第3章

数据挖掘概述

随着计算机软硬件技术的发展，尤其是计算机网络的发展与普及，计算机处理和存储的数据，正在以难以预计的速度增长；另外，随着社会经济的不断发展，商业竞争日趋白热化，人们迫切需要从数据中获得有用的知识来帮助进行科学决策。针对“数据丰富而知识贫乏”这一窘境，数据挖掘应运而生。

数据挖掘使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，并且能够找出与过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。通过数据挖掘，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，从而使大型数据库作为一个丰富、可靠的资源为知识的提取服务。

3.1 数据挖掘的起源与发展

3.1.1 数据挖掘的起源

为解决上述问题，来自不同学科的研究者汇集到一起，开始着手开发能够处理不同数据类型的更有效的、可伸缩的工具。这些工作都是建立在研究者先前使用的方法学和算法之上，并在数据挖掘领域达到高潮。特别地，数据挖掘利用了来自如下一些领域的思想：①统计学的抽样、估计和假设检验；②人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。

一些其他领域也起到重要的支撑作用。特别地，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能(并行)计算的技术在处理海量数据集方面常常是重要的。分

布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

图 3-1 展示了数据挖掘与其他领域之间的联系。

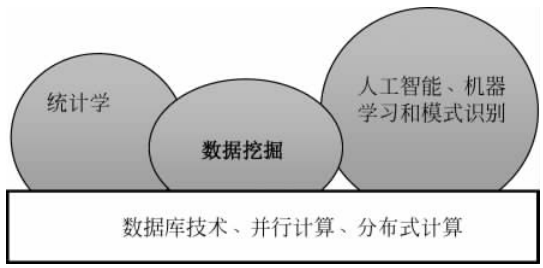

图 3-1 数据挖掘汇集了许多学科的知识

3.1.2 数据挖掘的发展

经过十几年的研究和实践，数据挖掘技术已经吸收了许多学科的最新研究成果，从而形成了独具特色的研究分支。毋庸置疑，数据挖掘研究和应用具有很大的挑战性。像其他新技术的发展历程一样，数据挖掘也必须经过概念的提出、概念的接受、广泛研究和探索、逐步应用和大量应用等阶段。从现状看，大部分学者认为数据挖掘的研究仍然处于广泛研究和探索阶段。一方面，数据挖掘的概念已经被广泛接受。在理论上，一批具有挑战性和前瞻性的问题被提出，吸引越来越多的研究者；另一方面，数据挖掘的广泛应用还有待时日，需要深入的研究积累和丰富的工程实践。

随着数据挖掘概念在学术界和工业界的影响越来越大，数据挖掘的研究向着更深入和更实用的技术方向发展。从事数据挖掘研究的人员主要在大学、研究机构，也有部分在企业或公司。所涉及的研究领域很多，研究集中在学习算法的研究、数据挖掘的实际应用以及有关数据挖掘理论等方面。进行的大多数基础研究项目是由政府资助进行的，而公司的研究更注重和实际商业问题相结合。

数据挖掘的概念从 20 世纪 80 年代被提出后，其经济价值就已经显现出来，而且被众多商业厂家所推崇，形成初步的市场。一份最近的 Gartner 报告中列举了在今后 3～5 年内对工业将产生重要影响的 5 项关键技术，其中数据挖掘和人工智能排名第一。同时，这份报告将并行计算机体系结构研究和数据挖掘列入今后 5 年内公司应该投资的 10 个新技术领域。另外，目前的数据挖掘系统也绝不是像一些商家为了宣传自己的商品所说的那样神奇，仍有许多问题需要研究和探索。把目前数据挖掘的研究现状描述为鸿沟(Chasm)阶段是比较准确的。所谓 Chasm 阶段是说数据挖掘技术在广泛被应用之前仍有许多“鸿沟”需要跨越。例如，就目前商家推出的数据挖掘系统而言，它们都是一些通用的辅助开发工具。这些工具只能给那些熟悉数据挖掘技术的专家或高级技术人员使用，仅对专业人员开发对应的应用起到加速或横向解决方案(Horizontal Solution)的作用。但是，数据挖掘来自于商业应用，而商业应用又会由于应用的领域不同而存在很大差异。大多数学者赞成这样的观点：数据挖掘在商业上的成功不能期望通过通用的辅助开发工具，而应该是数据挖掘概念与特定领域商业逻辑相结合的纵向解决方案(Vertical Solution)。

分析目前的研究和应用现状，数据挖掘在如下几个方面需要重点开展工作。

1. 数据挖掘技术与特定商业逻辑的平滑集成问题

谈到数据挖掘和知识发现技术，人们大多引用“啤酒与尿布”的例子。事实上，目前关于数据挖掘的确很难找到这样的其他经典例子。数据挖掘和知识发现技术的广阔应用前景，需要有效和显著的应用实例来证明。因此包括领域知识对行业或企业知识挖掘的约束与指导、商业逻辑有机地嵌入数据挖掘过程等关键课题，将是数据挖掘与知识发现技术研究和应用的重要方向。

2. 数据挖掘技术与特定数据存储类型的适应问题

不同的数据存储方式会影响数据挖掘的具体实现机制、目标定位、技术有效性等。指望一种通用的应用模式结合所有的数据存储方式发现有效知识是不现实的。因此，针对不同数据存储类型的特点，进行针对性研究是目前流行而且也是将来一段时间所必须面对的问题。

3. 大型数据的选择与规格化问题

数据挖掘技术是面向大型数据集的，而且源数据库中的数据是动态变化的，数据存在噪声、不确定性、信息丢失、信息冗余、数据分布稀疏等问题，因此挖掘前的预处理工作是必需的。数据挖掘技术又是面向特定商业目标的，大量的数据需要选择性地利用，因此针对特定数据挖掘问题进行数据选择、针对特定挖掘方法进行数据规格化是无法回避的问题。

4. 数据挖掘系统的构架与交互式挖掘技术

虽然经过多年的探索，数据挖掘系统的基本架构和过程已经趋于明朗化，但是受应用领域、挖掘数据类型以及知识表达模式等的影响，在具体的实现机制、技术路线以及各阶段或部件（如数据清洗、知识形成、模式评估等）的功能定位等方面仍需细化和深入研究。由于数据挖掘是在大量的元数据集中发现潜在的、事先并不知道的知识，因此和用户进行交互式探索性挖掘是必然的。这种交互可能发生在数据挖掘的各个不同阶段，从不同角度或不同程度进行交互。所以良好的交互式挖掘（Interaction Mining）也是数据挖掘系统成功的前提。

