SQL Server 2005 数据挖掘实例分析

SQL Server 2005 Data Mining Analysis by Living Example

王 欣 徐腾飞 唐连章 等编著

中国水利水电出版社
www.waterpub.com.cn

SQL Server 2005 数据挖掘实例分析

王 欣 徐腾飞 唐连章 等编著

中国水利水电出版社

内 容 提 要

数据挖掘的目的在于使用所发现的模式帮助解释当前的行为或预测未来的结果。数据挖掘过程涉及下列7个研究方面：数据仓库及OLAP技术、数据预处理、使用SQL Server Data Mining、关联规则、分类和预测、聚类分析及时序和序列数据的挖掘。

本书对数据挖掘和知识发现的各个方面都进行了必要的解说，侧重于用SSAS进行数据挖掘模型的建立、挖掘结果的分析与检验，以及解释与验证结果。本书对主要的挖掘技术提供了详细的SQL Server 2005数据挖掘的实例，读者通过案例来实验性地建立和检验数据挖掘模型。

本书适合希望学习SQL Server 2005数据挖掘技术的读者，可以作为数据挖掘工程师的参考用书。本书适合作为高校教学数据挖掘的教程，也是公司培训不可多得的参考用书。

图书在版编目（CIP）数据

SQL Server 2005数据挖掘实例分析 / 王欣等编著. —北京：中国水利水电出版社，2008

ISBN 978-7-5084-5346-0

Ⅰ. S… Ⅱ. 王… Ⅲ. 关系数据库－数据库管理系统，SQL Server 2005 Ⅳ. TP311.138

中国版本图书馆CIP数据核字（2008）第032683号

书　　名	SQL Server 2005数据挖掘实例分析
作　　者	王　欣　徐腾飞　唐连章　等编著
出版发行	中国水利水电出版社（北京市三里河路6号　100044） 网址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电话：（010）63202266（总机）、68331835（营销中心）、82562819（万水）
经　　售	全国各地新华书店和相关出版物销售网点
排　　版	北京万水电子信息有限公司
印　　刷	北京蓝空印刷厂
规　　格	787mm×1092mm　16开本　16印张　358千字
版　　次	2008年3月第1版　2008年3月第1次印刷
印　　数	0001—4000册
定　　价	28.00元

前　言

随着计算机技术，特别是数据库技术的快速发展和广泛应用，各行各业积累的数据量越来越大，传统的数据处理方式已经很难充分利用蕴藏在这些数据中的有用知识，于是数据挖掘技术应运而生。

数据挖掘（Data Mining）又称为数据库中的知识发现，可以把数据转化为有用的信息以帮助制定决策，从而在市场竞争中获得优势地位。数据挖掘是一个过程——一个不断把商业经验和知识与数据相结合的过程。数据挖掘的目标是找到能够帮助他们做出对其成功至关重要的决策的信息。例如，他们想知道这样一些情况："现在客户中哪些会对我们的新产品感兴趣?"，"这个贷款申请有合理的信用风险吗？"等等。数据挖掘中应用的方法包括概念描述、分类与预测、关联规则、聚集和神经网络等。

基于数据挖掘技术，微软公司于 2005 年 12 月 2 日发布了新一代企业级应用平台 Microsoft SQL Server 2005、Visual Studio 2005。使用 SQL Server 2005 Analysis Services（SSAS）可以很方便地创建复杂的数据挖掘解决方案。SSAS 工具提供了设计、创建和管理数据挖掘模型的功能，并且使客户端能够访问数据及挖掘数据。

数据挖掘的目的在于使用所发现的模式帮助解释当前的行为或预测未来的结果。数据挖掘过程涉及下列 7 个研究方面：数据仓库及 OLAP 技术、数据预处理、使用 SQL Server Data Mining、关联规则、分类和预测、聚类分析及时序和序列数据的挖掘。

