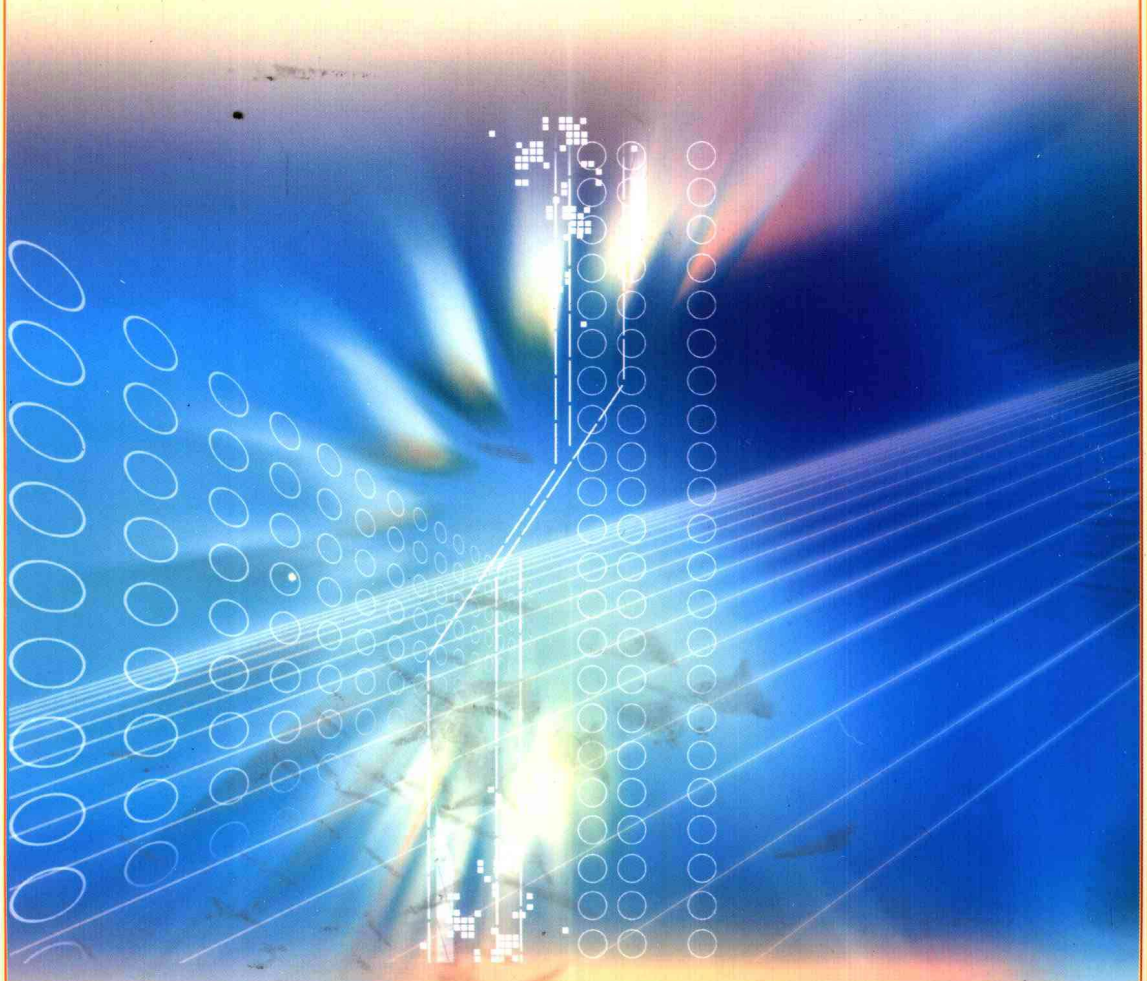
国家自然科学基金资助项目(50674086)
高等学校博士学科点专项科研基金项目(20060290508)