5. 数据挖掘语言与系统的可视化问题

对 OLTP 应用来说，结构化查询语言 SQL 已经得到充分的发展，并成为支持数据库应用的重要基石。但是，对于数据挖掘技术而言，由于诞生的时间较晚，加之它相比 OLTP 应用的复杂性，开发相应的数据挖掘操作语言仍然是一件极富挑战性的工作。可视化要求已经成为目前信息处理系统的必不可少的技术，对于一个数据挖掘系统来说，它更是重要的。可视化挖掘除了要和良好的交互式技术结合外，还必须在挖掘结果或知识模式的可视化、挖掘过程的可视化以及可视化指导用户挖掘等方面进行探索和实践。数据的可视化从某种程度来说起到了推动人们主动进行知识发现的作用，因为它可以使人们从对数据挖掘的神秘感变成可以直观理解的知识和形象的过程。

6. 数据挖掘理论与算法研究

经过十几年的研究，数据挖掘已经在继承和发展相关基础学科（如机器学习、统计学等）已有成果方面取得了可喜的进步，探索出了许多独具特色的理论体系。但是，这绝不意味着挖掘理论的探索已经结束，恰恰相反，它留给了研究者丰富的理论课题。一方面，在这些大的理论框架下有许多面向实际应用目标的挖掘理论等待探索和创新；另一方面，随着数据

挖掘技术本身和相关技术的发展，新的挖掘理论的诞生是必然的，而且可能对特定的应用产生推动作用。新理论的发展必然促进新的挖掘算法的产生，这些算法可能扩展挖掘的有效性。如针对数据挖掘的某些阶段、某些数据类型、大容量元数据集等更有效；可能提高挖掘的精度或效率；可能融合特定的应用目标，如 CRM、电子商务等。因此，对数据挖掘理论和算法的探讨将是长期而艰巨的任务。特别是，像定性定量转换、不确定性推理等一些根本性的问题还没有得到很好的解决，同时需要针对大容量数据的有效和高效算法。从上面的叙述可以看出，数据挖掘研究和探索的内容是极其丰富和具有挑战性的。

3.2 数据挖掘所要解决的问题

前面提到，面临新的数据集带来的问题时，传统的数据分析技术常常遇到实际困难。下面是一些具体的问题，它引发了人们对数据挖掘开展研究。

(1) 可伸缩。由于数据产生和收集技术的进步，数吉字节、数太字节甚至数拍字节[①]的数据集越来越普遍。如果数据挖掘算法要处理这些海量数据集，则算法必须是可伸缩的。许多数据挖掘算法使用特殊的搜索策略处理指数级搜索问题。为实现可伸缩可能还需要实现新的数据结构，才能以有效的方式访问每个记录。例如，当要处理的数据不能放进内存时，可能需要非内存算法。使用抽样技术或开发并行和分布算法也可以提高可伸缩程度。

(2) 高维性。目前，经常会遇到具有成百上千属性的数据集，而不是几十年前常见的只具有少量属性的数据集。在生物信息领域，微阵列技术的进步已经产生了涉及数千特征的基因表达数据。具有时间或空间分量的数据集也经常具有很高的维度。例如，考虑包含不同地区的温度测量结果的数据集，如果在一个相当长的时间周期内反复地测量，则维度(特征数)的增长正比于测量的次数。为低维数据开发的传统的数据分析技术通常不能很好地处理这样的高维数据。此外，对于某些数据分析算法，随着维度(特征数)的增加，计算复杂性迅速增加。

(3) 异种数据和复杂数据。通常，传统的数据分析方法只处理包含相同类型属性的数据集，或者是连续的，或者是分类的。随着数据挖掘在商务、科学、医学和其他领域的作用越来越大，越来越需要能够处理异种属性的技术。近年来，已经出现了更复杂的数据对象。这些非传统的数据类型的例子包括含有半结构化文本和超链接的 Web 页面集、具有序列和三维结构的 DNA 数据、包含地球表面不同位置上的时间序列测量值(温度、气压等)的气象数据。为挖掘这种复杂对象而开发的技术应当考虑数据中的联系，如时间和空间的自相关性、图的连通性、半结构化文本和 XML 文档中元素之间的父子联系。

(4) 数据的所有权与分布。有些时候，需要分析的数据并非存放在一个站点，或归属一个机构，而是地理上分布在属于多个机构的资源中。这就需要开发分布式数据挖掘技术。分布式数据挖掘算法面临的主要挑战包括：①如何降低执行分布式计算所需的通信量；②如何有效地统一从多个资源得到数据挖掘结果；③如何处理数据安全性问题。

(5) 非传统的分析。传统的统计方法基于一种假设-检验模式，即提出一种假设，设计

① Gigabytes、Terabytes、Petabytes 分别是 10^9B，10^{12}B，10^{15}B。

实验来收集数据，然后针对假设分析数据。但是，这一过程费时费力。当前的数据分析任务常常需要产生和评估数千种假设，因此需要自动地产生和评估假设，这促使人们开发了一些数据挖掘技术。此外，数据挖掘所分析的数据集通常不是精心设计的实验的结果，并且它们通常代表数据的时机性样本，而不是随机样本。而且，这些数据集通常涉及非传统的数据类型和数据分布。

3.3 数据挖掘的定义

数据挖掘是一门涉及面很广的交叉学科，融合了模式识别、数据库、统计学、机器学习、粗糙集、模糊数学和神经网络等多个领域的理论，因此可从多个视角来看待它。

从技术角度来看，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们不知道的但又是潜在有用的信息和知识的过程。这个定义有如下含义：数据源是真实的、大量的，并且可能是有噪声的；所发现的信息是用户感兴趣的知识；发现的知识是用户能够理解并使用的。在数据挖掘中，原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至可以是分布在网络上的异构数据。挖掘出来的知识可用于查询优化、信息管理、决策支持和过程控制等，还可用于数据自身的维护。数据挖掘把人们对数据的应用从低层次的简单查询，提升到从数据库中挖掘知识，从而提供决策支持。

从商业角度来看，数据挖掘就是按企业的既定业务目标，对大量的企业数据进行探索和分析，以揭示隐藏的、未知的规律性并将其模式化，从而支持商业决策活动。数据挖掘技术只有面向特定的商业领域才有应用价值，是一种新的商业信息处理模式，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和处理，从中提取出辅助商业决策的关键信息和知识。

从以上定义，可以得到数据挖掘具有以下特点。

（1）数据量巨大。如何高效地存取大量数据，如何在特定应用领域中找出特定的高效率算法，以及如何选取数据子集，都成为数据挖掘工作者要重点考虑的问题。

（2）动态性。许多领域的行业数据所包含的规律时效性很强，随着时间和环境的变化规律也在改变。这种数据和知识的迅速变化，就要求数据挖掘能快速做出相应的反应以及时提供决策支持。