本书对数据挖掘和知识发现的各个方面都进行了必要的解说，侧重于用 SSAS 进行数据挖掘模型的建立、挖掘结果的分析与检验，以及解释与验证结果。为了更好地理解数据挖掘过程，本书对主要的挖掘技术提供了详细的 SQL Server 2005 数据挖掘的实例，读者通过实例来实验性地建立和检验数据挖掘模型。

本书读者

本书阐述了数据挖掘的部分原理以及使用 SSAS 进行数据挖掘的基本方法和各种可视化工具。本书还针对不同的挖掘模型设计了实用的案例，帮助读者深入理解数据挖掘和熟悉 SSAS。对于数据挖掘的用户而言，本书将成为他们的入门工具和实践指南。相信大多数数据库管理人员、IT 专业人员和数据挖掘方面的学生都会从本书中获益。

本书内容

全书分为 9 个章节，细致地讲解了 SQL Server 2005 数据挖掘的原理和实务，帮助读者快速入门学习深奥的数据挖掘知识。本书的内容包括：

第 1 章　数据挖掘基本知识：提供关于数据挖掘的多学科领域的导论，讨论导致需要数据挖掘的数据库技术的发展道路和数据挖掘应用的重要性；考察挖掘的数据类型，包括

关系、事务和数据仓库数据，以及复杂数据类型，如数据流、时间序列、序列、图形、社会网络和多重关系数据、时空数据、多媒体数据、文本数据以及 Web 数据；根据所挖掘的知识类型，对数据挖掘任务进行一般分类。

第 2 章　数据仓库及 OLAP 技术：介绍了数据仓库和联机分析处理的基本概念、系统结构和一般实现，以及数据仓库和数据挖掘的关系；更深入地考察数据仓库和 OLAP 技术，详细地研究数据立方体的计算方法；讨论数据仓库和 OLAP 的进一步探查，如发现驱动的立方体探查，复杂数据挖掘查询的多特征立方体和立方体梯度分析；讨论另一种数据泛化和概念描述方法——面向属性的归纳。

第 3 章　数据预处理：介绍挖掘之前的数据预处理技术，包括描述性数据汇总的各种统计方法，包括数据的中心趋势和散布的度量。加强了数据清理方法的介绍，讨论了数据集成和变换、数据归约的方法，包括动态和静态离散化概念分层的使用；介绍概念分层的自动产生。

第 4 章　使用 SQL Server 2005 进行数据挖掘：通过 Business Intelligence Development Studio 的使用，数据源、数据源视图、数据挖掘对象的管理，数据查看及模型评估，介绍 SSAS 的特性以及设计、创建和管理数据挖掘模型的功能。

第 5 章　关联规则：介绍挖掘关系数据库中的频繁模式、关联和相关性的方法。除介绍诸如购物篮分析等基本概念外，进一步介绍了 Microsoft 关联规则挖掘模型。通过完整的实例阐述了 Microsoft 关联规则的挖掘步骤以及结果分析。

第 6 章　分类和预测：介绍数据分类和预测方法，包括决策树归纳、贝叶斯分类、后向传播的神经网络技术。还介绍了 Microsoft 决策树挖掘模型、Microsoft 贝叶斯挖掘模型、Microsoft 神经网络挖掘模型。通过决策树、神经网络两个实例介绍完整的挖掘方法和结果分析。

第 7 章　聚类分析：主要介绍数据的聚类方法，包括划分方法、层次方法、基于密度的方法等。通过对 Microsoft 聚类挖掘模型及基于该模型的案例的介绍，阐述如何利用 Microsoft 聚类挖掘技术进行挖掘、分析、可靠性检验等。

第 8 章　时序和序列数据的挖掘：主要讨论流数据、时间序列数据和序列数据（包括事务序列和生物学序列）的挖掘，简要介绍了 Microsoft 顺序分析挖掘模型。

第 9 章　数据挖掘的应用和发展趋势：总结本书介绍的概念，并讨论数据挖掘的应用和发展趋势。添加了一些新的数据挖掘材料，涉及生物学和生物医学数据分析、其他科学应用、入侵检测和协同过滤。除了具有挑战性的研究问题之外，还讨论了数据挖掘对社会的影响，如隐私和数据安全问题。

