第一章　Data Mining 简介

§1.1　前　　言

当前,无论在学术界还是产业界 Data mining 都是一个相当时髦和红火的专题。Data mining 的汉语名称有 数据采掘、数据淘金和数据采矿,还没有一个一致公认的译法，一般还是喜欢用英文‘Data mining’这个词,今后我们常用 DM 这样的简写来表示，因为很难找到一个确切的汉语词汇与之对应。

直观上说，Data Mining 就是要采掘出隐藏在原始数据中对决策有用的信息，为管理和研究服务。难怪很多人称之为数据淘金。下面将会给出更详细的说明。

《商业周刊》中文版[①] 1997 年第 7 期发表的《在原始数据中寻找关系》一文(John W. Verity),讲的就是 Data Mining 及其在各个领域中的应用。虽然文章主要讲的是关于数据搜索技术、数据仓库技术，但讲的内容确是 Data Mining 在银行、保险、电信及超市等方面的具体应用问题,并给出了由于进行了 Data Mining 而获得了巨大回报的具体案例，如 MCI 通讯公司，Wal－Mart 百货公司等。这说明一开始 Data Mining 就是作为一个新兴的现代技术出现的。

信息技术的发展，企业、政府机构可以很方便地收集到大量的资料。例如超级市场的每一顾客、他的每一次购物的状况就自动汇集到电脑中，所以几乎不费太多的成本就收集到非常庞大的数据。又如银行客户在每次刷卡时,就自动将取款数量、用途、日期等种种信息输入电脑。因此,管理人员面对的数据量是非常之大的，用大量数据已不足于反映，通常称之为海量数据。数据大到几十万、上百万,这时很简单的运算,也会成为很困难的事,例如要将这上百万个数据按大小排个次序，就成了耗时的运算，更不用说进一

① 台湾地区出版物。

步的分析了。“量变引起质变”。面对海量数据，相应的分析方法必须要有新的思路，这是数据采掘面对的难题之一。与量的增大的同时，往往资料涉及的面也非常宽。例如保险公司不仅有投保人的资料，还可以存有他的家庭、亲属、所在单位的种种资料，从统计上看，资料的维数很高，这是数据采掘面对的另一个难题。

另一方面，海量数据中确实隐含着各种各样的信息，这些信息往往人们凭直觉与经验是难于发现的，而电脑的特点是不怕多、不怕烦，总是可以耐心地仔细处理，数据越多，对研究目标越了解，也就越容易发现有用的规律。事实上，人们的经验也就是从自己经历的过去资料中，凭自己的感觉归纳出来的一些被自己的实践证明是有效的规律，只是缺乏理论的依据，或依据不足，或是没有深入去研究分析而已。而电脑就可以汇集种种经历，可以汇集种种发现、综合的方法，由一些程序、算法来处理数据——这就是总结经验。

数据采掘正是在这种情况下，从一些个案的处理，克服了许多困难，利用了并行算法、人工智能、统计分析的技术，综合成一种新的、能快速处理大量、海量数据的技术，被美国麻省理工学院(MIT)评为在未来发展中最有前途的十大技术之一。

§1.2 什么是 Data Mining

Data Mining 是目前 IT 行业发展最快的产业，并且许多不同领域的专家，如统计学家，金融学家等，对 Data Mining 也产生了极大的兴趣。计算机技术，统计分析方法、各类算法及行业知识的结合推动了 Data Mining 技术的快速发展。

关于 Data Mining 的定义，有各种说法：

Hand *et al*(2000) 的定义“Data mining is the process of seeking interesting or valuable information in large data bases”(数据采掘是在庞大的数据库中找出有意义或有价值信息的方法)；

Bhavani (1999) 的定义：“Data mining is the process of posing various queries and extractions useful information, patterns, and trends often previously unknown from large quantities of data possibly stored in databases.”(数据采掘是从储存在数据库的大量数据资料中，设置盘问，提取以前未知的信息、模式和趋势的方法)；

Bhavani (1999) 的定义：“The process of discovering meaningful new

correlation, patterns, and trends by sifting through large amount of stored data, using pattern recognition technologies and statistical and mathematical techniques" 数据采掘是从大量储存的数据中，利用模式识别、统计和数学的技术、筛选发现新的有意义的关系、模式和趋势的方法)；

Kovalerchuk & Evgenii Vityaev 的定义："These techniques are now applied to discover hidden trends and patterns in financial databases" (这些技术现在用于发现潜藏在金融数据库中的趋势与模式)；

