HZ BOOKS

微软公司核心技术书库

Microsoft

SQL Server 2000
数据挖掘技术指南

(美) Claude Seidman 著　刘 艺　王鲁军　蒋丹丹　等译

Data Mining With Microsoft SQL Server 2000 Technical Reference

机械工业出版社
China Machine Press

微软公司核心技术书库

SQL Server 2000 数据挖掘技术指南

（美） Claude Seidman 著

刘 艺 王鲁军 蒋丹丹 等译

淡菊资讯工作室 审校

机械工业出版社
China Machine Press

本书讲述了数据挖掘及其基础理论，并通过两个数据库实例介绍如何建立数据挖掘模型。主要内容包括：数据挖掘介绍、数据挖掘方法、数据挖掘应用编程等等。本书内容全面、深入浅出，集学术性和实用性于一体，适用于从事数据挖掘的IT工作者。

Claude Seidman:Data Mining with Microsoft SQL Server 2000 Technical Reference.
Copyright © 2002 by Microsoft Corporation.
Original English language edition copyright © 2001 by Microsoft Corporation; Published by arrangement with the original publisher, Microsoft Press, a division of Microsoft Corporation, Redmond, Washington, U.S.A. All rights reserved.

本书中文简体字版由美国微软出版社授权机械工业出版社出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。
版权所有，侵权必究。

本书版权登记号：图字：01-2001-4445

图书在版编目（CIP）数据

SQL Server 2000 数据挖掘技术指南/（美）希德曼（Seidman, C.）著；刘艺等译. -北京：机械工业出版社，2002.1
（微软公司核心技术书库）
书名原文：Data Mining with Microsoft SQL Server 2000 Technical Reference
ISBN 7-111-09519-7

Ⅰ.S… Ⅱ.①希… ②刘… Ⅲ.关系数据库-数据库管理系统，SQL Server 2000 Ⅳ.TP311.138

中国版本图书馆CIP数据核字（2001）第078838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：张鸿斌
北京第二外国语学院印刷厂印刷·新华书店北京发行所发行
2002年1月第1版第1次印刷
787mm×1092mm 1/16·18.25印张
印数：0 001-5 000册
定价：35.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

N1532-06

译 者 序

数据挖掘技术是一门非常有用、也非常有趣的数据处理技术。它通过一些专门的算法，借助计算机强大的运算能力，从浩瀚的数据海洋中揭示出鲜为人知的秘密规律，从而对未来进行有限和有效的预测。

通过微软SQL Server 2000，我们可以充分享用Analysis Services提供的数据仓库和数据挖掘服务。微软数据挖掘技术集成了大量的数据挖掘工具，隐匿了统计学计算公式的复杂性，使得用户可以轻松实现自己的数据挖掘需要。本书正是为这一需要而编写的一本关于微软数据挖掘的力作。本书作者Claude Seidman是一位数据库开发员、数据库管理员以及培训专家，有超过14年的实际应用开发经验（这正是某些国内作者所缺少的）。在这本书中，处处可以体现作者丰富的应用经验，以及严谨朴实的写作风格，本书是一本集学术性和实用性于一体的好书。

本书的作者假定读者已经熟悉关系型数据库并略知OLAP，所以本书并没有详细介绍SQL Server 2000及其Analysis Services。如果读者需要了解这方面的知识，请参阅我编写的《全面精通SQL Server 2000》（中国水利水电出版社 2001年出版）有关数据仓库和OLAP的章节。

参加本书翻译工作的人员除封面署名外还有吕常魁、卢峰等，同时也要感谢王春生、赵厚良、吴英、洪蕾、谈亮对我们的大力支持。由于时间仓促，在翻译中难免有不妥之处，恳请读者不吝赐教，我们将表示感谢。

我的电子邮件：my_reader@sina.com或liuyi@chinanewdream.com 。

刘 艺

2001年8月2日于南京

前　言

今天，当我们在amazon.com网站购书，点击因特网上的广告条，或是在邮件中接受预先核准的信用卡，我们不可避免地注意到大公司是如何了解我们的喜好的。在线图书销售商似乎“知道”我喜欢读什么书，听什么音乐。他们的广告条简直就像在呼唤我的名字，当我查看水中呼吸器传动装置的广告时，我不得不惊讶他们怎么知道我喜欢使用水中呼吸器潜水。我提到这些例子是为了说明数据挖掘技术的应用，并且表明这个新技术使用得多么广泛。数据挖掘现在比以前用得更为广泛，几乎对任何需要了解其利益的业务都是可用的。