本书特色

- ☑ 简单而详细的例子。本书通过列举不同数据挖掘技术如何建立模型的简单、详细的例子，揭去了数据挖掘的神秘面纱。
- ☑ 章节之间是独立的，读者可以按自己的兴趣选择阅读顺序，实现按需阅读，提高问题的解决能力。

☑ 精选数据挖掘经典分析方向，核心讲解必要的原理，将深奥的数据挖掘原理浅显地讲解出来。

☑ 与实际分析项目结合，全书以作者从事的实际分析项目为蓝本，讲解复杂数据挖掘的具体实践。

致谢

本书由王欣（西南交通大学）、徐腾飞、唐连章编著，姚新军负责前期的策划和后期质量监控。王欣从事数据仓库与数据挖掘领域的教学、研究与应用工作，对 SSAS 有着丰富的实践经验和独特的理解。在本书的编写过程中，参与具体工作的还有万雷、王斌、厉剑梁、殷世钦、江广顺、李强、吴志俊、杜长城、余松、刘羽宇、郭敏、董茜、陈鲲、王晓、陈洪军、余伟炜、王呼佳、许志清、张赛桥、夏惠军。还要感谢中国水利水电出版社计算机编辑室的老师们的辛苦努力，正是因为你们辛苦的付出，才使本书能在第一时间和读者见面。

本书的内容涉及面广，专业性强，虽几经斟酌，多方查找资料，但由于作者水平有限，难免有错误和不当之处，敬请各位读者批评指正。

作　者

2008 年 1 月

目　　录

第 1 章　数据挖掘基本知识

数据挖掘作为一个新兴的多学科交叉应用领域，正在各行各业的决策支持活动中扮演着越来越重要的角色。本章将从数据管理技术演化的角度介绍数据挖掘的由来、作用和意义。同时还将介绍数据挖掘系统的结构、数据挖掘所获得的知识种类，以及数据挖掘系统的分类。最后还简要介绍了当前数据挖掘领域尚存在的一些热点问题。

本章内容包括：

- ☑ 数据挖掘的概念
- ☑ 数据挖掘的存储对象
- ☑ 基本数据挖掘任务
- ☑ 数据挖掘系统的分类
- ☑ 数据挖掘的主要问题

1.1　数据挖掘的概念

数据挖掘，比较公认的定义是 W.J.Frawley、G.PiantetskyShapiro 等人提出来的：数据挖掘就是从大型数据库的数据中提取人们感兴趣的知识。这些知识是隐含的、实现未知的潜在的有用信息，提取的知识表示为概念、规则、规律、模式等形式。

这里把数据挖掘的对象定义为数据库，更广义的说法是：数据挖掘意味着在一些事实或者观察数据的集合中寻找模式的决策支持过程。数据挖掘的对象不仅可以是数据库，也可以是文件系统，或者其他任何组织在一起的数据几何，例如 WWW 信息资源。本书在讨论数据挖掘时采用数据库观点，即着重强调大型数据库（SQL Server 2005）中有效的和可规模化的数据挖掘技术。一个算法是可以规模化的，对于给定的内存和磁盘空间等可利用的系统资源，其运行时间随数据库大小线性增长。通过数据挖掘，可以从数据库提取有趣的知识、规律或高层次信息，并可以从不同角度观察或浏览。发现的知识可以用于决策、过程控制、信息管理、查询处理等。因此，数据挖掘被信息产业界认为是数据库系统重要的前沿之一，是信息产业最有前途的交叉学科。

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。

数据挖掘是信息技术自然进化的结果。如图 1-1 所示，进化过程的见证是数据库工业界开发的以下功能：数据收集和数据库创建、数据管理（包括数据存储和提取，数据库事

务处理）以及数据分析与理解（设计数据仓库和数据挖掘）。例如数据收集和数据库创建机制的早期开发已成为后来数据存储和提取、查询和事务处理有效机制开发的必备基础。随着提供查询和事务处理的大量数据库系统广泛付诸实践，数据分析和理解自然成为下一个目标。