Berry and Linoff(1997) 的下面这段话会让我们对 Data Mining 有更深刻的了解。分析报告给你的是后见之明 (hindsight)；统计分析给你的是先机(foresight)；Data Mining 给你识见(insight)。

从上述定义得出 数据采掘(data mining)所要处理的问题，就是在庞大的数据库中寻找出有价值的隐藏事件，加以分析，并将这些有意义的信息归纳成结构模式，作为企业在进行决策时之参考。此外，数据采掘看重的是数据库的再分析，包括模式的建构或是资料特征的判定，其主要目的就是要从数据库中发现先前关心却未曾获悉的有价值信息 (Hand, 1998)。事实上，数据采掘并不只是一种技术或是一套软件，而是数种专业技术的综合应用。

Data Mining 是指找寻隐藏在资料中的信息，如趋势 (Trend)、特征(Pattern)关系(Relationship)的过程，也就是从资料中发掘信息或知识(有人称为 Knowledge Discovery in Databases, KDD)，也有人称为"资料考古学"(Data Archaeology)、"资料模式分析"(Data Pattern Analysis)或"功能相依分析"(Functional Dependency Analysis)，目前已被许多研究人员视为结合数据库系统与机器学习技术的重要领域，许多产业界人士也认为此领域是一项能增加企业潜能的重要途径。这一领域蓬勃发展的原因是因为现代的企业已搜集了大量资料，包括市场、客户、供货商、竞争对手以及未来趋势等重要信息，但是数据的超载与无结构化，使得企业决策单位无法有效利用现存的资料，甚至会使决策行为产生混乱与误用。如果能通过数据采掘技术，从巨量的数据库中，采掘出不同的信息与知识出来，作为决策支持之用，就一定能成为企业竞争的优势。

目前，已有不少 Data Mining 的软件工具，有些销售得还相当火爆，但是，对于这种 Data Mining 的产品应该有一个正确的认识，就是它不是一个无所不能的魔法。它不是在那边监视你的资料的状况，然后告诉你说你的数据库里发生了某种特别的现象。也不是说有了 Data Mining 的工具，就连不了解业务、不了解资料所代表的意义、或是不了解统计原理的人也可以

此为试读，需要完整PDF请访问：www.ertongbook.com

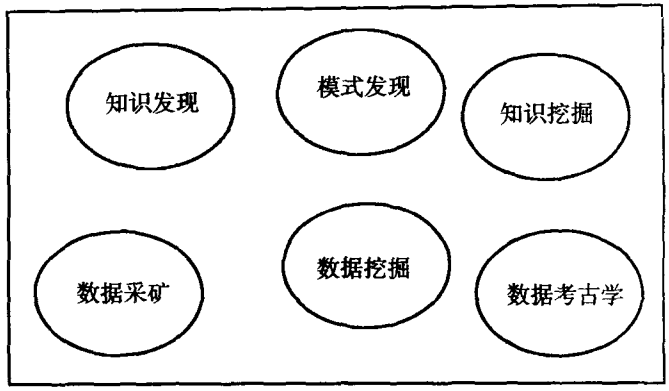

图 1.2.1 Data Mining 的不同名称

做 Data Mining。Data Mining所采掘出来的信息，也不是你可以不经确认，就可以照单全收应用到业务上的。事实上，Data Mining 工具是用来帮助业务分析策画人员从资料中发掘出各种假设（Hypothesis），但是它并不帮你确认（Verify）这些假设，也不帮你判断这些假设对你是否真有价值。

数据采掘的工作虽然是近年来数据库应用领域中相当热门和时髦的技术，但 Data Mining 使用的分析方法，如预测模型（回归、时间序列）、数据库分割（Database Segmentation）、连接分析（Link Analysis）、偏差侦测（Deviation Detection）等，美国政府从第二次世界大战以前，就在人口普查以及军事方面使用过。近几年来，随着信息科技超乎想象的发展，新工具的出现，例如关系式数据库、对象导向数据库、柔性计算理论（包括 Neural network、Fuzzy theory、Genetic Algorithms、Rough Set 等）、人工智能的应用（如知识工程、专家系统），以及网络通讯技术的发展，使从资料堆中采掘宝藏，常常能超越归纳的关系，使 Data Mining 成为企业智能的一部份。