（3）适用性。数据挖掘的规律适用于一部分数据，但不可能适用于全部数据，这是因为外部的环境不可能完全相同。

（4）系统性。数据挖掘不是一个简单算法，而是一个较为复杂的系统，它需要业务理解、数据理解、数据准备、建模、评估等一系列步骤，是一个不断循环和不断完善的系统工程。

3.4 数据挖掘的过程

在数据挖掘中，被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图 3-2 中各步骤是按一定顺

序完成的，当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成。在整个数据挖掘过程中，60％的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后续挖掘工作仅占总工作量的 10％。

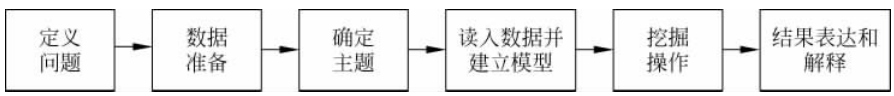

图 3-2　数据挖掘的一般流程

从大量的、不完全的、有噪声的、模糊的甚至随机的实际应用数据中提取出隐含在其中的非常有用的信息、模式（规则）和趋势的数据挖掘过程主要包括 6 个步骤，各步骤的大体内容如下。

（1）定义问题。首先明确定义将要解决的问题。数据挖掘者要熟悉所研究行业的数据和业务问题，缺乏这些，就不能够充分发挥数据挖掘的价值，很难得到正确的结果。模型的建立取决于问题的定义，有时相似的问题，所要求的模型几乎完全不同。

清晰地定义出业务问题，认清数据挖掘的目的，是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

（2）数据准备。有些人将数据挖掘看作是一个不可思议的过程，认为它吞进的是原始数据，吐出来的是“钻石”。数据准备正是这个过程的核心。这一阶段又可分为三个子步骤：数据集成，数据选择，数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理，解决语义模糊性，处理数据中的遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合，缩小处理范围，提高数据挖掘的质量，因此需要搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。而数据预处理则是为了克服目前数据挖掘工具的局限性，提高数据质量，同时将数据转换成一个适用于特定挖掘算法的分析模型。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

（3）确定主题。数据挖掘是一个经常需要回溯的过程，因此没有必要在数据完全准备好之后才开始进行数据挖掘。随着时间的推移，你所使用的数据、你对它们分组的方式以及数据清洗的效果等都将改变，并有可能改进整个模型。这一步会涉及了解研究主题的局限性，选择待完成的良好研究主题，确定待研究的合适的数据元素，以及决定如何进行数据操作等。

（4）读入数据并建立模型。一旦确定要输入的数据之后，接着就是要用数据挖掘工具读入数据并从中构造出一个模型。根据所选用的数据挖掘工具的不同，所构造出的数据模型也会有很大的差别。

（5）挖掘操作。依照上述准备工作，利用选好的数据挖掘工具在数据中查找，这个搜索过程可以由系统自动执行，自底向上搜索原始事实以发现它们之间的某种联系，也可以加入用户交互过程，由分析人员主动发问，从上到下地找寻以验证假设的正确性。数据挖掘的搜索过程需要反复多次，通过评价数据挖掘结果不断调整数据挖掘的精度，以达到发现知识的目的。

（6）结果表达和解释。根据最终用户的决策目标对提取出的信息进行分析，把最有价值的信息区分出来，并通过决策支持工具提交给决策者。

数据挖掘过程的分步实现，不同的阶段会需要有不同专长的人员，他们大体可以分为以下三类。

（1）业务分析人员：要求精通业务，能够解释业务对象，并能根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

（2）数据分析人员：要求精通数据分析技术，对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

（3）数据管理人员：要求精通数据管理技术，并能从数据库或数据仓库中搜集数据。

从上可见，数据挖掘是一个多种专业人员相互配合的工作过程，也是一个在资金上和技术上高投入的过程。这一过程要反复进行，在反复的过程中，不断地趋近事物的本质，不断地优选问题的解决方案。

20 世纪 90 年代后期，当时的数据挖掘市场是年轻而不成熟的，但是这个市场显示出了爆炸式的增长。三个在这方面经验丰富的公司 Daimler Chrysler、SPSS、NCR 发起并建立了一个社团，目的是建立数据挖掘方法和过程的标准。在获得了 EC（European Commission）的资助后，他们开始实现他们的目标。为了征集业界广泛的意见，共享知识，他们创建了 Special Interest Group（SIG）。SIG 组织开发并提炼出 CRISP-DM（Cross-Industry Standard Process for Data Mining），如图 3-3 所示，同时在 Mercedes-Benz 和 OHRA（保险领域企业）中进行了大规模数据挖掘项目的实际试用。SIG 还将 CRISP-DM 和商业数据挖掘工具集成起来。SIG 组织目前在伦敦、纽约、布鲁塞尔已经发展到二百多个成员。

当前 CRISP-DM 提供了一个数据挖掘生命周期的全面评述，包括项目的相应周期、它们各自的任务和任务之间的关系。在这个描述层中，识别出所有关系是不可能的。所有数据挖掘任务之间关系的存在依赖于用户的目的、背景和兴趣，最重要的还有数据。SIG 组织已经发布了 CRISP-DM Version 1.0 Process Guide and User Manual 的电子版，可以免费使用。

一个数据挖掘项目的生命周期包含 6 个阶段。这 6 个阶段的顺序是不固定的。我们经常需要前后调整这些阶段。这依赖于每个阶段中特定任务的产出物是否是下一个阶段必需的输入。图 3-3 中的箭头指出了最重要的和依赖度高的阶段关系。

图 3-3 中的外圈象征数据挖掘自身的循环本质——在一个解决方案发布之后一个数据挖掘的过程才可以继续。在这个过程中得到的知识可以触发新的、经常是更聚焦的商业问题。后续的过程可以从前一个过程中得到益处。

（1）业务理解。最初的阶段集中在理解项目目标和从业务的角度理解需求，同时将这个知识转化为数据挖掘问题的定义和完成目标的初步计划。将知识转化为定义和计划。

（2）数据理解。数据理解阶段从初始的数据收集开始，通过一些活动的处理，以熟悉数据，识别数据的质量问题，首次发现数据的内部属性，或是探究引起兴趣的子集以形成隐含信息的假设。

（3）数据准备。数据准备阶段包括从未处理数据中构造最终数据集的所有活动。这些数据将是模型工具的输入值。这个阶段的任务有可能执行多次，没有任何规定的顺序。任务包括表、记录和属性的选择，模型工具的转换和数据的清洗。