如果你在因特网搜索引擎中键入“data mining”（数据挖掘），得到的结果多得你都没有时间阅读完。对公司来说访问它们大量的数据，需要快速的计算机、便宜和无限的存储空间以及更好的通信技术。这样才能精确定位数据和收集数据，并使之具有意义。在业务中使用数据挖掘来查找所有的信息，从在线购物模式到信用历史，精明的商家正在将这一技术转变成黄金。

微软的SQL Server 2000就像许多其他的大型关系型数据库系统那样，从不断增长的廉价存储媒介以及越来越强大的主流服务器中受益。企业级存储媒介不再是问题，公司存储着多年的详细交易信息在销售点的系统中，还存储着因特网日志记录甚至是音频和视频数据流。

自动的数据挖掘提供了带有包装好统计公式的工具，使得数据库专家不必理解挖掘工作背后的统计学就能使用这些工具。使用这些工具，数据挖掘操作员可以点击按钮并处理上亿字节信息，如果不是关于人口统计、购物习惯、经济指标的千亿字节信息，要回答“谁最喜欢我的电话售货活动，为什么？”这样的问题，喝一杯咖啡的时间就可以处理完毕。

正如在本书中看到的那样，微软数据挖掘工具几乎就这样简单易用。本书讲述了数据挖掘及其基础理论，并通过两个数据库实例教你如何建立自己的数据挖掘模型。

本书读者对象

本书适用于从事设计、实现、使用Analysis Services并打算使用微软数据挖掘工具的IT工作者。本书假定读者熟悉关系型数据库并有点熟悉OLAP数据库。系统设计员能够使用本书来理解Analysis Services。数据库管理员能够使用本书来理解如何设置数据挖掘环境，包括关系型数据库的数据源。OLAP管理员能够使用本书来理解如何从OLAP中提供实例数据用于数据挖掘模型，以及如何使用该模型增强现有OLAP立方体的分析能力。另外，应用程序开发人员将能使用本书来编写执行管理任务及预测查询的前端应用程序。

本书主要内容

第一部分“数据挖掘介绍”概括了数据挖掘活动取得成功的途径并说明了其重要性。数

此为试读，需要完整PDF请访问：www.ertongbook.com

据挖掘是一个在理论上和实践中都很丰富的主题，已经有许多介绍性读物涉及这一主题。本书的前几章既从理论上也从实践中介绍了数据挖掘。第一部分讨论了成为逻辑基础的关键要素，以及包括微软Analysis Services在内的数据挖掘工具背后的统计学基础。同时还讨论了我们为什么使用数据挖掘，以及数据挖掘处理是如何工作的。

第二部分“数据挖掘方法”从更加技术化和特定产品的角度来观察数据挖掘。介绍了如何使用所有的微软向导以及其他交互式工具来设计和创建数据挖掘模型。你将学习如何创建“挖掘就绪”表以及有效的数据挖掘模型。你还将学会如何“训练”模型和解释结果以得到对数据的更深理解。

第三部分“数据挖掘应用编程”是供那些创建应用程序的开发人员参考的，这些应用程序使用Analysis Services引擎或PivotTable Services来管理数据挖掘模型。对用于管理的应用程序，开发者可以学会如何创建带有和Analysis Manager同样功能的应用程序。对于依赖现存模型进行预测的应用程序，开发者可以学习如何通过编程来提交预测查询。这是微软数据挖掘最令人激动的一个方面，因为这样你就能够使用微软的Visual Basic、Visual C++、C#或ASP来创建复杂的前端数据挖掘应用程序。

了解了这些之后，让我们开始学习数据挖掘吧。

目　录

第三部分 数据挖掘应用编程

附 录

第一部分 数据挖掘介绍

让计算机去处理你的数据，你不知道它们将会得出什么结果——这就是全部要点。

Edmund X. DeJesus, BYTE杂志高级编辑

数据挖掘，也称为知识挖掘（Knowledge Discovery，KD），是对巨大的数据集进行寻找和分析的计算机辅助处理过程，在这一过程中发现先前未曾发现的模式，然后从这些数据中发掘某些内涵信息，包括描述过去和预测未来趋势的信息。

