAP，用户有需求时，只需利用工具就能找到各种数据，例如管理角度的交叉分析、数据排名、预算及实际值的比较等，不需要再排队等待信息人员撰写程序，而以简单的操作方法就可获得想要查询的数据。
- 数据挖掘(Data Mining)：指利用分类、关联性、序列分析、群集分析、机器自我学习及其他统计方法，从数据库中庞大的数据中，找出隐藏的、未知的、但却对企业经营十分有用的信息。有些企业的历史数据以百万、千万计，要分析起来相当困难，而利用数据挖掘工具，可从庞大的信息中提取有用的信息，以客观的统计分析方法快速而且正确地找出企业需求的经营信息，得到正确的销售模式、客户关系和行销策略等，以利于企业掌握正确的经营动态，增加利润并减少支出。

数据挖掘与 OLAP 的不同在于 OLAP 主要是忠实客观地呈现出用户想查询的众多因素分析汇总得出的报表，而报表的翻译将由用户判断。数据挖掘则能够更进一步利用统计等方法对数据再分析，以获得更加深入的理解，帮助用户得知原因。且数据挖掘拥有预测功能，可由已有的数据预测未来。

目前市面上有许多强调数据挖掘、OLAP 的产品，必须提醒企业的是，数据挖掘功能固然吸引人，但前提是有完整、正确的数据。因此，通常在建立数据仓库系统基础后，再加此功能以达到加倍效果。若没有建立完善的数据仓库，数据未经过整理综合，OLAP 及数据挖掘是不会有太大的使用空间的。

有些企业有数以百万笔的历史数据，若要分析，相当困难，而且容易错失企业应有的商机。而数据挖掘工具则能从庞杂的信息中提取有用的数据，通过公正客观的统计分析，快速而且正确地得知企业经营信息，找出销售模式，准确掌握未来的经营动态。

数据挖掘指通过高等统计工具的使用，从数据库或其他计算机储藏中识别出对商业有用的样本或与商业有关的程序，收集与顾客相关的数据，利用统计分析与人工智能，对大量数据进行筛选、推演与模型建造等程序，以揭露隐含在数据与模式中的可把原始数据转换成商机并成为决策依据的新知识。从 CRM 的整体结构来说，数据挖掘是整个 CRM 的核心，也是构成商业智能的基础。

完整的数据挖掘不单可以做到准确的目标市场行销。当分析的工具和技术成熟时，加上数据仓库提供大量储存顾客数据的能力，可让数据挖掘做到大量针对个人的数据定制，从而准确地对顾客作一对一的行销。企业对顾客有充分的了解，才能有效地和顾客建立关系，进而有效地进行行销，创造商机。数据挖掘是 CRM 中的基础和核心，通过数据挖掘，能有效地提供行销、销售、服务的决策支持，让工作人员可以得到充分的信息来行动，并预测在适当的时间、地点，提供给顾客合适的产品及服务。

一旦企业对顾客的了解程度提高，针对目标市场行销的准确度将大幅提高，从而直接影响到成交的比例。在目前彼此竞争激烈的企业之间，如能了解顾客需求，就可以有效地过滤无效的样本，在未接触顾客以前就能知道哪些顾客会是未来成交的对象，于是减少以往漫无目的的行销策略，较其他竞争者获得先机。

进入 21 世纪，由于网络及数据库系统比较发达，企业得以收集大量的、实时的顾客数据。然而面对成千上万的数据，大多数企业却只是选择采用有效的技术去储存它，却不能对数据进行进一步分析，以进行商业策略的应用。

数据挖掘能为企业界做什么

在企业累积了一定数量的用户数目之后，若要进一步掌握客户消费模式，了解市场商机所在，就必须通过数据仓库、数据挖掘等系统进行客户分析。通过客户关系管理系统，企业将推出量身定做的产品与服务，进而提升客户忠诚度，从而增添业绩。

直效行销

数据挖掘技术最早是应用在邮件领域。如何利用最少的行销成本将行销讯息传递到正确的顾客身上，除了运气及经验成分之外，还需要进一步地对数据挖掘的结果进行更精准的 DM 发送。

顾客关系管理

顾客至上是许多服务业奉为经典的口号，但是事实上，顾客总是会质疑：为什么我和你们交易那么久了，你们却永远不认识我？就像即使是合作了 10 年的银行，当顾客要申请一个定期存款时，依然要填写冗长的表格。凭借长期累积的顾客交易信息，数据挖掘可以进一步建立顾客的消费模式，除了监控现有商品的消费情形之外，同时还能够根据模型主动促销顾客可能感兴趣的商品。