（4）建立模型。在这个阶段，可以选择和应用不同的模型技术，模型参数被调整到最佳的数值。有些技术可以解决一类相同的数据挖掘问题。有些技术在数据形成上有特殊要

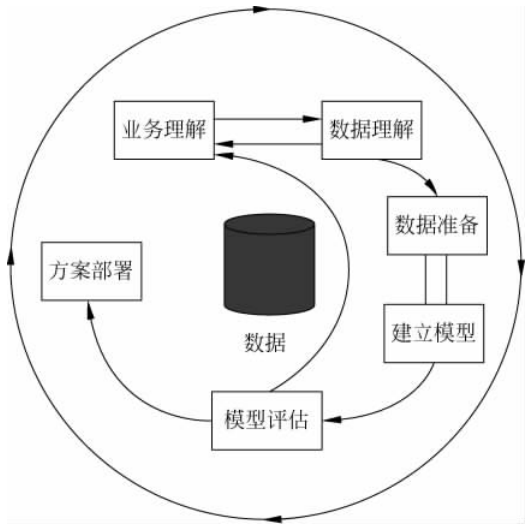

图 3-3　CRISP-DM 的组成架构

求，因此需要经常跳回到数据准备阶段。

（5）模型评估。到项目的这个阶段，就已经从数据分析的角度建立了一个高质量显示的模型。在开始最后部署模型之前，重要的事情是彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键任务是确定是否有重要业务问题没有被充分地考虑。在这个阶段结束后，必须达成一个数据挖掘结果使用的决定。

（6）方案部署。通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得知识，并以便于用户使用的方式重新组织和展现。根据需求，这个阶段可以产生简单的报告，或实现一个比较复杂的、可重复的数据挖掘过程。在很多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作。

3.5　数据挖掘系统

3.5.1　数据挖掘系统的分类

数据挖掘源于多个学科，因此数据挖掘研究产生了大量的、各种不同类型的数据挖掘系统。这样，就需要对数据挖掘系统进行分类。这种分类可以帮助用户区分数据挖掘系统，确定最适合其需求的数据挖掘系统。根据不同的标准，数据挖掘系统可以进行以下分类。

1. 根据数据挖掘的数据库类型分类

由于数据库系统本身可以根据不同的标准分类，因此，数据挖掘系统可以进行相应的分类。根据数据模型分类，可以分为关系的、事务的、面向对象的、数据仓库的数据挖掘系统；根据所处理数据的特定类型分类，可以分为空间的、时间序列的、文本的、多媒体的或 Web 的数据挖掘系统。

2. 根据挖掘的知识类型分类

该类数据挖掘系统依据所挖掘出的规则而分类，这些规则有分类规则、特征规则、聚类分析、关联规则、孤立点分析、时间序列模式分析等。

3. 根据挖掘方法分类

根据所采用的挖掘方法的不同，分为面向数据库的方法、机器学习方法、统计学方法、模式识别方法、可视化方法等。具体地，可以分为模糊集方法、神经网络方法、统计方法、粗糙集方法、决策树、生物智能方法等。

4. 根据数据挖掘应用分类

不同的应用需要有针对该应用的特别有效的方法，因此数据挖掘系统还可以根据其应用领域来分类，从而出现了诸如股票市场数据挖掘系统、DNA 序列数据挖掘系统、电信行业数据挖掘系统、旅游数据挖掘系统、医药销售数据挖掘系统、保险行业数据挖掘系统等。

3.5.2 数据挖掘系统的发展

一般来讲，开发数据挖掘系统是一个由多功能部件组成的、多种类技术相互合作的系统性研发过程。粗略地说，数据挖掘系统的发展可分为三个阶段：独立数据挖掘软件（1995 年之前），横向数据挖掘工具集（1995 年起），纵向数据挖掘解决方案（1999 年起）。

（1）独立数据挖掘软件。独立的数据挖掘软件出现于数据挖掘技术发展的早期，研究人员每开发出一种新型的数据挖掘算法，就会形成一个相应的软件原型，这些原型系统会不断尝试和不断完善。这类软件要求用户对具体的算法和数据挖掘技术有相当的了解，还需要负责大量的数据预处理工作。

（2）横向数据挖掘工具集。随着数据挖掘和知识发现技术的不断发展和研究的不断深入，人们逐渐认识到随着数据量的增加和应用领域的拓宽而涌现的一些新问题，诸如：现实领域中的问题多种多样，单靠少数几个数据挖掘算法难以解决；有待挖掘的数据通常不符合算法要求，需要有数据清洗、转换等数据预处理操作配合，才能得出有价值的模型。因此需要大量多领域、多方法、多技术的结合，由此积累了许多数据挖掘模型和算法，从而出现了一批集成化的数据挖掘工具集。从 1995 年开始，软件开发商提供了“工具集”的数据挖掘软件。由于这类工具并非面向特定的应用，而是通用的算法集合，所以称之为横向数据挖掘工具。典型的数据挖掘工具有 SPSS Clementine、IBM Intelligent Miner、SAS Enterprise Miner、Oracle Darwin、SGI MineSet 等。

（3）纵向数据挖掘解决方案。随着横向数据挖掘工具的使用日益广泛，人们发现只有熟悉数据挖掘算法的专家才能使用这类工具。如果对数据挖掘技术及算法不了解，就难以开发出好的应用系统。从 1999 年开始，大量的数据挖掘工具研制者开始提供纵向的数据挖掘解决方案。这种方案的核心是针对特定的应用提供完整的数据挖掘解决方案，优点是挖掘目标明确、针对性强、挖掘模型选择方便、系统研制快捷。由于和特定的商业领域相联系，因此数据挖掘技术的应用成为企业信息系统的一部分。

根据以上所述，按照数据挖掘系统的特征和发展趋势，可将数据挖掘系统归纳为 4 代。4 代数据挖掘系统的特征、所采用的数据挖掘算法数量、集成的功能、分布计算模型的方式和数据挖掘模型等方面如下叙述。

(1) 第一代数据挖掘系统。在第一代数据挖掘系统中,数据挖掘通常作为一个独立的应用,系统仅支持一个或少数几个数据挖掘算法,这些算法被用来挖掘向量数据,这些数据模型在挖掘时一次性地调入内存进行处理,通常在单台机器上运行。

(2) 第二代数据挖掘系统。第二代数据挖掘系统支持数据库和数据仓库的集成,同时它们具有高性能的接口,具有很好的可扩展性。第二代数据挖掘系统通过支持数据挖掘模式和数据挖掘查询语言来增加系统的灵活性,能够挖掘大数据集、更复杂的数据集以及高维数据。

(3) 第三代数据挖掘系统。第三代数据挖掘系统能够挖掘 Internet/Extranet 的分布式和高度异质的数据,并且能够有效地将其同操作系统集成。这一代数据挖掘系统的关键技术之一是对建立在异质系统上的多个预言模型以及管理这些预言模型的元数据提供支持。