还要说明的是，DATA MINING 和统计分析是有不同的。其实 Data Mining 技术中的 CART、CHAID 或模糊计算等等理论方法，也都是由统计学者根据统计理论所发展衍生，Data Mining 有相当大的比重是由高等统计学中的多变量分析所支撑。但是为什么 Data Mining 的出现会引发各领域的广泛注意呢？主要原因在相较于传统统计分析而言，Data Mining 有下列几项特性：

- 目标是海量数据的处理，不是一般意义上的统计分析；
- 分析的任务是找出特征、规律、联系，而不是验证；
- 必须多种技术结合，而不只是统计分析。

§1.3 Data Mining 综合的技术领域

Data Mining 综合了利用了表 1.3.1 中所述的技术领域。表中各技术领域都有专著论述，下面我们简单介绍一下这些技术在进行数据采掘中的作用。

表 1.3.1 **Data mining** 综合的技术领域

Database systems, Data Warehouses, OLAP
Parallel Processing
Machine learning
Visualization
Statistical and data analysis methods
Mathematical programming
High performance computing
Decision support

数据库管理研究人员利用智能化查询等数据搜索技术，使得 Data mining 工作变得容易和便利，以保证工作即时完成。数据仓库是另一种主要的数据库管理技术，它着重于整合不同的数据源，组织和管理海量数据以便进行数据分析和寻找规律，而以往数据库的设计着重于方便查找。

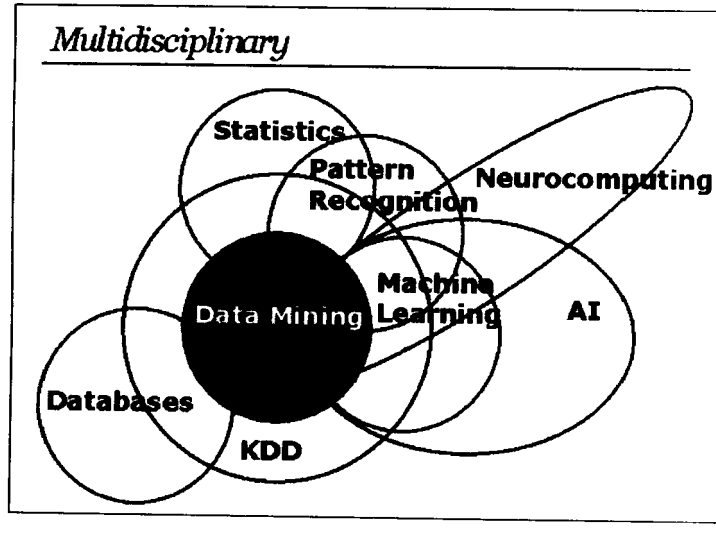

图 1.3.1 **Data Mining** 的综合技术领域

统计分析研究除了将许多的统计及数据分析方法用于数据采掘以外，还将统计方法和机器学习方法结合在一起，为数据采掘发展更复杂的统计分析工具（现在，许多的统计分析软件都有市场化的数据采掘工具产品）。机器学习的概念是让机器从已观测到的模式中学习各种各样的规则，然后再用这些规则来解决问题。但数据采掘通常面临的是海量的数据，所以，进行数据采掘工作时，必须将数据库管理与机器学习技术结合起来应用。

利用计算机可视化技术，可以进行交互式数据采掘工作。

决策支持系统是一系列工具和过程，用来帮助管理者进行决策并指导他们进行管理。

利用数学规划和高性能计算方法发展的技术能使得数据采掘算法顺利实现。当然，有了高性能的硬件会更好。

注意到，数据采掘正在不断发展，别的技术也不时地对数据采掘产生影响,如协作,代理和分布式目标管理技术等。

§1.4 Data Mining 的功能

这一节介绍一下 Data Mining 的功能，有时也称为 Data Mining 的任务。

随着 DM 的蓬勃发展，它的功能会越来越多。本节给出关于 DM 功能的一种分类方法，它分为下列五项：

- 分类 (Classification)
- 估计与预测(Estimation and Prediction)
- 聚类 (Clustering)
- 关联 (Association)和序列发现(Sequence Discovery)
- 描述 (Description)

这些功能大都可以用成熟的计量及统计分析方法来实现。现将它们的意义及可能使用的技巧简述如下：

1. 分类 (Classification)