交叉销售

“从已经花了 10 块钱的顾客身上再赚 10 块钱，绝对远比向没掏出钱的顾客下手容易。”数据挖掘能够找出商品之间购买的关联性，以建立起交叉销售的行销策略，从而降低商品行销的成本，让 20%的客户能够制造出 80%的收益。

信用评估

研究报告指出，全美移动电话通信费用呆账额一年高达两亿美元。无论是通信费用付款、保险费用估算、核查信用卡额度或是个人及公司的贷款，若是能够利用数据挖掘的模型预测出谁是潜在的不良顾客，便能够早一步提出预防措施并减少企业损失。

财务预测

利用时间序列分析或是人工神经网络，可以建立起季节性或非季节性的财务数字预测，同时也能够预先估计促销活动对于销售数字及获利的影响，从而让企业界尽量不作出错误的决策。

然而，企业界面临的难题却是：接受数据挖掘这样的观念容易，但是实际将它落实在企业中却相当困难。首先要面对的就是人才的问题，由于数据挖掘所涉及的技术层面涵盖了数据库结构、统计学、人工智能、行销学，甚至是各企业的“专业领域技能及窍门”，同时要针对企业不同的问题，利用不同的分析工具进行行销策略的拟定及评估。由于数据挖掘的概念仍相当新颖，像这样跨领域的人才相当稀少，美国时代杂志将数据挖掘分析员评为 21 世纪最热门行业的第 5 名。

由于人才缺乏，因此企业界只能退而求其次，依靠自动化及半自动化的软件包进行数据库分析。事实上，目前许多企业使用的所谓的数据挖掘软件包，其实都仅是数据存取与数据报表或是多维联机分析处理工具(OLAP，online analytical processing)而已。这些工具虽然能够自动、快速地分析数据，但是只能提供简单的总数、平均数或是标准差而已，这只能够称得上是数据仓库部分的清点基本功夫，当然无法进行更复杂的数据库知识发现(KDD，knowledge discovery in databases)。

此外，撇开市面上未成熟的工具不谈，真正用来进行数据挖掘的软件工具定价相当高，同时市面上并没有一项完全涵盖所有数据挖掘任务、负担相当高的软件购置和人力成本，使企业无法立竿见影地看到数据挖掘所带来的效益，因而造成数据挖掘任务的中断。

综合应用多种工具

没有一种工具能够解决所有的问题，更何况数据挖掘涉及统计、数据库管理以及人工智能等广泛领域，因此惟有拥有充足的工具才能解决所有企业将面临的问题。除了使用市面上所售的数据挖掘工具外，同时还有针对各种专业领域以及特殊案例设计的规划分析工具、技术，包括了统计、人工神经网络、模糊系统、专家系统及灰色系统等。

与网络工具及企业内外部数据综合

若是要进一步结合消费者较深层的信息，或是对其他网页的活动记录进行数据挖掘工作，单靠企业内部的客户交易记录是不够的。因此需要综合企业内外部的信息作为信息源，全方位地建立正确的模型，以提高分析的准确度。

解析技巧

- 统计学
- 人工神经网络
- 遗传算法
- 决策树
- 关联性分析
- 面向记忆的推理
- 购物篮分析 / 规则建构

- 模糊集理论
- 专家系统
- 灰色系统

近年来公司与其顾客间的关系已经演变成一个非常重要的研究议题，因此顾客关系管理的概念于近几年内被提出，并渐渐地将过去的顾客分析概念纳入其范围中。因为如此，顾客关系管理也在企业中成为新宠，不再只是针对顾客来讨论，而是提供一个综合企业内部所有问题的整体全面解决方案，其领域包括顾客行销、销售以及顾客服务等，并且综合人员、流程与信息科技形成一个具有革命性冲击的新观念。

由于传统的行销观念已经被新出现的课题重新给予定义，因此许多竞争者纷纷朝向新的课题来谋求更多的发展空间。各企业主要的生存之道在于争取顾客、维系顾客以及有效增加会带来利润的顾客。除此之外，最重要的是如何建立一个面向顾客的企业，以维持市场上的竞争优势，这些都必须要从中找出目标顾客群以及充分地将数据综合分析，以便作出更好的相关决策。除此之外，如何在找到顾客之后确认其需求以及如何维系并创造新的顾客群已成为顾客关系管理当中最重要的两个步骤，也是未来研究所着重的两项主要目标。