(4) 第四代数据挖掘系统。第四代数据挖掘系统能够采用多个算法挖掘嵌入式系统、移动系统和普遍存在的计算设备所产生的各种类型的数据,使系统的集成度更高,计算方式和数据模型更加复杂。

3.6 数据挖掘的功能和方法

3.6.1 数据挖掘的功能

数据挖掘是一门交叉学科,融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。数据挖掘的主要功能有以下几点。

1. 关联分析

关联分析的目的是找出数据集中属性值之间的联系,形成关联规则。为了发现有意义的关联规则,需要给定两个阈值:最小支持度和最小可信度。在这个意义上,挖掘出的关联规则就必须满足最小支持度和最小可信度。关联规则是在 1993 年由 R. Agrawal 等人提出的,然后扩展到从关系数据库、空间数据库和多媒体数据库中挖掘关联关系,并且要求挖掘出通用的、多层次的、用户感兴趣的关联规则。随着应用和技术的发展,几年来对挖掘关联规则的技术提出了更新的要求,如在线挖掘、提高挖掘大型数据库的计算效率、减小 I/O 开销、挖掘定量型关联规则等。

2. 概念描述

一个概念通常是对一个包含大量数据的数据集总体情况的描述。概念描述就是通过对与某类对象关联数据的汇总、分析和比较,对此类对象的内涵进行描述,并概括这类对象的有关特征。这种描述是汇总的、简洁的和精确的,当然也是非常有用的。概念描述分为特征性描述和区别性描述。前者描述某类对象的共同特征,后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性;生成区别性描述则涉及目标类和对比类中对象的共性。

3. 数据总结

数据总结的目的是对数据进行浓缩,给出数据集的紧凑描述。数据挖掘是从数据泛化

的角度来研究数据总结的。数据泛化是一种把数据库中的相关数据从低层次抽象到高层次的过程。用户有时希望可以从高层次的视图上浏览数据，因而需要对数据进行不同层次上的泛化以适应各种查询及处理需求。目前，数据泛化的主要技术有面向属性的归纳技术和多维数据分析方法。

4. 分类分析

类刻画了一类事物，这类事物具有某种意义上的共同特征，并明显与不同类事物相区别。分类分析就是通过分析示例数据库中的数据，为每个类别做出准确的描述或建立分析模型或挖掘出分类规则，然后用这个分类规则对其他数据库中的记录进行分类。从机器学习的观点来看，分类技术是一种有指导的学习，即每个训练样本的数据对象已经有类标识，通过学习可以形成与表达数据对象与类标识间对应的知识。目前已有多种分类分析模型得到应用，主要有神经网络方法、Bayesian 分类、决策树、统计分类方法、粗糙集分类、SVM 方法、覆盖算法等。在数据挖掘中这些方法均遇到数据规模的问题，即大多数方法能有效解决小规模数据库的数据挖掘问题，但当应用于大数据量的数据库时，会出现性能恶化、精度下降的问题。

5. 聚类分析

聚类是把一组个体按照相似性归成若干类别，它的目的是使得属于同一类别的个体之间的差别尽可能小，而不同类别上的个体间的差别尽可能大。聚类结束后，每类中的数据由唯一的标志进行标识，各类数据的共同特征也被提取出来，用于对该特征进行描述。提高聚类效率、减少时间和空间开销，以及如何在高维空间进行有效数据聚类是聚类研究中的主要问题。聚类分析的方法很多，如 k-平均算法、k-中心点算法、基于凝聚的层次聚类和基于分裂的层次聚类等。采用不同的聚类方法，对于相同的记录集合可能有不同的划分结果。

分类和聚类技术不同，前者总是在特定的类标识下寻求新元素属于哪个类，而后者则是通过对数据的分析比较生成新的类标识。

6. 时间序列分析

时间序列分析中的相似模式发现分为相似模式聚类和相似模式搜索两种。相似模式聚类是将时间序列数据分隔成等长或不等长的子序列，然后用模式匹配的方法进行聚类，找出序列中所有相似的模式。相似模式搜索是指给定一个陌生子序列，在时间序列中搜索所有与给定子序列模式最接近的数据子序列。时间序列分析主要应用于天气数据预报、金融市场数据分析、医疗诊断分析、科学工程数据以及通信信号、雷达信号数据处理等方面。

7. 偏差分析

偏差分析包括分类中的反常实例、例外模式、观测结果对期望值的偏离以及量值随时间的变化等，基本思想就是对数据库中的偏差数据进行检测和分析，检测出数据库中的一些异常记录，它们在某些特征上与数据库中的大部分数据有着显著不同。通过发现异常，可以引起人们对特殊情况的格外关注。异常模式包含：出现在其他模式边缘的奇异点；不满足常规类的异常实例；与父类或兄弟类不同的类；观察值与模型推测出的期望值有明显差异的例子等。偏差分析方法主要有基于统计的方法、基于距离的方法和基于偏移的方法。孤点数据的发现可以应用在信用卡使用、金融欺诈防范、医学数据分析等领域中。

8. 建模

通过数据挖掘，建造出描述一种状态或活动的数学或物理模型。机器学习中的数据挖掘就是对一些自然现象进行建模，重新发现科学定律，如 BACON 系统。基本的思路是：采用数据驱动，通过启发式约束搜索，依赖于理论数据项，应用一些通用的发现方法，找出概念之间的内在联系并表示出来，从而探索出理论模型。

3.6.2 数据挖掘的方法

由于数据挖掘应用领域十分广泛，因此产生了多种数据挖掘的算法和方法，如决策树方法、模糊集方法、神经网络方法、粗糙集方法、统计分析方法、可视化方法等。有时对于某一数据库很有效的算法对另一数据库有可能完全无效，因此，应针对具体的挖掘目标和应用对象而设计不同的算法。目前具有代表性的方法有以下几类。

1. 决策树方法

决策树表示形式简单，所发现的模型也易于为用户理解，是挖掘分类知识中最流行的方法之一。它利用信息论中的信息熵作为结点分类的标准，建立决策树的一个结点，再根据属性当前的值域建立结点的分支。决策树的建立是一个递归过程。在知识表示方面具有直观、易于理解等优点。最早的决策树算法是 ID3 方法，它对较大的数据集处理效果较好。在 ID3 的基础上，Quinlan 又提出了改进的 C4.5 算法。

2. 模糊集方法

模糊集方法是利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析，是一种应用较早的处理不确定性问题的有效方法。系统的复杂性越高，模糊性越强。模糊集理论是用隶属度来刻画模糊事物的亦此亦彼性的。