煤矿安全监测监控数据知识发现方法

孟凡荣　编著



中国矿业大学出版社
China University of Mining and Technology Press

国家自然科学基金资助项目（50674086）
高等学校博士学科点专项科研基金项目（20060290508）

煤矿安全监测监控数据知识发现方法

孟凡荣　编著

中国矿业大学出版社

内 容 提 要

知识发现作为从大量数据中发现有用知识的技术，可以应用于煤矿安全监测监控数据的处理，并从中找出对煤矿灾害防治有用的知识。

本书以煤矿安全监测监控数据为研究对象，利用先进的知识发现方法，从大量的监测监控数据中发现对煤矿安全有意义的知识。主要内容包括：知识发现的基本概念、基本任务、一般处理过程和知识发现的应用领域；煤矿安全监测监控系统现状、组成和数据的获取与转换方法；数据预处理的一般方法和具体应用在安全监测监控数据的处理方法；煤矿安全监测监控数据知识发现系统的体系结构；以及基于云理论的煤矿安全监测监控数据关联规则挖掘、基于语义描述的煤矿安全监测监控数据聚类分析和基于粗糙集与神经网络的煤与瓦斯突出预测方法。

本书结构紧凑，内容丰富，可作为计算机专业及安全监测监控和安全技术工程等相近专业研究生教材，也可以供相关专业的工程技术人员参考使用。

图书在版编目(CIP)数据

煤矿安全监测监控数据知识发现方法/孟凡荣编著.
徐州：中国矿业大学出版社，2008.11
ISBN 978-7-5646-0109-6

Ⅰ.煤… Ⅱ.孟… Ⅲ.①矿山安全－监测系统－数据处理②矿山安全－监控系统－数据处理 Ⅳ.TD76

中国版本图书馆 CIP 数据核字(2008)第 170123 号

书　　名　煤矿安全监测监控数据知识发现方法
编　　著　孟凡荣
责任编辑　李士峰　周　丽
出版发行　中国矿业大学出版社
（江苏省徐州市中国矿业大学内　邮编 221008）
网　　址　http://www.cumtp.com　**E-mail**：cumtpvip@cumtp.com
排　　版　徐州中矿大印发科技有限公司排版中心
印　　刷　徐州中矿大印发科技有限公司
经　　销　新华书店
开　　本　787×1092　1/16　**印张** 10.5　**字数** 249 千字
版次印次　2008 年 11 月第 1 版　2008 年 11 月第 1 次印刷
定　　价　42.00 元

前　言

煤炭是我国重要的基础能源和原料，在国民经济中具有重要的战略地位。在我国一次能源结构中，煤炭将长期是主要能源。改革开放以来，煤炭工业取得了长足发展，煤炭产量持续增长，生产技术水平逐步提高，煤矿安全生产条件有所改善，对国民经济和社会发展发挥了重要的作用。在今后很长时期内，煤炭在我国一次能源消费结构中仍将占第一位。因此，煤矿安全是我国安全生产工作的重中之重，始终受到党中央、国务院的高度重视。煤炭工业的持续、稳定、健康发展，直接关系着我国能源安全和全面建设小康社会目标的实现。

随着煤炭需求的增加以及煤炭工业生产规模的不断扩大，煤矿事故已经引起全社会的共同关注，特别是煤矿重大和特大恶性事故的时有发生，不仅直接关系国家的能源安全，影响煤炭工业的健康发展，同时也直接关系着煤炭职工的生命安全和职业健康，关系和谐社会的构建。提高煤矿安全保障程度是煤炭工业持续、稳定、健康发展的前提和基础，是落实以人为本的科学发展观的具体体现，是构建和谐社会的内在要求。预防和控制煤矿重大与特大事故的发生、促进煤矿安全生产形势的根本好转成为社会各界关注的重要问题。根据调查分析，有关专家认为煤矿生产管理和安全监测预警技术水平不能满足要求的主要原因之一是对涉及全矿井复杂的数据资源知识挖掘的深度不够。