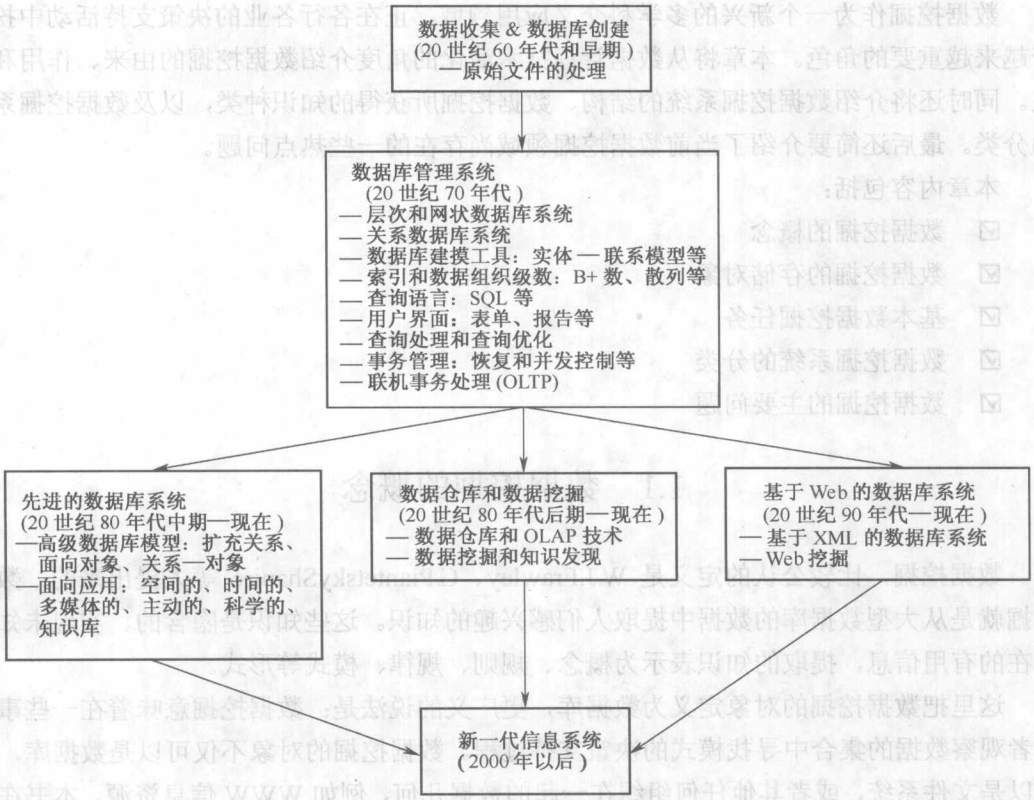

图 1-1 数据库技术的进化

与数据挖掘关系密切的研究领域包括归纳学习、机器学习和统计分析。特别是机器学习被认为和数据挖掘的关系最密切。两者的区别主要在于：数据挖掘的任务是发现可以理解的知识；机器学习关心的是提高系统性能，因此训练神经网络来控制一根倒立行走的木棍是一种机器学习过程，而不是数据挖掘。数据挖掘的对象是大型数据库，一般来说机器学习处理的数据集要小得多，因此效率问题对数据挖掘来说至关重要。

从数据挖掘的定义可以看出，作为一个学术领域，数据挖掘和知识发现（KDD，Knowledge Discovery in Databases）具有很大的重合度，大部分学者认为数据挖掘和知识发现是等价的概念，人工智能领域习惯称为 KDD，数据库领域习惯称为数据挖掘。也有学者把 KDD 看作发现知识的完整过程，而数据挖掘只是这个过程的一个部分。这里，我们倾向于前一种观点，认为数据挖掘从理论上和技术上继承了知识发现领域的成果，同时又有着独特的内涵，数据挖掘更着眼于设计高效的算法以达到从巨量数据中发现知识的目的。如图 1-2 所示，基于这样的观点，典型的数据挖掘系统具有以下主要部分：