因此，根据一份 1998 年 Customer Retention Practice Newsletter 的报道“典型的企业中有 80%的利润是 20%的顾客所创造出来的”，可以发现顾客在企业中所扮演的角色的重要性已经超越传统的型态，更说明了为何公司对于其主要顾客无不使用各种竞争手段与行销方法来加以维系与开拓新的族群。换言之，顾客对于各企业的影响力已经超越以往，不但成为企业的指标，更从被动变主动。另外，对于大型的企业来说，虽然已拥有较大的顾客群，但是仍需不断地寻找潜在的目标顾客群，以期为公司带来更丰厚的利润。而顾客忠诚度在潜在的无限商机中是相当重要的一个核心环节。因为除了可以利用已有的顾客群来刺激重复购买之外，还可通过分析其特征属性来预测潜在的顾客差异。但是如何能够从顾客关系管理的观念出发来真正达到提升顾客忠诚度的目标呢？我们认为，创造顾客利润与重点维系顾客应该是首要的开源节流目标。

顾客维系和顾客利润是企业赚钱的重要指标，但是定义却仍然相当模糊，因此我们提出了明确而简单的定义。所谓“顾客维系”，就是了解顾客过去的历史性消费行为，并且在经过数据分析后，让公司的决策者可以分辨出哪些顾客维系住的概率最大。一般来说，顾客需要经过 5~10 次的重复购买才有可能成为主要的顾客群，因此只要增加 5%的顾客维系率，就有可能相对地增加 60%~100%的利润，由此可以看出顾客维系的重要性。顾客利润是一种管理与了解顾客的能力以及顾客所表现出来的潜在价值，它也是由某些元素所构成的，如价格、奖金与资产、交叉销售、服务的成本、交易数据量的大小、期望的损失值、寿命与消耗等。任何一个元素都可以提升顾客利润。因此，决策者便可以针对上述任一元素来着手，以期提升目前的顾客利润，进而提高客户关系管理的整体附加价值。

总之，数据挖掘技术能够进一步运用统计等方法对数据进行再分析，以获得更加深入的了解，并且具有预测功能，可从已有的数据来预测未来。

目　录

第 1 章　数据挖掘简介

1.1　什么是数据挖掘

数据挖掘是近年来随着人工智能和数据库技术的发展而出现的一门新兴技术。它是从大量的数据中筛选出隐含的、可信的、新颖的、有效的信息的高级处理过程。

数据挖掘是面向事实的，例如在经典的“尿布与啤酒”的例子(美国 WalMart 超市，细节请参阅 1.6 节)中，经理和行销专家在事前并没有与“尿布与啤酒”相关的知识。在数据挖掘中，数据分为训练数据、测试数据和应用数据三大部分，而这三部分的比例依据经验来确定(例如 1:1:8)。数据挖掘力图在训练数据中发现事实，并以测试数据作为检验和修正理论的依据，而最后把知识应用于数据中。数据挖掘的关键性思路为实事求是。“实事”即“数据”，“求”就是去发现、去挖掘、去探索，“是”即数据中隐藏的规律。

例如在“尿布与啤酒”的例子中，数据挖掘假设事物之间总有关系，并用支持度和信赖度来衡量其强弱。数据挖掘运用了 Apriori(验证)算法或其改进型算法，检查数据库中商品同时销售的事实，发现了下列规则，即客户购买尿布(支持度＝5%)，则客户同时购买啤酒(信赖度＝70%)，由此可见数据挖掘强调“面向事实”。而“数据挖掘”在方法论上强调“面向数据”，由于它充分运用了自动化的数据收集技术与速度快、容量大的计算机，从而具有处理大量复杂数据库的能力，而这可能是数据挖掘的研发比较顺利的原因。

总之，数据挖掘利用了分类、关联性分析、序列分析、群集分析、机器学习、知识发现及其他统计方法，从数据库庞大的数据中，找出隐藏的、未知的、但却对企业经营十分有用的信息。有些企业的历史数据是以百万、千万计，要分析起来相当困难，而利用数据挖掘工具，可从庞大的数据中筛选出有用的信息，以客观的统计分析方法，快速准确地找出企业所需要的经营信息，得到正确的销售模式、客户关系及行销策略等。