在这一部分，我将概括地讨论数据挖掘——数据挖掘是什么，数据挖掘不是什么。我还要仔细回顾许多有关数据挖掘方法论的重要的原理和定义，包括数据挖掘模型的角色、统计学和数学的运算法则。我还要解释数据挖掘如何在Analysis Services体系中配置，以及它和SQL Server 2000关系型数据库引擎、OLAP引擎如何互相作用。

第1章 了解数据挖掘

最近我和一个汽车销售公司的CEO和CIO探讨它们公司的数据库。我们主要是想找到存储它们公司大量的销售数据和重要的公司信息的最佳方法。当我们细看数据库中成千上万条记录时，那位CEO满怀敬畏的说："我打赌这里的信息有很多非常有价值，如果我们有一千年的时间来分析它们。"这句话非常典型地汇总了很多公司领导、天文学家、医生和金融商的意见，它们存有大量的具有潜在价值的数据却苦于没有好的方法让这些数据为他们服务。数据挖掘充分利用了当今服务器的运算能力，将堆积成山的数据转换为有用的信息。

计算机不断增长的物理存储能力，再加上强有力的处理能力，使得仅仅数年前看来难以想象的纷繁复杂的数据分析成为可能。不久以前，仅仅是大型公司和大学才有超级计算机和大型主机用于实现有用的数据挖掘任务。相对于超级计算机的价格而言，随着服务器的功能更强大，价格更便宜，较小的公司也能利用服务器的强大功能挖掘它们存储的数据，赢得市场竞争中的优势。为了以创新主义的、超现实主义的种种方法来挖掘数据，我们必须首先理解那些可用的技术，以及对特定的数据存储如何应用这些技术。

1.1 什么是数据挖掘

数据挖掘一个发现过程，它在非常大的数据库中发掘隐藏其中的有意义的某种模式和关系。对表和记录的浏览几乎不能引导你得到有用的模式，即通过自动处理对数据作典型分析，也就是通常在数据挖掘术语中称为的知识挖掘（Knowledge Discovery，KD）。知识挖掘是数据挖掘

此为试读，需要完整PDF请访问：www.ertongbook.com

的组成部分，利用计算机的强大功能加上操作员的人类天生能力，最终的目标就是得到可视化的显而易见的模型。对于自动的数据挖掘，计算机只是发掘出数据中存在的模型和趋势，而负责利用这些挖掘结果的人要确定哪些模型是真正相关的、有用的。

数据挖掘能够发现描述性的和预见性的信息。你选择哪种类型的信息用于数据挖掘，很大程度上取决于你想用它的结论完成什么工作。当你寻求预测信息时，目标是从信息中挖掘出能够提供关于未来事件的线索，例如，如果一个轿车经销商要知道她能从一辆跑了68 000英里的1998 Ford Mustang赢利多少，她就是在寻求预测信息。如果该经销商已经保存了几年内销售的价格记录，那么这一数据仓库可以进行挖掘，然后用来确定买入价格和预期销售价格。将诸如轿车的已使用时间和型号之类的变量输入计算机，从先前的销售情况将得到一个预期价格。

同样的一个轿车经销商如果要给她的本年度创最高利润的销售员一个夏威夷奖励游（注意，我说的是创最高利润，而非总销售额最高的销售员），那么，她就需要关于其雇员的销售历史的描述性信息。这种信息就其本身而言并无预测价值，但它确实以某种难以通过普通方法看出的模式准确描述了以前的事件，因此为决策提供了建立在新发现的关系基础上的机会。

1.2　为何使用数据挖掘

数据挖掘是一种提供商业优势的活动，也是对于涉及到对公司数据库内部所蕴涵的信息进行发掘的那些不断提出问题的解决方案：

- 增长的磁盘空间。
- 关系型数据库管理系统（RDBMS）引擎的不断提高。
- 在线事务分析处理（OLAP）的增强。

在信息技术领域，这对任何人来说都不奇怪，磁盘空间越来越大，同时也越来越便宜。

> **你磁盘上的生命**
>
> 英国电信正在进行将一个人的所见所闻全部存储在磁盘上的研究！引用英国电信的未来学家Ian Pearson的话——“超过80年的生命，我们要处理10兆兆字节……”，尽管这听起来超乎现实，但是它表明磁盘的存储空间已不再是数据挖掘者所顾虑的事情了。