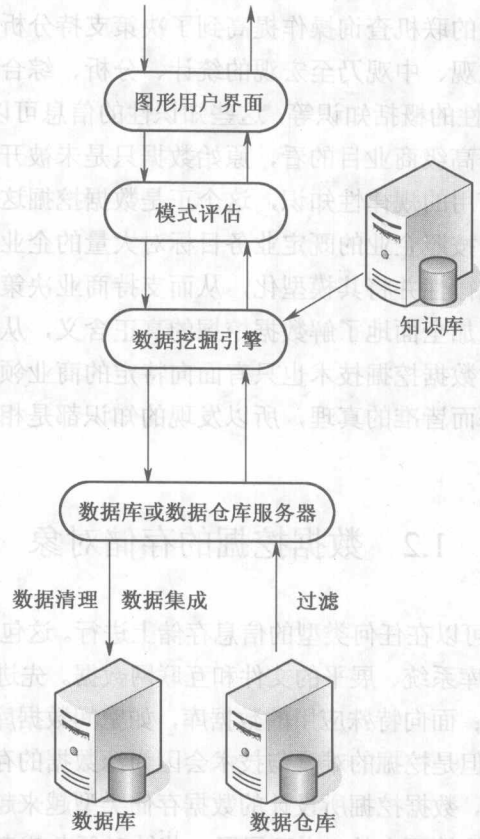

图 1-2　典型的数据挖掘系统结构

- 数据库、数据仓库或其他信息库。这是一个或一组数据库、数据仓库、展开的表或其他类似的数据库。可以在数据上进行数据清理和数据集成。
- 数据库或数据仓库服务器。根据用户的数据挖掘请求，数据库或数据仓库服务器负责提取相关数据。
- 知识库。这是领域知识，用于指导搜索或评估结果模式的兴趣度。这种知识可能包括概念分层，用于将属性或属性值组织成不同的抽象层。用户确信方面的知识也可以包含在内。可以使用这种知识，根据非期望性来评估模式的兴趣度。领域知识的其他例子还有兴趣度限制或阈值和元数据（例如，描述来自多个异种数据源的数据）。
- 数据挖掘引擎。这是数据挖掘系统的基本部分，由一组功能模块组成，用于特征、关联、分类、聚类分析、演变和偏差分析。
- 图形用户界面。该模块在用户和挖掘系统之间通信，允许用户与系统交互，指定数据挖掘查询或任务，提供信息，帮助搜索聚焦，根据数据挖掘的中间结果进行探索式数据挖掘。此外，该部分还允许用户浏览数据库和数据仓库模式或数据结构，评估挖掘的模式，以不同形式使模式可视化。

从商业应用的角度来看，数据挖掘是一种新的商业信息处理技术，数据挖掘技术把人

们对数据的应用从低层次的联机查询操作提高到了决策支持分析预测等更高级的应用上，它通过对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，发现数据间的关联性、未来趋势以及一般性的概括知识等，这些知识性的信息可以用来指导高级商务活动。

从决策分析和预测等高级商业目的看，原始数据只是未被开采的矿山，需要挖掘和提炼才能获得对商业目的有用的规律性知识，这个正是数据挖掘这个名字的由来，所以从商业角度看，数据挖掘就是按照企业的既定业务目标对大量的企业数据进行深层次的分析，以解释隐藏的、未知的规律性并将其模型化，从而支持商业决策活动。从商业应用角度刻画数据挖掘可以使我们更加全面地了解数据挖掘的真正含义，从它的提出之日起就具有很强的商业应用目的，同时数据挖掘技术也只有面向特定的商业领域才有应用价值。数据挖掘并不是要发现放之四海而皆准的真理，所以发现的知识都是相对的，并且对特定的商业行为具有指导意义。