数据挖掘技术能够进一步运用统计等方法对数据进行再分析，以获得更深入的了解，并具有预测功能，可借助已有的数据来预测未来。

1.2　新世纪的统计科学

统计学是一门大约有一百年历史，并经常推陈出新的学科。它的研究范围涵盖了数据收集、统计推论与随机现象的模式化三大领域。由于统计学充分运用了自动化的数据收集技术与计算机，其目前面临的最大课题将是处理大量、复杂数据库的能力。由于传统的统计推论局限在小样本，无法处理大量、复杂的数据库。在新千年，从数据中搜寻结构化与较粗糙的信息的重要性将与日俱增。对大量的数据进行数据挖掘将是 21 世纪最热门的行业之一。而目前此种科技整合研究的趋势将在生物信息学、信息基因学、制药与高科技等新兴科技方面大放异彩。

1.3 数据挖掘是掌握商机的命脉

有些企业有数以百万计的历史数据，要精密分析相当困难，容易错失企业应有的商机；而数据挖掘工具能从庞杂的信息中筛选出有用的数据，以公正客观的统计分析快速准确地得知企业经营的信息，从而找出销售模式，正确掌握未来的经营动态。

数据挖掘通过高等统计工具的使用，从数据库或其他电子文件中识别出对商业有用的样本或关系的程序，并收集与客户相关的数据，利用统计分析与人工智能技术，针对大量数据进行筛选、推导与模型构造等操作，以揭露隐含在数据与模式中的金矿，从而把原始数据转换成商机，成为决策依据的崭新知识。从 CRM 的整体结构来说，数据挖掘是整个 CRM 最重要的一个阶段，也是构成商业智能整体解决方案的基础。

完整的数据挖掘不但可以做到准确的目标市场行销，当分析工具及技术成熟时，加上数据存储提供大量存储客户数据的能力，可使数据挖掘进行大规模的针对个人客户的定制，准确地对客户作一对一的行销。只有企业对客户有充分的了解，才能有效地和客户建立亲密的关系，进而有效地进行行销，创造商机。数据挖掘是 CRM 中的关键性阶段，透过数据挖掘，能有效地提供行销、销售和服务的决策支持，让工作人员得到充分的信息而展开行动，并在适当的时间和地点给客户提供适当的产品及服务。

企业一旦提高了对客户的了解程度，针对目标市场行销的准确度就会大幅提高，这将直接影响到成交的比例。在目前彼此竞争激烈的企业之间，如能了解客户的需求，就可以有效地过滤无效的原始数据，而在未接触客户以前就能知道客户可能是未来成交的对象，减少以往漫无目标的行销策略，从而较其他竞争者优先获得商机。

1.4 点 式 行 销

点式行销即把客户当成独特的个体来处理，它是一种量身定做的行销方式。虽然量身定做的行销方式可能发展很慢，但在 2005 年前后，消费者会习惯共同参与设计其购买的产品。

点式行销，或真正的量身定做行销，其成功率通常高于 50%，它代表个人化的一对一行销，充分地为个体服务，满足其独特的要求。在此模型之下，行销人员只寻求已对产品提出精确要求的客户；而客户常常因行销人员能够符合其对产品或服务的精确要求而主动联络。

点式行销的反应率高于 50%。由于其对客户要求的了解随着时间而增长，因此可以和客户建立起根深蒂固的紧密关系。亚马逊成功的关键之一便是它对客户的了解。任何上网登录的客户都会依据其过去的采购记录获得想购买的书、音乐或录像带的目录。亚马逊网站首页的横幅广告就是特别为不同客户根据其存储在亚马逊数据库中过去的关系及行为而设计的。此种精巧的响应技术称为数据挖掘技术，它能够解析客户的行为以产生量身定做的信息。

点式行销是行销策略的一种，照理应该可以在网络之外进行。但是它只有在互联网

之类的媒体上才有实用价值，而人们可以通过网络和计算机实时互动，而在互动中所留下的“足迹”足以让网站归纳出人们的购物偏好。

1.5　门户网站的信息挖掘法

用户往往先进入门户网站，然后再进入其他网站，因此只要不是一开始就进入的网站，都不能称为门户网站。而门户网站的重要性在于能够掌握用户的第一手信息。例如，什么人进来逛过网站、看过哪些东西之类的信息，用户的数据全部收集在门户网站中。