所谓分类，就是按照分析对象的属性，建立类组(class)。它根据一些变量的数值做计算，再依照结果作分类。（计算的结果最后会是几个少数离散值,然后按不同值分类,例如将一组资料分为‘可能会响应’或是‘可能不会响应”两类）。分类常常被用来处理邮寄对象筛选的问题。我们会用一些已

经分好类的资料来研究它们的特征，然后再根据这些特征对其他未经分类或是新的数据做预测。这些我们用来寻找特征的已分类资料可能是来自我们的现有的历史性资料，或是将一个完整数据库做部份取样，再经由实际的运作来测试；譬如利用一个大的邮寄对象数据库的部份取样来建立一个分类模型（Classification Model），以后再利用这个模型来对数据库的其它资料或是新的资料作预测。例如，将信用申请者的风险属性，区分为高度风险申请者，中度风险申请者及低度风险申请者。使用的技术有决策树（decision tree），记忆基础推理（memory-based reasoning）等。

2．聚类（clustering）

面对海量的资料，首要的任务是将它合理地归类。如果已知要求，于是对资料可以设问，按回答的不同给予分类，这就是上面所说的分类。如果事先没有任何要求，象全国各地环境监测的资料，就只能按资料反映的情况，比较接近的划归一类，这种归类的方法称之为聚类（clustering）聚类分析的算法往往按距离的远近来归类，也可以按相似程度的大小来归类。只有合理地聚类后，每一类内就可以找出有关的特征，否则是难于发现真正有用的信息。很自然，不同类型的问题可以给出不同的聚类原则，从而找到不同的特征。例如世界上的居民可以按民族归类，可以按肤色归类，可以按国家归类，也可以按宗教信仰归类，……这些不同的聚类原则自然会找出不同的特征。

3．估计与预测（Estimation and Prediction

估计（estimation）是根据已有的长期累积的资料来推测某一属性未知的真值。例如按照信用卡申请者的教育程度、行为和性别来推估其信用卡的消费量。使用的技巧包括统计方法中的相关分析、回归分析及人工神经网络方法。预测（prediction）是根据对象属性之过去观察值来估计该属性未来之值。例如，由顾客过去的刷卡消费量来预测其未来刷卡消费量。使用的技巧包括回归分析、时间序列分析及人工神经网络方法等。

回归是使用一系列的现有数值来预测一个定量指标的可能值。若将范围扩大亦可利用逻辑斯蒂回归（Logistic Regression）来预测定性变量，特别在广泛运用现代分析技术如人工神经网络或决策树理论等工具，预测的模式已不受传统线性的局限，在预测的功能上大大增加了选择工具的弹性与应用范围的广度。

此为试读，需要完整PDF请访问：www.ertongbook.com

时间序列预测是用指标本身现有的历史数值来预测未来的数值。Time-Series Forecasting 的特点在于它所分析的数值都与时间有关，可以处理有关时间的一些特性，譬如时间的阶段性（例如每个礼拜五个或六个工作天）、季节性、节日、以及其它的一些特别因素如过去与未来的关连性有多少等等。

4. 关联(Association)和序列发现(Sequence Discovery)

关联是要找出在某一事件或是资料中会同时出现的东西。关联（Association)主要是要找出下面这样的信息 如果 A 是某一事件的一部份，则 B 也出现在该事件中的机率有 X%。（例如：如果一个顾客买了低脂乳酪，那么这个顾客同时也买低脂牛奶的机率是 85%。）确定那些相关对象应该放在一起。例如超市中相关之盥洗用品(牙刷、牙膏和牙线),放在同一间货架上。在客户行销系统上，此种功能可以用来确认交叉销售（cross-selling)的机会，以设计出吸引人的产品群组。

序列发现 Sequence Discovery)与关联(Association)关系很密切,所不同的是序列发现 Sequence Discovery)中相关的对象是以时间区分开来(例如：如果做了 X 手术,则 Y 病菌在手术后感染的机率是 45%。又例如：如果 A 股票在某一天上涨 12%，而且当天股市加权指数下降，则 B 股票在两天之内上涨的机率是 68%)。

5. 描述(Description)

描述的功能是对复杂的数据库提供简要的描述。最简单的例子就是变量的均值和方差。这个功能的主要目的是为了在使用别的功能时对数据先有较好的了解。在建立任何模型之前先做数据描述的工作是十分重要的，因为这会告诉我们怎样去建模。许多的商业数据采掘软件包也提供有用的画图软件来帮你对数据作可视化处理。另外，经理们经常使用更加复杂的采掘工具(比如 market basket analysis , tree-based models ）来更好地理解数据和开发模型。