1.2 数据挖掘的存储对象

原则上讲，数据挖掘可以在任何类型的信息存储上进行。这包括关系数据库、数据仓库、事务数据库、先进的数据库系统、展平的文件和互联网数据。先进的数据库系统包括面向对象的对象——关系数据库；面向特殊应用的数据库，如空间数据库、时间序列数据库、文本数据库和多媒体数据库，但是挖掘的难度与技术会因为源数据的存储类型的不同而不同。特别是近些年来的研究表明，数据挖掘所设计的数据存储类型越来越丰富，国内外学者除了一些有通用价值的模型、架构等研究外，也开展了一些针对复杂或者新型数据存储方式下的挖掘技术或算法的研究。下面将针对一些主要的数据存储类型中的数据挖掘问题进行介绍。

1.2.1 关系数据库

数据库系统，也称数据库管理系统（DBMS），由一组内部相关的数据——数据库及一组管理和存储数据的软件程序组成。软件程序涉及如下机制：数据库结构定义，数据存储，并行，共享或分布的数据访问，面对系统瘫痪或未授权的访问，确保数据的一致性和安全性。

关系数据库是由一系列数据表组成的，它本身的发展是相当成熟的，它有成熟的语义模型（如实体－关系模型），有成熟的 DBMS（如 Oracle），有成熟的查询语言（SQL 语言），而且有一批可视化的工具可以使用或借鉴。随着关系型数据库应用的普及和深入，人们在思考更高层次地利用它的问题，即关系型数据库的数据挖掘问题，在一个关系型数据库中，可以根据挖掘目标获得想要的知识类型或模式，例如广义知识、关联知识、类知识、预测型知识和特异性知识等。

关系数据库是表的集合，每个表都赋予一个唯一的名字。每个表包含一组属性（列或字段），并通常存放大量元组（记录或行）。关系中的每个元组代表一个被唯一关键字标识的对象，并被一组属性值描述。语义数据类型，如实体—关系（E-R）数据模型，将数据库作为一组实体和它们之间的联系进行建模，通常称为关系数据库构造 E-R 模型。

例 1.1 教务管理数据库由下列关系表描述：Xjjbxx（学籍基本信息），Dm_kc（课程

代码），Xscjzk（学生成绩状况）和 Xkmd（选课名单），这些表的片段在表 1-4 中给出。

关系表 Xjjbxx 的一组属性如表 1-1 所示，包括学生的唯一标识号（Xh）、姓名（Xm），性别（Xb）、院系代码（Yxdm）、院系名称（Xsh）、专业代码（Zydm）、专业名称（Zyh）、年级（Dqszj）、生源所在地（Syszd）等。

表 1-1　关系表 Xjjbxx

Xid	Xh	Xm	Xb	Yxdm	Xsh	Zydm	Zyh	Dqszj	Syszd
1035	20030293	吴飞	男	0157	管理学院	015789	管理科学与工程	2003	北京
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

关系表 Dm_kc 的一组属性如表 1-2 所示，包括课程代码（Kcdm）、课程名称（Kcmc）、学分（Xf）、开课院系（Kkyx）等。

表 1-2　关系表 Dm_kc

Sid	Kcdm	Kcmc	Xf	Kkyx
89736	01578935	数据库原理	4	信息学院
…	…	…	…	…
…	…	…	…	…

关系表 Xscjzk 的一组属性如表 1-3 所示，包括学号（Xh）、姓名（Xm）、课程代码（Kcdm）、课程名称（Kcmc）、课程类别（Kclb）、成绩（Cj）、学分（Xf）、学年（Xn）、学期（Xq）、开课教师（Kkjs）、教学班号（Jxbh）、分制（Fz）等。

表 1-3　关系表 Xscjzk

cid	Xh	Xm	Kcdm	Kcmc	Kclb	Cj	Xf	Xn	Xq	Kkjs	Jxbh	Fz
87632	2003293	吴飞	01578935	数据库	必	87	4	2006	1	黄海	1	百分
…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…