随着国家对煤矿安全生产的重视和关注，煤矿企业十分重视对煤炭生产安全的监测监控，先后构建了多种形式的监测监控系统，这些监测监控系统在日常运行过程中产生了大量的以文本、图纸、数值、声音、图像等形式的监测监控数据资源，其中蕴含着丰富的、与安全生产相关联的知识。这些数据具有资源数量巨大，来源不同，数据格式和存取形式不同，数据之间耦合度不同等特点。如何从这些复杂数据中挖掘与安全生产和经营管理密切相关的知识，对指导安全生产和提高经济效益十分重要。

因此，以煤矿安全生产和灾害防治理论为指导，以煤矿各种安全监测监控复杂数据为数据源，通过对现有煤矿安全生产过程中各种复杂数据的统计整理，及时有效地引入相近学科中的新理论和新方法，通过知识学习，并从中提取关键信息，不但可以为煤矿安全生产和决策指挥提供直观、可靠的知识，而且可以为煤矿安全生产建设决策支持保障系统做好基础工作，也为数字矿山基础理论研究提供了一定的技术支持。

本书共分为 7 章。第 1 章主要介绍煤矿监测监控技术和煤矿安全监测监控数据的复杂性，介绍知识发现的产生、基本概念、基本任务、一般处理过程和知识发现的应用领域，给出了本书的研究目标及内容。第 2 章介绍煤矿安全监测监控系统的现状和组成、分析存在的问题和煤矿安全监测监控数据的主要特点，同时阐述煤矿安全监测监控数据的属性、获取和转换方法。第 3 章叙述数据预处理的基本概念和一般方法，对相关的安全监测监控数据源进行了数据的集成，并完成不完备数据填补方法和离群点处理方法的实现。第 4 章针对煤

矿安全监测监控数据知识发现系统的特点，在分析典型知识发现系统结构的基础上，提出了煤矿安全监测监控数据知识发现系统的系统模型和体系结构。第5章概述了关联规则的基本概念、分类和挖掘过程，分析了典型的关联规则挖掘算法；在云理论和关联规则研究的基础上提出了一种基于云模型的属性空间软化分方法，研究传统的关联规则挖掘方法和煤矿安全监测数据本身的特点，对 Apriori 算法进行了改进，提出了适用于对煤矿安全监测数据进行挖掘的算法。第6章主要介绍本体知识和聚类分析知识，针对基于语义描述的煤矿安全监测数据的特点，给出了基于语义描述的煤矿安全监测数据的混合相似性度量方法；基于该相似性度量方法，给出了改进的 CURE 聚类算法。第7章在粗糙集、神经网络等相关理论研究的基础上，提出了基于粗糙集与神经网络的煤与瓦斯突出预测模型，并设计了模型的关键模块。算法首先利用粗糙集理论对突出样本数据进行属性约简，提取最简的规则集，然后利用最简规则集对神经网络进行构建，这样既减少了神经网络的学习训练时间，又提高了预测的准确度。

本书是作者近年来在数据库技术、数据挖掘和知识发现研究成果的基础上总结而成的。前人的研究成果和国内外许多论文著作给予作者很大的启示和帮助，在此表示谢意。值此著作完成之际，深深感谢夏士雄教授的理解、帮助和支持，感谢马小平教授的的关心和支持，感谢王恩元教授等有益的建议、指导和帮助。著作的完成，得到了同研究室的周勇博士、张磊博士、牛强、闫秋艳、张辰等教师的大力帮助，得到了作者诸多研究生的大力支持和协助，在此一并表示深深的谢意。同时感谢中国矿业大学出版社对本书出版所给予的大力支持。

书中提出的模型、方法和算法等还有待于今后进行更深入和细致的研究。

由于作者水平有限，书中不足之处在所难免，恳请读者批评指正。

作　者

2008 年 7 月

目　录

第1章 绪 论

煤炭是我国重要的基础能源和原料，在国民经济中具有重要的战略地位。改革开放以来，煤炭工业取得了长足发展，煤炭产量持续增长，生产技术水平逐步提高，煤矿安全生产条件有所改善，对国民经济和社会发展发挥了重要的作用[1]。在今后很长时期内，煤炭在我国一次能源消费结构中仍将占第一位。煤炭作为主体能源，是我国能源安全的基石。根据专家的预测，到2020年，要使国内生产总值在2000年的基础上再翻两番，将消耗40亿t标准煤。考虑到节能因素，2020年我国能源消费总量也将高达30亿～32亿t标准煤，其中原煤22亿～23亿t，甚至更多。因此，煤矿安全是我国安全生产工作的重中之重，始终受到党中央、国务院的高度重视[2]。煤炭工业的持续、稳定、健康发展，直接关系着我国能源安全和全面建设小康社会目标的实现。