磁盘的存储空间越来越便宜，数据存储对存储空间已不再考虑。对于更多的日常案例来说，你只需看看银行和信用卡公司。它们通常是把每一个发生在一个帐户生命周期上的事务进行存储、归档。很显然，这些公司希望利用这些数据充分研究它们的顾客，从而发现一个理想顾客的特征。考虑到那些大信用卡公司一个月要打印几十万张记录单，自动的数据挖掘是在它们硬盘上储存的大堆信息中寻找任何一种有意义的信息的唯一希望。

为了对公司硬盘上爆炸式增长的数据作出响应，RDBMS引擎对查询的响应时间已经有了大幅度的提高。可是，一旦数据存储了较长时期，它对于基于总体信息的查询比起单条信息来说就更为有用。例如，一条较大的连锁商店，就它的每个区域和每个产品类型的销售总额与其单个产品销售分析来说，前者显然更有意义。尽管大部分遵从SQL的RDBMS引擎可以在这个层次上聚簇数据，但是，它并不是这些引擎最直接的最优化处理。进行查询时，这些引擎按给定的条件进行专门的优化以查找数据集，无须始终执行数学计算。

OLAP数据库原来是为了减轻频繁使用的计算集合的问题特别创建的。与RDBMS不同，OLAP的设计是用来以这样一种方式预先计算和存储数据集，该方式允许查询简单返回预处理表中的一些结果。这样不仅消除了对昂贵的处理功能的需求也充分利用了丰富的磁盘存储空间。

除了OLAP系统提供的存储便利之外，特别的存储和显示功能使得用户可以通过使用上下滚动的聚簇视图来访问大量的数据档案。虽然，OLAP确实给许多公司提供了一种较好的处理信息的方法，但是，它无法告诉这些公司去找什么。例如，大型的汽车经销商花费了大量的时间分析他们所销售的每一种样式和型号的汽车的利润率，底线决定了来年的样品陈列室。经销商们分析那些明显的因素，如价格、里程表读数（如果是二手车）、型号和样式，但最成功的经销商雇用那些有经验的职员，他们可以识别那些不明显但却同样非常重要的因素，如颜色，马力以及变速器类型。

即使是最有经验的商人也会犯一些低级错误，因为有许多难以估计的隐藏的因素都会起作用。这些因素可能包括小车在停车场停留的时间；车子在停车场与其他车子位置关系的安排（如果，你将跑车停在大卡车的旁边，那些小车将很难被发现）；以及天气、季节乃至道琼斯工业指数。如你所见，这一清单几乎没完没了。事实上，如果考虑所有的因素那简直就需要一个政府经济预测组，这足以让绝大部分商人对深入研究他们的数据望而却步。

OLAP作为数据挖掘工具，尽管功能强大，它还是需要初始的假设以给出导航数据努力的方向。之后数据又被用来证实或推翻这些努力背后的理论。在包含大量表和列的数据库中，使用OLAP常常导致试凑法的运用，这不仅花费大量的时间，而且往往是产生一些普普通通的结果。由于使用了手工的关系型数据库挖掘，使得用OLAP发现模式和有意义的关系常常受到编写假设条件的人和所分配的发掘数据时间的限制。OLAP数据挖掘的优势在于它允许计算机在分析中检测每一个能想象得到的因素，这些因素可能影响输出，并导出这一分析的结果。这一过程本质上是找出一个实例的各特征之间的关系，也就是可以描述实例共性特征的那些。如果我们要在汽车经销商的早期实例中应用数据挖掘过程，那么在预测一个结果比如价格范围时，实例中涉及车辆销售的每一个特征因素或变量，如颜色和样式都应考虑，并进行分析，所有这些都可能会被认为是可以用来推测每一辆车各特征因素的关系以及车价的一个结果。

无论做什么用，自动处理都需要向经销商提供各种可能影响结果的情况，即使是那些一眼看起来没什么关系的因素。例如，根据数据情况来看，经销商可能会发现将粉红福特卡车放在场地角落旧Yugos车旁时要卖的好一些，但仅仅是星期三，而且不是十一月的星期三。因为涉及到的因素众多，也因为具有一种倾向，先入为主地把不可能的因素自然而然排除，这一类关联极易被人忽略，而计算机却不会忽略它们。