关系表 Xkmd 的一组属性如表 1-4 所示，包括选课编号（Xkbh）、课程代码（Kcdm）、课程名称（Kcmc）、教学班号（Jxbh）、学号（Xh）、姓名（Xm）、学分（Xf）、开课教师（Kkjs）等。

表 1-4　关系表 Xkmd

Sid	Xkbh	Kcdm	Kcmc	Jxbh	Xh	Xm	Xf	Kkjs
46734455	653921	01578935	数据库原理	1	20030293	吴飞	4	黄海
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

表可用于表示多个关系表之间的联系。对于上述例子，Xkmd（选课名单）表示由学生选课这个事件将多个关系表建立联系。

关系数据可以通过数据库查询访问。数据库查询使用如 SQL 这样的关系查询语言，或借助图形界面书写。对于后者，用户可以使用菜单指定包含在查询中的属性和属性上的限制。一个给定的查询被转换成一系列的关系操作，如联机、选择和投影，并被优化，以便有效地处理。可以下达这样的命令："显示上季度销售商品的列表"。关系查询语言也可以包含聚集函数，如 sum、ave、count、max 和 min，这些函数可以实现"按分店分组显示上个月的总销售额"、"12 月份发生多少笔销售事务？"或"哪位员工是销售冠军？"之类的查询。

当数据挖掘用于关系数据库时，可以进一步搜索趋势或数据模式，例如，数据挖掘系统可以分析顾客数据，根据顾客的收入、年龄和以前的销售信誉等信息预测新顾客的信誉风险。数据挖掘系统也可以检测偏差，例如，与以前的年份相比哪种商品的销售出人预料。这种偏差的原因可以进一步考察（例如，包装是否有变化，价格是否大幅度提高等）。

关系数据库是数据挖掘的最流行的、最丰富的数据源，因此它是进行数据挖掘研究的主要数据形式。关于关系型数据库中的数据挖掘，已经积累了很多方法和成果，目前的研究更倾向于针对关系型数据库的特点集成多种技术来解决实际应用问题。这些问题如下：

- 多维知识挖掘问题。传统的事务数据库所研究的知识一般是单维的，例如"购买计算机的人也购买打印机"这样的知识，它刻画了以"购买"行为作为聚焦点（维）的商品间的关联。但是，在关系型数据库中，仅有这样的知识可能还不够，例如人们可能还想进一步知道"什么样购买计算机的人也购买打印机的可能性更大"。因此，类似于"收入高的人在购买计算机时也购买打印机"这样的知识更加有价值。由于关系型数据库可以存储包含输入情况等的客户基本资料以及客户购买记录，所以这样的知识是有可能获得的。这样的知识是多维的，因为它有两个聚焦点：购买和收入。
- 多表挖掘和数量数据挖掘问题。多表挖掘和数量数据挖掘问题是关系数据库有别于传统的事务数据库挖掘中的两个重要问题。从逻辑上说，关系型数据库是一系列表的集合。因此，在关系型数据库的挖掘中，除了要考虑表内属性的关联外，也必须考虑表间属性的关联。传统的事务数据库挖掘所研究的技术和算法一般是基于单表的。因此，在关系型数据库挖掘中必须考虑多表的挖掘技术。另外，在关系型数据库中，可能具有数量属性（如工资）。这些数量属性给传统的数据挖掘方法，如 Agrawal 的规则发现架构，提出了新的挑战。
- 多层知识挖掘问题。数据及其关联总是可以在多个不同的概念层上来理解它。在一定的背景知识下，一个关系型数据库可以在多个概念层次上来挖掘相关的知识。1995 年，Srikant 和 Agrawal 建立了以广义知识挖掘框架来研究多层知识挖掘的思想，并提出了 R-兴趣度等概念。另一个比较有代表性的工作是 Han 等对大型数据库的多层知识挖掘问题的研究。
- 知识评价问题。1996 年，Chen 和 Han 发现按着 Agrawal 的规则发现理论进行强