不过，为了有效运用这些信息，就必须活用“信息挖掘”。所谓“信息挖掘”，是指将收集到的庞大信息利用计算机从各个方面进行筛选，藉此可调查用户的上网频率、嗜好、购物趋向、年龄段等信息，进而从共同项目中取样。

经过取样之后的共同项目可以运用在门户网站的首页设计，如此才可针对不同属性的用户量身定做。

从企业的角度来看，如果在美国在线公司(American on Line，AOL)和雅虎(YAHOO)刊登广告，便可自动筛选出该公司所需求的客户层，因此该公司只要在这些人经常接触的网站刊登广告，营业效率就会提高。事实上，无论是雅虎还是销售书籍、CD 和录像带等的亚马逊书店(amazon.com)等的首页，都已运用了 Broad Vision 公司的一对一个人行销技术，针对不同用户提供不同的首页。虽然用户不会注意到这些差异，但是通过这种模式，网站广告今后将吸引更多的潜在客户上网浏览网页。

1.6　尿布与啤酒

购物篮分析(Market Basket Analysis)的吸引人之处在于它运用了关联规则。它相当清楚而实用，因为它说明了物质商品之间的相关性和它们为什么会组合在一起。

以下即为应用购物篮分析产生的实例(在美国大型超市 WalMart 发生的实例)：在星期四，消费者通常同时购买尿布和啤酒。

这个例子说明了购物篮分析最常产生的三种信息：有用的、明显的但无法解释的。

上述例子告诉我们，年轻夫妇通常在星期四的晚上准备好周末所需的物品：买尿布给婴儿用、买啤酒给丈夫喝(可以很直接地联想到，啤酒是为了周六晚间的篮球赛或棒球赛所准备的)。对 WalMart 的老板而言，如果将尿布和啤酒的货架放在一起，则可以大大地提高利润。因为规则易懂，原因合理，所以人人都可以自行推行其他策略，例如也可以将配啤酒的花生、洋芋片等零食陈列在附近的货架上。

所以属性之间的关联往往能给人出乎意料的信息，若透过数据挖掘软件得知：尿布→啤酒的支持度为 0.01，信赖度为 0.8，表明有 1%客户买了尿布和啤酒，而且在所有买了尿布的客户中，有 80%买了啤酒，通常人们认为风马牛不相及的啤酒和尿布之间竟然有如此大的关系(有了这一知识之后，人们进一步调查，才发现先生们在为婴儿买尿布时，总不忘记为自己顺便买些啤酒)。

1.7 千禧年新问题：不缺信息缺知识

随着数据库技术的成熟和信息应用的普及，人类累积的信息量正在成指数地增长。全世界每天存入数据库的数据数量超过万兆字符。

例如，在超级市场通过条形码扫描，把每一宗商品交易输入数据库中，一个中型超市贩卖的商品就有数万种，每天的交易量上万笔，如此大量的数据，传统的数据库不能很好地回答老板所关心的问题：商品在不同季节或一天的不同时间中的销售量有何变化规律？商品 A 销售量的增加是否会同时带动商品 B 的销售？如何调整商品的资金比例以达到最佳的资源调配？各种商品的销售之间是否存在一定的关联(如美国 WalMart 超市通过“知识发现”(Knowledge Discovery)技术意外地发现，尿布和啤酒常常摆在一起销售，原来先生们为小孩买尿布时又随手带回啤酒，而 WalMart 将尿布和啤酒放在同一货架，使得销售量双双增长)?

如今我们有太多的数据而总嫌知识不够。无怪乎未来学家奈斯比特(John Naisbitt)惊呼：“人类正被信息淹没，却饥渴于知识。”面临浩瀚无边的数据，人们呼唤从数据的汪洋大海中去芜存精、去伪存真。而“从数据库中发现知识”(KDD)及其核心技术——数据挖掘便应运而生了。

专家认为 KDD＝数据预先处理＋数据挖掘＋解释评估。由于预先处理和解释评估的研究较为成熟，所以目前 KDD 的研究和执行难点与重点都集中在核心的数据挖掘上。

1.8 为客户量身定做

量身定做是一个人行销中不可或缺的一部分，因为每个人的兴趣、能力、风格差异很大。计算机如果要帮助我们完成信息革命，就必须考虑这些差异。但量身定做将影响人们的隐私，而且会促使隐私保护政策随着时间而改变。在有关隐私的辩论持续不衰之际，越来越多组织可能发现亚马逊和雅虎的做法有可取之处。这些公司率先搜集个人信息，尽可能去了解顾客的偏好。它们很早就懂得：要成功，所获的信息必须保密，以取得客户的信任。此种做法之所以会在全球各地流行，因为它既保护了人们的隐私，也较容易执行。