目前已有各种各样的方法（算法）来实现 Data Mining 的上述功能（任务)。这些方法包括诸如回归分析、时间序列、判别分析、因子分析和聚类分析等一些统计分析方法；也有粗集(rough set),模糊逻辑(fuzzy logic)，或人工神经网络(Neural Network)、决策树理论(Decision Trees)以及规则归纳法(Rules Induction)为基础的方法等,例如,表 1.4.1 就列出了实现 Data

mining 各项功能所常用的一些工具。

表 1.4.1 Data mining 功能及常用的工具举例

功　能	采掘工具举例	应用举例
Classification	Neural networks, logistic regression, tree-based models, decision tree, memory-based reasoning	Mailing decisions, target marketing, credit decisions
Clustering	Neural networks, multivariate statistics,	Segmentation
Estimation and Prediction	Linear and nonlinear regression, neural networks, hazard models, collaborative filtering	Customer scoring, sales forecasting, customer service, various marketing decision models,
Association	Market basket analysis, set theory, link analysis	Promotion design, shelf space allocation,
Description	Traditional statistics, market basket analysis, tree-based models	Exploratory data analysis

§1.5 Data Mining 的应用

国际上 Data Mining 应用的行业包括了金融业、电信业、网络相关行业、零售商、制造业、医疗保健及制药业等。为了便于查找，下面我们给出在网上可以查到的数据采掘行业应用分类：

- 综合 General

  客户分类(Customer Segmentation)

  Retention／Acquisition

  Database marketing

  Lifetime value of a customer

  Cross selling

- 银行业

  Banking

  Credit scoring

  Credit Card Fraud Detection Portfolio Analysis

  Cash Planning

- 保险与保健行业

Insurance & Health Care

Claim Analysis

Fraudulent Behavior

- 电信行业

Telecommunications

Call Behaviour Analysis

Churn Management

Fraud Detection

- 零售/行销行业

Retail/Marketing

Market Basket Analysis

Category Management

Credit Scoring

- Manufacturing and Utilities

Process Management

Demand Patterns

Capacity Planning

Inventory Planning

表 1.5.1 数据采掘应用领域分类表

Applications of Data Mining		
Customer-focused	Operations-focused	Research-focused
Life-time Value	Profitability Analysis	Combinatorial Chemistry
Market-Basket Analysis	Pricing	Genetic Research
Profiling & Segmentation	Fraud Detection	Epidemiology
Retention	Risk Assessment	
Target Market	Portfolio Management	
Acquisition	Employee Turnover	
Knowledge Portal	Cash Management	
Cross-Selling	Production Efficiency	
Campaign Management	Network Performance	
E-Commerce	Network Performance	
	Manufacturing Processes	

利用 Data mining，企业能够从巨大数据库中采掘到从未发现的信息，并从使用中获利。例如，一个发行管理共同基金(mutual funds)的企业要采掘出它会有的潜在客户，就需要整合客户的帐户、人口统计、生活方式等资料。也就是说要能把数据库中人口资料细分成为一些关键子集合：都市化情况、婚姻状态、家庭所得、年龄、风险偏好等。最后，依据数据采掘的结果，按不同的类从事促销活动，成功的把共同基金推展到市场。又如拥有汽车的新婚夫妻很可能购买儿童专用汽车椅，这个现象很容易被理解，并不需要应用到 Data Mining ；但如考虑到另一个问题，这些夫妻会购买何种颜色的儿童专用汽车椅？这个问题就困难许多，需要用到许多的变量来预测可能的结果，这时 Data Mining 便提供了非常大的帮助。

目前国外企业界把 Data Mining 应用在许多领域。例如，行销、财务、银行、制造厂、通讯等。并且在产学合作下，发展出许多实用的系统，例如 MDT、Coverstory and Spotlight、NichWork visualization system、LBS、FALCON、FAIS、NYNEX、TASA 等等。这些数据采掘系统，应用非常广泛，例如有一个应用在行销领域的例子：经由记录客户的消费记录与采购路线，超级市场可以设计出更吸引顾客购买的环境。根据数据采掘出来的信息，现在超级市场的厨房用品，是按照女性的视线高度来摆放。根据研究指出：美国妇女的视线高度是 150 公分左右，男性是 163 公分左右，而最舒适的视线角度是视线高度以下 15 度左右，所以最好的货品陈列位置是在 130 至 135 公分之间。在商业上，有许多特征是很难理解的，但若了解到这些信息，就会增加企业的竞争能力。一般行销部门较典型的问题是：