1.3 当前数据挖掘是如何使用的

数据挖掘对于那些收集了大量历史资料的机构来说尤其具有价值。银行、保险公司、信用卡公司乃至天文学家使用该技术从大量的难以处理的数据样本中获得有评判价值的信息。

最常见的数据挖掘的应用之一就是个人信用风险评估。当申请者填写借贷申请表时，他们需要提供他们的社会安全号码、地址和一般的鉴别信息。同时，申请者还被要求提供一些其他方面有关他们自己的信息，例如，有些问题是关于申请者是否是租贷人或房产所有人，申

请人在目前的地址住了多长时间或他为某个雇主工作了多长时间，他的婚姻状况和受教育情况，等等。

因为金融机构有大量的顾客可以用以收集数据，所以他们是应用数据挖掘技术的首选，这就可以分析那些数据，发现申请者的种种个人特征与失败贷款可能的相关性。

不用说，允许计算机评估和检查所有影响该结果的变量使得金融机构能够同时处理成千上万的贷款申请，而只需要用过去人力的一小部分。

数据挖掘技术也广泛应用在零售业来决定如何安排产品的排放位置。例如，一个零售商店要找出一个高尔夫设备的最大销售额方案。通过检查过去四年或更长时间，就会发现买高尔夫设备的顾客们通常需要买双男式鞋。具备了这一信息，零售店可能会决定让高尔夫俱乐部挨着男鞋部以期获得最大销售机会。

条形码技术使得区分连锁商店中购买的每一件商品成为可能。对这些数据的研究，可以得出有关的各种商品购买的关系。例如，如果一个仓储商店要找出销售更多啤酒的方法，就在其连锁商店中找出买啤酒的消费者还买些什么商品，如果发现那些买啤酒的顾客同时常常还买些纸尿片，那么这个仓储商店可能就会通过对纸尿片的优惠来促进啤酒销售。

> **注意**　这个啤酒和纸尿片的例子不是我自己的发明。由于某些偶然的、琐碎的原因，可能因为看起来是滑稽可笑的想象和推想——它恰恰成为数据挖掘文献中讨论相关数据挖掘算法的一个经典例子。

令人感兴趣的是数据挖掘应用于许多领域，如医疗诊断和气象学。那些应用的规则，就像其面对的收益那样面对挑战，这是从令人生畏的一堆数据中发现意义的挑战。简而言之，任何商务的或学术上收集、研究的大量数据都是数据挖掘的候选对象。

1.4　术语定义

数据挖掘常和其他数据存储和数据处理技术联系在一起，如数据仓库、OLTP。这些技术的术语是通用的，有些词可以互换。简明起见，在这里逐一介绍这些技术以及它们的共用的术语，尤其是那些与数据挖掘有关的术语。

- **数据挖掘**（Data Mining）　简要地说，数据挖掘就是通过对大量的历史存储数据的分析和分类，从中得到有意义的模式和关系的过程。
- **数据仓库**（Data Warehousing）　数据仓库就是一个数据存储中心，这些数据是从OLTP数据库的操作数据中提取出来的。与数据仓库不同，OLTP系统是设计用来存储高速处理事务的操作数据的。因为在这些数据库中数据存放的结构对于客户端来说难以理解，所以从中获取信息也比较困难。把这些难以理解的数据转换到一个数据仓库就可以使信息放在一个更容易获取的框架结构中。与OLTP系统比较，数据仓库无须改变已存在的数据就可以接收新数据。因此，该存储结构为容纳海量信息而设计，它们以支持快速检索高效事务处理的结构化方式存储信息。
- **挖掘模型**（Mining Models）　挖掘模型是指由数据挖掘算法编译的数据子集的物理结构，同时还包括对原始数据集的描述。数据挖掘需要一个结构体系，它应包含在基础数据库中呈现的模式。该结构然后成为进行预测的基础，这种预测是在对缺值处进行“填空”的新