在很多场合，数据挖掘任务所面临的数据具有同样的模糊性和不精确性，因此把模糊数学理论应用于数据挖掘则顺理成章。使用模糊集方法可以对已挖掘的大量的关联规则的有用性、兴趣度等进行评判，也可用于分类、聚类等数据挖掘任务。

3. 神经网络方法

神经网络是指一类计算模型，它模拟人脑神经元结构及某些工作机制，利用大量的简单计算单元连成网络来实现大规模并行计算，它有并行处理、分布存储、高度容错、自组织等诸多优点，因此它是数据挖掘中的重要方法。近年来人们研究从训练后的神经网络中提取规则的方法，从而推动了神经网络在数据挖掘分类问题中的应用。神经网络的知识体现在网络连接的权值上，它是一个分布式矩阵结构；神经网络的学习体现在神经网络权值的逐步调整上。在数据挖掘中应用最多的是前馈式网络。它以感知器、反向传播模型、函数型网络为代表，可用于预测、模式识别等方面。

4. 粗糙集方法

粗糙集是一种刻画具有信息不完整、不确定系统的数学工具，能有效地分析和处理不精确、不一致、不完整等各种不完备信息，并从中发现隐含的知识，揭示潜在的规律。粗糙集的核心概念是不可区分关系以及上近似、下近似等。对于给定的一个信息表，粗糙集的方法是通过等价类的划分寻找信息表中的核属性和约简集，然后从约简后的信息表中导出分类/决

策规则。对信息表进行属性约简，获得和原信息表相同信息分布的子表，提高了数据挖掘的效率，并且使得获得的知识更为简单、易于理解。属性约简是数据挖掘中数据预处理阶段的重要环节。

粗糙集理论具有良好的数学性质和可解释性，但在应用于实际数据时，还需要解决复杂度高、数据中的噪声等问题。

5. 统计分析方法

统计方法是从事物的外在数量上的表现去推断该事物可能的规律性，统计分析的本质是以数据为对象，从中获取规律，为人类认识客观事物，并对其发展趋势进行预测、决策和控制提供有效的依据。统计分析方法在数据挖掘中有许多应用，理论也最为成熟。常见的统计方法有回归分析、判别分析、差异分析、聚类分析、描述统计、相关分析和主成分分析等。

6. 可视化方法

可视化是把数据、信息和知识转化为可视的表示形式的过程，其内涵是将数据通过图形化、地理化真实而形象地表现出来并且找出数据背后蕴含的信息，其本质是从抽象数据到可视结构的映射。

可视化技术是20世纪80年代后期提出的一个全新的研究领域。通过丰富的图形表现能力，可视化技术能够准确地表达原始数据、挖掘过程、挖掘结果，使用户可以深入地理解问题并选择更适当的数据挖掘算法，达到深入剖析数据的目的。其特点为：信息可视化的焦点在于信息；信息的数据量很大；信息的来源多种多样等。可视化数据挖掘拓宽了传统的图表功能，使用户对数据的剖析更清楚。

7. 生物智能算法

生物智能算法在优化与搜索应用中前景广阔，用于数据挖掘中，常把任务表示成优化或搜索问题，利用生物智能算法可以找到最优解或次优解。生物智能算法主要包括以下几个方面。

（1）遗传算法。遗传算法是由John Holland于1975年提出的一种有效地解决最优化问题的方法，是一种基于生物进化理论的技术。其基本观点是“适者生存”，用于数据挖掘中，则常把任务表示为一种搜索问题，利用遗传算法强大的搜索能力找到最优解，是一种仿生全局优化方法。遗传算法作用于一个由问题的多个潜在解（个体）组成的群体上，并且群体中的每个个体都由一个编码表示，同时每个个体均需依据问题的目标函数而被赋予一个适应值。遗传算法是多学科结合与渗透的产物，它广泛应用在计算机科学、工程技术和社会科学等领域。

（2）蚁群算法。蚁群算法是由意大利学者Dorigo M.等人在20世纪90年代初首先提出来的。它是一种新型仿生类进化算法，是继模拟退火、遗传算法、禁忌搜索等之后的又一启发式智能优化算法。蚂蚁有能力在没有任何提示的情况下找到从巢穴到食物源的最短路径，并且能随环境的变化，适应性地搜索新的路径，产生新的选择。蚁群算法成功地应用于求解TSP、二次分配、图着色、车辆调度、集成电路设计及通信网络负载等问题。

（3）粒子群优化算法。粒子群优化（PSO）算法是一种基于群体智能的随机优化算法，源于对鸟群或鱼群群体运动行为的研究。由于PSO算法概念简单、易于实现、调整参数少，现已广泛地应用于许多工程领域。然而，粒子群优化算法具有易于陷入局部极值点、进化后

期收敛慢、精度较差的缺点，为了克服粒子群优化算法的缺点，目前出现了大量的改进粒子群优化算法。

(4) 人工鱼群算法。人工鱼群算法(AFSA)是李晓磊等人于2002年提出的一种基于动物自治的优化方法，是集群智能思想的一个具体应用。它的主要特点是不需要了解问题的特殊信息，只需要对问题的解进行优劣的比较，通过各人工鱼个体的觅食、聚群和追尾等局部寻优行为，最终在群体中使全局最优解突显出来。该算法具有良好的求解全局极值的能力，收敛速度较快。

3.7 数据挖掘的典型应用领域

数据挖掘技术源于商业的直接需求，并在各种领域都有广泛的使用价值。数据挖掘已在金融、零售、医药、通信、电子工程、航空、旅馆等具有大量数据和深度分析需求、易产生大量数字信息的领域得到广泛使用，并带来了巨大的社会效益和经济效益。它既可以检验行业内长期形成的知识模式，也能够发现隐藏的新规律。将数据挖掘用于企业信息管理，虽然面临着很大的挑战和许多亟待解决的问题，但有充分的理由相信，这些问题将随着各应用领域的信息化推进逐步得到解决，数据挖掘的应用前景十分乐观。

1. 金融领域的应用

在金融方面，银行和金融机构往往持有大量关于客户的、各种服务的以及交易事务的数据，并且这些数据通常比较完整、可靠和高质量，这大大方便了系统化的数据分析和数据挖掘。在银行业中，数据挖掘被用来建模、预测，识别伪造信用卡，估计风险，进行趋势分析、效益分析、顾客分析等。在此领域应用的数据挖掘，可以进行贷款偿付预测和客户信用政策分析以调整贷款发放政策，降低经营风险。信用卡公司可以应用数据挖掘中的关联规则来识别欺诈。股票交易所和银行也有这方面的需要。对目标客户群进行分类及聚类，以识别不同的客户群，为不同的客户提供更好的服务，以推动市场。此外，还可以运用数据分析工具找出异常模式，以侦破洗钱和其他金融犯罪活动。智能数据挖掘利用了广泛的高质量的机器学习算法，能够在应付大量数据的同时保证理想的响应时间，使得市场分析、风险预测、欺诈管理、客户关系管理和竞争优势分析等应用成为可能。