- 除了已经购买的产品外，我的客户还可能购买哪些产品？
- 我的最有价值客户中，他们的共通特征为何？
- 当我的客户有可能转向其它竞争同业时，哪些变量能测量出这样的信息？

国外企业界实际发展 Data Mining 时，效能并不能预期，因为有许多因素影响着。例如，不充足的教育训练、不适当的支持工具、资料的无效性、过于丰富的模式(patterns)、易变与具有时间性资料或空间导向资料(spatially oriented data)、复杂的资料结构、资料的可度量性(scalability)等 这说明资料与知识的发掘是一项信息技术程度很高的工作，面对易变的环境，没有现成的模型马上可用，也不要期望按一定的计算程序即能成功。因此，我们要认识到一些潜在的因素，如资料取舍、实体关系性、数量多寡、复杂性、数据质量、变迁、专家意见等种种因素，才能做好数据采掘工作。

此为试读，需要完整PDF请访问：www.ertongbook.com

Data Mining 在各领域的应用非常广泛，只要该产业拥有具分析价值的数据仓库或数据库，就可利用 Mining 工具进行有目的的分析。在国外一般较常见的应用案例多发生在零售业、直销界、制造业、财务金融保险、通讯业以及医疗服务等。

Data Mining 对每个公司来说都是一 :计划，而将它列为 高投资的计划，要聘请各家公司到底用 做什么样的事是相 当不容易的。根据 Two Crows Corp. 最 ata Mining 主要的 三个应用方式都分别是：Customer Profil rketing、以及 Market-Basket Analysis。

在 Customer Profiling 方面，即如何获得新顾客？我们希望找出客户的一些共同的特征，希望能藉此预测哪些人可能成为我们的客户，以帮助行销人员找到正确的行销对象。Data Mining 可以从现有客户资料中找出他们的特征，再利用这些特征到潜在客户数据库里去筛选出可能成为找们客户的名单，作为行销人员推销的对象。行销人员就可以针对这些名单寄发广告资料，既可以降低成本，又提高了行销的成功率。

Market-Basket Analysis 主要是用来帮助零售业者了解客户的消费行为，即如何增加顾客的消费额？譬如哪些产品客户会一起购买，或是客户在买了某一样产品之后，在多长时间之内可能购买另一产品等等，利用关联性产品销售(cross-selling)和连贯性销售(Continuity-Selling)方法，来提高客户的终生价值(Live Time Value)。利用 Data Mining，零售业者可以更有效的决定进货量、库存量，以及在店里要如何摆设货品，同时也可以用来评估店里促销活动的成效。

客户关系管理是 Data Mining 的另一个常见的应用方式，即如何留住他们？我们可以由一些原本是我们的客户，后来却转向成为我们竞争对手的客户，分析他们的特征，再根据这些特征到现有客户资料中找出有可能转向的客户，然后公司必须设计一些方法将他们留住，因为毕竟找一个新客户的成本要比留住一个原有客户的成本要高出许多。

在销售资料中采掘顾客的消费习性，很容易由交易纪录找出顾客偏好的产品组合，还可找出流失顾客的特征，确定推出新产品的时机点，还可结合基本资料，并依品牌价值等级的高低来区分顾客，进而达到差异化行销的目的 制造业对 Data Mining 的需求多运用在质量管理方面，由制造过程中找出影响产品品质最重要的因素，来提高作业流程的效率。

近来国外的电话公司、信用卡公司、保险公司、股票交易商、以及政府单

位对于诈欺行为的侦查（Fraud Detection）比较关注，这些行业每年因为诈欺行为而造成的损失都非常可观。 Data Mining 可以从一些信用不良的客户资料中找出相似特征并预测可能的诈欺交易，从而达到减少损失的目的财务金融业可以利用 Data Mining 来分析市场动向，并预测个别公司的营运以及股价走向。Data Mining 的另一个独特的用法是在医疗业，用来预测手术、用药的疗效、以及医院服务、管理上的效益提高。