数据基础上形成的。通过从原始数据集收集信息，数据挖掘应用软件建立一个数据子集，该子集被编译用于一个数据挖掘算法。按照样本数据，该结果集就可以用来进行预测了。

- **模式**（Patterns） 模式就是指在一个数据库中出现频率足以揭示它们之间有关联的一系列事件。揭示这种关联通常也就是一个归纳推理的过程。例如，你可能发现一系列的数据表明一个顾客买啤酒时，她还要买纸尿片。如果这一事件的发生频率足够高，数据挖掘算法将能够确定它是一个可进行预测的模式，应该存储在一个模型中。通过这一方法，操作者浏览数据挖掘模型时就很清楚地看到那个买啤酒的顾客买纸尿片的可能性也是很大的。
- **实例**（Cases） 用作数据挖掘模型的每一项历史数据都是一个实例。例如，一个数据挖掘模型描述消费者在仓储商店的购买活动，那么，每一次的购买行为都是总结数据挖掘模型的经验的一个独一特定的实例。
- **数据挖掘算法**（Data-Mining algorithms） 一个数据挖掘算法将实例从原始数据转换为数据挖掘模型的数学和统计学的算法。数据挖掘模型最终的形式很大程度上依赖于对数据所应用的数据挖掘算法。你会发现，有很多可以采用的算法，但微软 SQL Server 2000所介绍的数据挖掘服务主要提供的是决策树和聚类分析。

1.5 数据挖掘方法

就像信息系统所涉及的所有原则一样，数据挖掘需要设计出一个计划，并能够按计划将最初的主意变成最终的实现。以下列出了一个数据挖掘计划的组成，如图1-1所示。

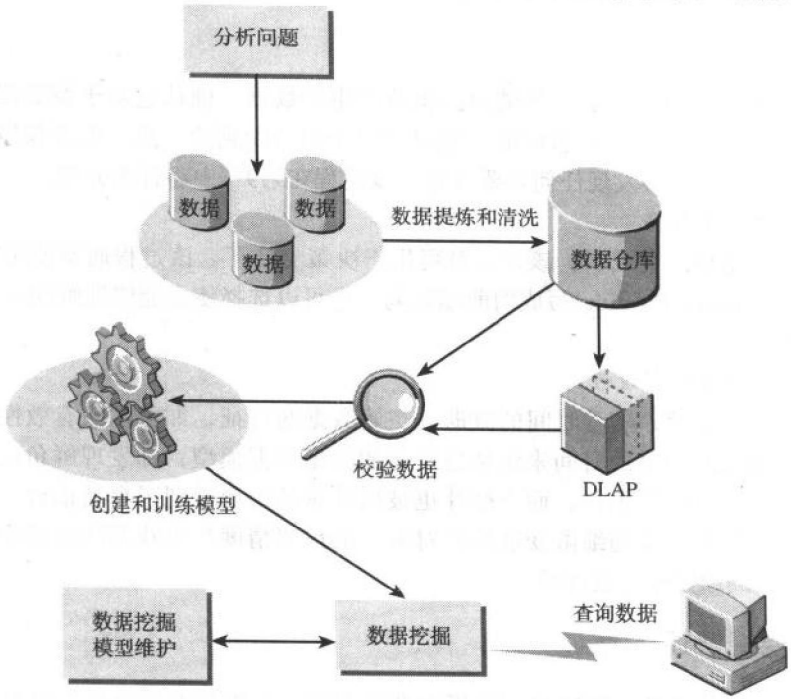

图1-1 数据挖掘方法

- 分析问题。
- 提取和清洗数据。
- 效验数据。
- 创建与调试模型。
- 对数据模型进行数据查询。
- 维护数据挖掘模型的有效性。

1. 分析问题

源数据库必须经过评估以确认其是否符合数据挖掘的标准。数据的质量和充足是决定数据是否合适的首要因素。另外，对数据挖掘的预期结果必须仔细分析以确认已有数据是否确实能够得出这一类别的信息。例如，一个仓储商店的连锁店的数据如果是来自收银机收集的数据，那就无法区分每一个购物车里，消费者买了些什么。一旦决定了预期结果，也就选择了这项工作的最优算法。

2. 提取和清洗数据

数据最初是从自身本源提取的，比如OLTP数据库，文本文件，Access数据库，以及电子表格。提取后的数据放在一个结构上与数据模型兼容的数据仓库中。通常，要用数据转换服务提取数据，以一个统一的格式清洗那些不一致的、不兼容的数据。

3. 校验数据

一旦提取和清理数据后，一个很好的做法是浏览一下你所创建的模型，以确保所有的数据是都已存在并且完整。

4. 创建和调试模型

当算法应用于模型，即产生了一个结构。浏览产生的数据，确认它对于源数据中事实的准确代表性，这是很重要的一点。虽然可能无法对于每个细节做到这一点，但是仅通过查看已生成的模型，就应该能够很容易发现任何重要特性。该过程在第9章中将详细介绍。