2. 网络金融交易应用

从网络金融角度来看，网络金融是指通过互联网进行的金融交易。这种交易具有速度快、交易量大、交易次数多、交易人所在地分散的特点。这种基于生产力水平的加速常常超出生产力本身的发展速度，使人类进入脆弱的虚拟经济时代。在股市交易中，人们的兴趣在于预测股市起伏，并且各种各样的算法都曾经被使用过。有的算法在一种情况下有效或在一段时间内有效，有的算法更能捕捉转瞬即逝的个股买/卖点或在众多股票中选出应买卖的股票。金融时序数据是一种常见的数据结构，在这一方面，已有不少学者研究了对其进行挖掘的一般性问题或框架。对股市进行动态数据挖掘，可以随时掌握由大量数据所反映的金融市场暗流。此外，还可以将监管搜索范围完全扩大到一般的网页上，借助一定的文字分析技术提高准确率。

另一方面的应用是研究股市炒作的快速检测算法和技术。互联网的出现和使用也只是

近十年的事，而标志着金融领域重要突破的中国股市的产生和发展也正好在这十余年。电子交易每天产生的海量数据已超出人工处理的能力，但这正使得应用计算机算法进行智能自动监控成为可能。从证监会的角度看，可以通过各种交易数据发现异常现象和相应的操作，识别出哪些是合法炒作，哪些是非法炒作。

3. 零售业务应用

在零售业方面，计算机使用率越来越高，大型超市大多配备了完善的计算机及数据库系统。零售业积累的大量销售数据、顾客购买历史记录、货物进出与服务记录等数据中真正有价值的信息是哪些？这些信息之间有哪些关联？回答这些问题就需要对大量的数据进行深层分析，从而获得有利于商业运作、提高竞争力的信息。数据挖掘技术有助于识别顾客购买行为，发现顾客购买模式和趋势，改进服务质量，取得更高的顾客保持力和满意程度，降低零售业成本。

通常企业所掌握的客户信息特别是以前购买行为的信息中，可能正包含着这个客户决定他下一个购买行为的关键信息，甚至是决定性因素。这个时候的数据挖掘的作用就体现为它可以帮助企业寻找到那些影响顾客购买行为的信息和因素。对这些丰富数据资源的挖掘，可有助于识别顾客购买行为，发现顾客购买模式和趋势，改进服务质量，取得更高的顾客满意程度，提高销量。

还有一个问题就是研究超市顾客的购买行为，这是一种典型的时间序列挖掘问题。在零售服务业中，直接给潜在的顾客寄广告是一种常见的办法。通过分析人们的购买模式，估计他们的收入和孩子数目，作为潜在的市场信息。在庞大的数据集中找出哪些人适合寄广告或折扣券，哪些人会喜欢哪一类的折扣券，哪些人应给予的折扣多一些，哪些产品摆在一起会比分别放在各自的类中卖得更快更多，这都成了数据挖掘的任务。

零售业中数据挖掘的成功应用包括：①销售、顾客、产品、时间和地区的多维分析；②对促销活动有效性的分析，以此提高企业利润；③对顾客忠诚度的分析，以留住老顾客，吸引新顾客；④挖掘关联信息，以形成购买推荐和商品参照，以帮助顾客选择商品，提高销量。

4. 医疗电信领域应用

在医疗领域中，成堆的电子数据可能已放在那儿很多年了，比如病人、症状、发病时间、发病频率以及当时的用药种类、剂量、住院时间等。在药物实验中，可能有很多种不同的组合，每种若均加以实验则成本太大，决策树方法可以用来大大减少实验次数，这种方法已经被许多大的制药公司所采用。生物医学的大量研究大都集中在 DNA 数据的分析上，人类大约有10^5个基因，一个基因通常由成百个核苷按一定序列组成，核苷按不同的次序可以组成不同的基因，几乎不计其数。因此，数据挖掘成为 DNA 分析中的强大工具，如对 DNA 序列间的相似搜索和比较；应用关联分析对同时出现的基因序列的识别；应用路径分析发现在疾病不同阶段的致病基因等。

电信业已经迅速从单纯的提供市话和长话服务演变为综合电信服务，如语音、传真、寻呼、移动电话、图形、电子邮件、互联网接入服务等。电信市场的竞争也变得越来越激烈和全方位化。目前，不管是住宅电话还是移动电话，每天的使用量很大。对电话公司来讲，如何充分使用这些数据为自己赢得更多的利润就成了主要问题。利用数据挖掘来帮助理解商业

行为、对电信数据多维分析、检测非典型的使用模式以寻找潜在的盗用者、分析用户一系列的电信服务使用模式来改进服务、根据地域分布疏密性找出最急需建立网点的位置、确定电信模式、捕捉盗用行为、更好地利用资源和提高服务质量，是非常必要的。借助数据挖掘，可以减少很多损失，保住顾客。

数据挖掘在电信业的应用包括：①对电信数据的多维分析；②检测非典型的使用模式以寻找潜在的盗用者；③分析用户一系列的电信服务使用模式来改进服务；④搅拌分析等。

3.8 数据挖掘的发展趋势

数据挖掘是一门综合性学科，一个多学科交叉的研究领域。它融合了数据库技术、人工智能、机器学习、统计学、知识工程、信息检索、高性能计算及数据可视化等许多学科的概念、理论、方法和技术。经过 20 年的研究和实践，数据挖掘已经吸收了许多学科的研究成果，成为独具特色的研究分支。数据挖掘的概念已经被广泛接受，并吸引了一大批学者投入到数据挖掘的研究领域。

经历了 20 年的发展，包括统计学、人工智能等在内的许多理论和技术成果已经被成功应用到数据挖掘中。数据挖掘的理论体系是由数据库、人工智能、数理统计、计算机科学以及其他方面的学者在探讨性的研究中创立的。这些理论本身的发展和应用为数据挖掘提供了有价值的理论和应用积累。

随着数据挖掘在学术界和工业界的影响越来越大，数据挖掘的研究向着更深入和实用的技术方向发展。从事数据挖掘研究的人员主要在大学、研究机构，也有部分在企业或公司。所涉及的研究领域很多，研究集中在学习算法的研究、数据挖掘的实际应用以及有关数据挖掘的理论等方面。