下面列举一些 Data Mining 的在国外别的方面运用的类型：

- 如果采用不同的价格策略，是否能增加市场占有率？
- 让我们获利高的客户们有什么共同的特征？
- 如何认定客户的信用风险状况？
- 如何设计更好的保险产品来吸引客户，让客户满意？
- 一个经纪人在一个星期中应该可以卖出多少共同基金？
- 根据以往审核的资料，寻找核发信用卡的规则
- 在 NBA 球赛资料中，找出球员的强弱点
- 从消费及缴费资料中，预警信用卡呆帐可能
- 从通话记录资料中，预警盗打电话可能
- 从宇宙飞船拍摄的影像资料，找寻星球上的火山
- 星际星体分类

等等。

第二章　Data Mining 数据仓库的设置

在进行数据采掘的工作之前，如何准备好你的数据库，也就是可供“采矿”的数据库，这是最令人关心的一件事。什么是数据仓库呢？而统计数据库和传统的数据库有什么不同呢？首先让我们来了解什么是数据仓库。

§2.1　数据仓库与数据超市

1. 数据仓库与数据超市（Data Warehouse and Data Markets）

要将庞大的资料转换为有用的信息，必须先有效地整储资料。随着科技的进步，功能完善的数据库系统就成了能很好汇集资料的工具。 META Group 在 1996 年 2 月提出的一份报告指出，95％的企业为了想提供信息给决策来用，都将建立数据超市（DataMart ,也就是小型的数据仓库）。

“数据仓库”,简单的说,就是搜集来自其它系统的有用资料,存放在一整合的储存区内。其实就是一个经过处理整合，容量特别大的关系型数据库，用以储存决策支持系统（ Decision Support System ）所需的资料，供分析使用。从信息技术的角度来看，数据仓库的目标是组织好资料，在恰当的时间,将恰当的资料交给恰当的‘人’。这是一种持续性的作业,不是做一次就结束的，与过去以交易为主（ transaction-oriented ）的系统就开发的方式而言也有所不同。而决策支持系统的资料来源，既包含企业内部的各个业务系统，还有外界提供的资料。数据仓库常见的问题就是如何将各个不同来源的数据库整合在一起，并且迅速地将资料搬运到数据仓库内　数据仓库（Data Warehouse）可以让您按需取得内部资料以及市场资料，得知您企业对市场的反应情况。这样的市场分析可以协助您在利润较高的市场投入更多资源，同时在竞争性高的市场中寻求适当的切入点。

许多企业选择较小型的部门或工作群组的 Data Warehouse 或称 Data-Mart,以兼顾低成本与快速上线能力。如果日后有需要,该 DataMart 可以

予以扩充，在资料量或 DataMart 数目扩充后，就可以成为整个企业的 Data Warehouse。一般而言，数据仓库必须能从多种异质(Heterogeneous)资料源中撷取作业资料，并将这些资料转换成 Informational Data(可供分析利用的加工资料）后储存在 DataMart 中。

2. 数据仓库与操作系统（Data Warehouse Vs. Operational System）

数据仓库与数据库的设计以及对系统预期的作用是很不同的。以往操作系统的组织架构的设计是为了支持线上交易处理（OnLine Transactional Processing, OLTP),例如订单输入、银行的存取款等作业,需具备快速的交易处理功能。而数据仓库的组织架构设计，则是随特定的主题（subject）而定,例如行销、产品、信用风险等。两者的比较列于表 2.1.1。

表 2.1.1 传统操作系统与数据仓库的比较表

	OLTP	Data Warehouse Systems
组织架构	依执行的交易而定(如订单处理、进销存处理等)	依主题而定(如产品、顾客等)
使用者的数量	大量的同时上线的使用者	经常的使用者人数可能只有数百人或更少
交易的数量/处理时间	交易时间通常很短,只存取数笔资料;交易处理在数秒间完成	处理查询的交易时间可能长达数分钟到数小时;查询次数少;但很复杂,通常是检索许多笔资料,并需要结合(join)多个数据库的表格。
数据库的大小	OLTP 的数据库通常小于数据仓库;因为历史资料会移出,不在线上	数据仓库的资料量通常远大于 OLTP 数据库,因为其中整合了数种资料来源(包括 OLTP),也包括历史资料
结构	高度正规化的数据结构,由许多表格组成,每一个表格设计是愈少字段愈好	表格数量较少(因为有特定的主题),但字段的数量很多
历史资料	线上只维持在较短期间的资料	维护的资料可能回溯到数年以前
更新处理	当新交易被处理或输入时,持续性的更新(近乎实时),	阶段性的更新,通常以批次的方式(例如为了反应 OLTP 资料的变更)