5. 查询数据挖掘模型数据

一旦合适的模型创建并生成了，该数据就可用于决策支持了。该过程通常使用VB或ASP通过OLE DB for Data Mining Provider写成的前端查询，也可以选择使用能够理解OLE DB for Data Mining的第三方报表工具。

6. 维护数据挖掘模型的有效性

数据挖掘模型组装好后，随着时间的流逝，初始数据的特征，如粒度或有效性，都可能会发生改变。例如，通过六个月的时间来组装连锁店的数据挖掘模型，却发现鲜鱼已经被从一开始的肉制品柜台换到了海产品柜台，而干酪片也被从乳制品柜台搬到了食品柜台。即使像以六瓶合装的可乐与六个单瓶可乐的细微变动都会对未来的预测精度产生戏剧性的影响，因为它的变化影响了作为基础的原始模型的性质。

1.6　微软数据挖掘概述

数据挖掘是决策支持工具，可以自行分析大型数据库。数据挖掘的设计有它自己的独特特点，能够从事其他数据分析工具无法解决的独特的决策支持问题。人们常常把数据挖掘和其他

工具混淆，如OLAP。这一节将讲述数据挖掘的主要组成和特点。

1.6.1 数据挖掘与OLAP

数据挖掘和OLAP都是微软分析服务的组成部分，都是决策支持工具，但它们是为不同的用处而设计的。OLAP主要是允许客户端设计汇总表用来存储数据，便于数据的修复和导航。许多供应商认为用户通过浏览汇总信息可以发现数据的有关信息，这也就是一种数据挖掘的解决方案，可以从中发现因果关系。但是，大多数情况下，用户仅仅是被带到已经非常了解的数据中而已。对用户来说，得到的结果只是汽车销售的时间、制造、型号等等的一个直观的交互式的展示报告。如果经销商在不同地区卖不同的小车和卡车，把这一信息转换为对商务活动的理解是很简单的。OLAP 可以用来尝试发现新的数据，但因为数据发现的工作实际是由客户端来做的，所以在OLAP的协助下所做的数据发现是比较有局限的，有偶然性，不完全。数据挖掘在客户端是否易于浏览汇总数据这一点上则不太在意，因为它主要是自动地发现可以应用到预测未来结果的新的模式和规则。鉴于这些区别，OLAP被认为是一种高效的存储和修复机制，而数据挖掘是一个知识发掘工具。

1.6.2 数据挖掘模型

源数据需要以优化预测的方式在已建立的变量基础上来构建。正如我前面提到的，该结构是通过一些算法来创建的。当一个算法应用于一个数据结构时，该结构的填充数据从某种意义上来讲反映了存在原始数据集内部的关联和模式，这就意味着基于这些数据的预测是非常容易得出的。微软采用了特殊数据结构来存储数据挖掘模型，要浏览它可以使用微软 Analysis Manager、采用OLE DB for Data Mining Services 的应用程序，或者诸如决策支持对象（DSO）的COM接口。所有这些工具允许创建、复制、改变和删除新的数据模型。通过OLE DB，与各种数据源建立直接的数据连接成为可能，比如微软 SQL Server、文本文件、微软 Access、微软 Excel，甚至是 Oracle 和 DB2。

> **模型（model）与模式（pattern）**
>
> 模型与模式极易混淆。Webster词典认为它们是同义词，但在数据挖掘的术语里，这两个概念是不能互换的。模式是重复出现的数据集，例如111211121112111....，例子中的重复使得我们比较容易也比较准确地预测到下一个将要出现的数字将是2。在数据挖掘的环境中，模型是指一个特定的存储已经过算法处理实例的数据结构，从这个意义上讲模型包含了在原始数据中找到的相同的模式。模型存储了模式，从而使得在上面的例子中，对于我们来说，很容易估计到下一个将要出现的数字。

1.6.3 数据挖掘算法

第6章“微软聚类”和第7章“第三方算法”将详细讨论数据挖掘算法。这些算法中的一些是由数据挖掘服务内部支持的，另外的一些算法由第三方供应商集成到数据挖掘服务中。算法的选择很大程度上还是基于你所要建立的模型，因此，我们将更着重于数据挖掘的过程而不是算法的技术细节。

此为试读，需要完整PDF请访问：www.ertongbook.com