分析目前的研究和应用现状，数据挖掘在如下几个方面需要重点开展工作。

(1) 数据挖掘理论与算法的研究。数据挖掘继承和发展了相关基础学科已有的成果，探索出许多独具特色的理论体系。但是，这绝不意味着数据挖掘理论的探索已经结束，相反地，它留给了研究者丰富的理论课题。一方面，在这些大的理论框架下有许多面向实际应用目标的挖掘理论等待探索和创新；另一方面，随着数据挖掘技术本身和相关技术的发展，新的挖掘理论的诞生是必然的，而且可能对特定的应用产生推动作用。新理论的发展必然促进新的挖掘算法的产生，这些算法可能扩展挖掘的有效性，如数据挖掘的某些阶段、某些数据类型、大容量源数据集等；可能提高挖掘的精度或效率；可能融合特定的应用目标，如 CRM、电子商务等。因此，对数据挖掘理论和算法的探讨将是长期而艰巨的任务。

(2) 复杂数据类型的挖掘问题。许多数据集中包含着复杂的数据类型，如关系型数据、半结构化数据、非结构化数据、复杂的数据对象、超文本数据和多媒体数据、空间和时间数据、视频数据、声音数据等，局域网和广域网上连接了许多数据源并形成了巨大的、分布式的、分层的和异构的数据库。这些复杂数据类型的数据集，对数据挖掘提出了新的挑战。目前，数据挖掘主要处理的是数值型数据和分类数据，针对非结构化数据、时空数据、多媒体数据的数据挖掘仍是迫切需要解决的问题。

(3) 数据挖掘语言与数据挖掘的可视化。标准的数据挖掘语言或其他方面的标准化工

作将有助于数据挖掘的系统化开发，改进多个数据挖掘系统和功能间的相互操作。可视化对于一个数据挖掘系统来说非常重要，除了要和良好的交互性技术结合外，还要在挖掘结果的可视化、挖掘过程的可视化以及可视化指导用户挖掘等方面进行探索和研究。数据挖掘语言和可视化将促进数据挖掘在企业和社会中的应用。

（4）数据挖掘的性能问题。数据挖掘的性能包括数据挖掘算法的有效性、可伸缩性和并行处理能力。数据挖掘算法的效率和可伸缩性是指为了有效地从数据库中抽取有用的知识，数据挖掘算法必须是有效的和可收缩的。也就是说，一个数据挖掘算法在大型数据库中的运行时间必须是可预计的和可接受的。许多现有的数据挖掘算法往往适合于常驻内存的、小数据集的数据挖掘，而大型数据库中存放了 TB 级的数据，所有数据无法同时导入内存。所以，从数据库的观点来看，有效性和可伸缩性是实现数据挖掘系统的关键问题。

（5）数据挖掘系统的架构。虽然经过多年的探索，数据挖掘系统的基本架构和过程已经趋于明朗，但是受应用领域、挖掘数据类型以及知识表达模式等的影响，在具体的实现机制、技术路线以及各阶段或部件（如数据清洗、知识形成、模式评估等）的功能定位等方面仍需细化和深入研究。目前新颖的数据挖掘框架日益受到重视，如云模型和数据场理论、双库协同机制、基于多智能体的主动型数据挖掘框架等。

（6）交互式数据挖掘技术。由于数据挖掘是在大量的元数据集中发现潜在的、事先并不知道的知识，因此和用户交互式地进行探索性挖掘是必然的。这种交互可能发生在数据挖掘的各个不同阶段，从不同角度或不同粒度进行交互。所以良好的交互式挖掘也是数据挖掘系统成功的前提。

（7）数据挖掘中的私有性问题。数据挖掘可能会导致对私有权的入侵，研究采用哪些措施防止暴露敏感信息是十分重要的。当从不同角度和不同抽象级上观察数据时，数据安全性将受到严重威胁。这时，数据保护和数据挖掘可能会造成一些矛盾的结果。例如，数据安全性保护的目标可能与从不同角度挖掘多层知识的需求相矛盾。

（8）数据挖掘中的不确定性问题。不确定性是客观事物的一个固有特征，尤其在实际应用中存在大量不确定数据。不确定性数据挖掘的任务就是发现隐含在这些不确定数据中的知识，寻找并且能够形式化地表现不确定性的规律性，至少是某种程度的规律性。如果数据挖掘模型不能准确地描述或者没有充分考虑数据挖掘对象的不确定性，那么由数据挖掘模型得到的结果是不可信的，甚至是错误的。

（9）数据挖掘中的动态性问题。传统的数据挖掘是从静态的数据库中发现知识，许多实际数据库系统中的数据不是稳定不变的，而是不断递增和变化的，这种改变可能使先前发现的模式无效，因此发现知识或模式也需要动态维护，及时更新。为了随时获得一个与数据相关的有效模式，需要以一定的不多的时间间隔重复同样的数据分析过程。由于某些数据挖掘过程的高成本，产生了对增量数据挖掘算法的研究需求。开发增量式数据挖掘算法并与数据库更新操作相结合，可以提高数据挖掘的效率，不必重新挖掘整个数据库。因此，需要研究新的动态数据挖掘算法来应对以增量形式获得的新数据。

数据挖掘将成为对工业生产乃至日常生活产生重要影响的技术之一。随着数据挖掘理论与方法的进一步完善和计算机处理能力的进一步提高，数据挖掘无论在理论上还是在应用上都将得到更大的发展，数据挖掘将产生深远的社会影响。一方面越来越多的研究人员将投入到数据挖掘的研究中；另一方面广大的用户也将逐渐看到它的价值。随着众多数据

挖掘研究人员对于技术的不断改进，软件供应商所提供的工具的不断完善，数据挖掘技术的应用和开发不再是专业人士的专利，而成为一项经过一定培训就可以为人们所利用的普及的工具。同时更多的软件隐含地把数据挖掘作为它们的功能部件，使用户感觉不到它们的存在，这种隐含的应用将成为普通大众执行数据挖掘的重要手段。

小结

本章介绍了数据挖掘的起源及其发展、定义、数据挖掘所要解决的问题、数据挖掘的过程以及数据挖掘系统。数据挖掘来自实际领域的需求，其理论与方法涉及多个学科知识的交叉，在生产实践、商业活动中获得了成功的应用，是数据智能化的积极推动因素。目前，各个领域都对数据挖掘提出了新的要求，也为数据挖掘的发展提供了强大的发展动力。

习题

1. 数据挖掘的特点是什么？怎么定义数据挖掘？
2. 数据挖掘的过程是什么？
3. 数据挖掘的基本功能有哪些？谈谈你对其的理解。
4. 数据挖掘方法有哪些？谈谈你对其的理解。
5. 上网查找数据挖掘的一些应用，并谈谈你对数据挖掘的大致认识。