由表 2.1.1可看出两者的差异性。因此在设置数据仓库时，所设计的架构中，对数据仓库中查询的响应时间可以大幅改善，而特定业务的使用者也能够依业务的决策需要，存取大量但只与自己相关，深入各种细节的适当

的资料，而不影响线上交易系统的效率或引起任何安全上的麻烦

数据仓库的架构，有两种基本的型态：企业级数据仓库与数据超市。企业级数据仓库所包含的是全企业的信息，这些信息是为了整体的资料分析而整合众多个操作系统的资料来源。一般而言，是由数个主题领域所组成，例如，客户、产品加上业务等，可用于战术(tactical)与策略(strategic)的决策支持。企业级数据仓库的信息包括实时的详细信息，也有汇总的信息，数据库大小的范围可能从 50 gigabytes (GB) 到 1 terabyte (TB)。企业级数据仓库的设置与管理往往非常昂贵且耗时；建立的方法通常是从上到下 (top down)由统筹的信息服务单位主导。

数据超市是企业级数据仓库的子集，建置的目的是便于企业中个别的部门或单位使用。与企业级数据仓库不同的是，数据超市通常只为了特定的决策支持的应用程序或使用群组，是由下到上(bottom up)利用部门的资源来设置。数据超市通常只有特定主题的汇总或详细资料。而数据超市中的资料却可以是企业级数据仓库的一个子集（独立的数据超市）或者可能直接使用来自运作中的资料来源。无论是数据仓库或数据超市，其组成与维护的程序是相同的，使用的技术部分也都类似。

3. 数据仓库的组件

一般来说，数据仓库中一定包括以下的组件：

- 运作的资料来源 (Operational data sources)
- 设计／开发工具
- 资料抽取(data extraction) 与转换 (transformation)工具
- 数据库管理系统(DBMS)
- 资料存取与分析工具
- 系统管理工具

为协助企业快速进入数据仓库系统，并降低企业采用数据仓库的风险，主要的数据库厂商纷纷推出已被广大企业肯定的数据超市 (data mart)产品。由于已趋成熟的数据仓库项目较为昂贵、复杂，部分客户为降低风险，转而热衷于数据超市的解决方案。基本上，对于部门别的数据仓库来说，数据超市是一个较为低廉，也更容易扩充的解决方案。

§2.2 统计数据库与数据采掘

在了解数据仓库的架构及功能后，就知道要使用数据库中的资料，必须有合适的存取及分析工具做为数据库系统前后端的接口，作为使用者和数据库间的交谈信道。以往数据库系统所能提供的资料查询、运算、逻辑判断及报表输出等的功能，对现在相当庞大的数据而言，这些已经是数据仓库的基本功能，现在再加上数据采掘技术的引进，高品质资料的需求已经日益殷切，随着网际网络的发展，网络数据库的概念日益成形，企业如何处理如此庞大的资料，是现今企业所必须面临的课题。

在传统的环境中，会有如此庞大数据库的机构，大概就属政府机关了，因为政府必须汇整全国的统计信息，以提供决策及施政建设的参考依据，政府中统计数据库的概念便会应运而生。而政府就如同一个大的企业，虽然运作的方式有所不同，但决策的思考模式却是类似的。统计数据库系统和传统数据库系统最大的不同，在于其针对各种统计资料的检误及所产生的功能，是传统数据库系统所没有的，因此提升了数据处理的效率及数据质量，这些经由统计数据库系统所产生的资料，正好可以作为数据采掘所需的初级资料，因此如果你要进行数据采掘的工作，如何加强数据库系统的功能，就是进行数据真正采掘的准备工作之一。我们以一个现有统计数据库的系统，来说明什么是统计数据库系统。

1. 统计数据库硬件

统计数据库的架构，由下方的图片可以得知，就硬件设备而言，和传统数据库并无太大的差异，仍是需要透过数据库的主机来做资料存取的动作

2. 统计数据库功能

与传统数据库差异较大的是功能方面，通常统计数据库的功能如下：

(1)弹性输入功能

- 输入格式、顺序可由使用者自行定义
- 检核公式可依字典属性自动产生在线检核
- 可输入前后注记
- 资料一致性的处理
- 最佳化储存结构

此为试读，需要完整PDF请访问：www.ertongbook.com