“数据挖掘”也让消费者更有能力找到真正需要的东西。他们可以用全球语义信息网(Global Semantic Web)的红色链接轻而易举地找到相关的产品和它们的特性。个性化的信息存取和自动化能力，要能符合这些目的。在遇到有合适的新信息出现时，它们会根据产品的特性和指定的偏好是否符合，提醒你注意。娱乐业是量身定做的另一大受益者。将来或许能把历年来 5 万部电影的数据库缩小，并且激活自动提醒功能，要系统去“看”有没有新片发行，以便找到引起你兴趣的“完美”选择，然后你就可以舒服地坐在摇椅上，通过网络租片子回来观赏。

在医疗保健、金融、政府、法律以及其他许多服务领域中，量身定做的天地将更为宽广，理由很简单。在经济系统，零售交易、娱乐等方面，这些活动居于主宰地位。企业对企业(B To B)的客户关系管理也会走向量身定做，未来的市场将十分庞大。

1.9　数据挖掘的数字物理学

要问数据挖掘的未来状况如何，就如同要预测新生婴儿的后裔会长成什么样子一样困难。虽然我们已进入数字信息时代的黎明阶段，但在“数据挖掘”领域，却仍处于初始阶段。

“数据挖掘”是了解、探索与征服数字数据新世界的核心活动之一。它是一种识别与发现数据中所隐藏的有用结构而计算机化的程序。所谓的结构与数据中的形式、模型和关系密切相关，而形式即为数据中子集的抽象化描述，模型为整个数据库的统计或描述，“关系”即为数据子集域(性质)之间的特定相关性(例如拥有汽车的人有可能也会遭遇车祸而动手术)。

“数据挖掘”的定义相当富有隐喻性，它也是目前所面临的重要的挑战性问题之一。例如，如果我们想要了解数据建模的意义，并从数据中推导出有用的信息，那这种模型是否存在？此种有趣或有用的总结意味着什么？对可能描述任何已知有限数据库的量化模型与形式，我们是否已选取了正确的描述方式？所采用的模型为什么比较好？推测与事实之间的差异是什么？

上述问题都是新兴信息型经济中的神秘问题，我们将数据视为最基本与最有价值的资产。若能精通高维数据挖掘的技术，则在数字信息时代将会鹤立鸡群，一枝独秀。所以“数据挖掘”在未来的人类历史中将与物质世界中的“航行”和“采矿”一样并驾齐驱。

人类记录与存储数据的历史虽已有 5000 年，但数据分析是相当新颖的领域，尤其在天文学与医学方面的科学领域中，其应用更为广泛。

早期的数据库，例如与企业、军事相关的记录，都是小型、低维的样本，只能反映少于 10 个的变量。若要将数据加以分析与模式化，则涉及到运用图表的数据可视化技术。而数字计算机的高速化与大量数据的存储功能，戏剧化地改变了此种困境。人类天生生活在低维的环境中，我们的直觉与本能只能处理 3～5 维的事物，最多也只能处理 10 维的事物，根本无法处理 100 维或 1000 维的事物，更何况在电子商务、网络、制造业、财务与科学研究方面的维数成千上万。

由于人类能构造计算工具来处理大量的数据库，所以人类的分析能力能延伸至高维空间。目前的数据挖掘算法运用统计模型来处理数据，从中找出可靠的形式；其所运用的工具相当复杂，用概率论、信息论、统计推定理论、不确定理论、图论与数据库技术等数学技术来建模。数据挖掘技术是一种搜索密集型技术，其通常涉及到对可接受的解作迭代收敛。而在高维空间中，由了解所隐藏的信息内涵，决定如何从数据中找出其间的因果关系。所以有价值的数据如同金钱一样需要好好地保存，数据银行(data banks)的革新势在必行。如同提供电话服务与电力服务的物理设备一样，人们可与数据银行互动来存取、改变与操作数据，而使数据存储用户能将数据转化为具有附加价值的信息与知识，此即数据挖掘的终极目标。

数据挖掘技术可用来过滤、选择、定制与传送正确格式和正确文本的信息，由于所有的组织与个人通过联网都能连接到所存储的大量知识库中，故在任何时刻都可存取数据而