随着煤炭需求的增加以及煤炭工业生产规模的不断扩大，煤矿事故已引起全社会的共同关注，特别是煤矿重大和特大恶性事故的时有发生，不仅直接关系国家的能源安全，影响煤炭工业的健康发展，同时也直接关系着煤炭职工的生命安全和职业健康以及和谐社会的构建。提高煤矿安全保障程度是煤炭工业持续、稳定、健康发展的前提和基础，是落实以人为本的科学发展观的具体体现，是构建和谐社会的内在要求。根据调查分析，有关专家认为煤矿生产管理和安全监测预警技术水平不能满足要求的主要原因之一是对涉及全矿井复杂的数据资源知识挖掘的深度不够。

1.1 煤矿监测监控技术

1.1.1 煤矿监测监控技术现状

国外煤矿监测监控技术起步较早，发展较快，工业以太网与现场总线相结合的监控系统体系结构已经相当成熟，并广泛应用于生产安全监控领域。我国煤矿监控系统已经从早期的引进、消化、吸收发展到了现有的自主创新开发。目前，在用的监测系统主要有KJ90、KJ95、KJF2000等等。随着信息化水平的提高、煤炭行业信息化和网络化的发展以及电子技术的迅猛发展，安全生产信息化也得到了高速发展，信息技术推动了安全生产技术和监测监控技术的发展，促进了各类传感器、数据传输技术在煤矿领域的应用。

近年来国内相继研制出的一批煤矿监控系统和地面工业集中控制系统，出现了集环境安全监测监控、煤矿生产监控、皮带监控、瓦斯抽放泵站监控和信息管理于一体的多功能综合性监测监控系统，技术水平有了较大提高。但是，我国煤矿监控系统还停留在集散式监控水平上，仍以主从式窄带低速通信体系结构和时分制通讯为主流技术。这些监控系统虽对改善我国煤矿安全生产状况起到了积极作用，但现有煤矿监控技术的体系结构、信息传输方

式、监控系统反应速度等方面的技术“瓶颈”日益突出，严重制约了煤矿现代化生产的需要和监控领域产业的发展。在传感器技术方面，日本成功开发了光干涉甲烷传感器，并得到了广泛的应用；英国、德国和美国等开发的红外甲烷传感器，使传感器反应速度和可靠性大为提高，也得到了广泛应用；我国相继研制开发出了热催化、电化学、光干等原理的瓦斯检测传感器。多年来我国一直还是以载体催化原理的传感器为主导，它具有价格便宜、可靠和更换维护方便等优点，并制定了相应的检定标准和方法。但是，受多种因素的影响，我国自行研制开发的传感器，在使用寿命、灵敏度、调校周期等技术性方面与国外先进水平相比差距还很大，集中表现在响应速度慢（检测反应时间要 30 s）、选择性差、受硫化氢气体干扰大、高浓度瓦斯易造成中毒而无法恢复、使用寿命短（半年到 1 年）、调校期短（7 天）、零漂严重等缺点。光电传感器与红外传感器仅仅处在研发探索阶段，试制出来的样机仍有许多关键性技术问题没有得到根本解决，离大规模推广应用还有很长一段距离。在灾害的监测预警方面，我国与世界其他先进的采煤国家，如美国、俄罗斯、法国、澳大利亚、波兰、日本等类似，也开展了煤矿危险性评价技术的研究，但国内外的研究大多都是针对单一灾害的危险源进行静态评价，缺乏动态分析和评价。

当前，我国煤矿监控系统的主要问题突出表现为：一是监控系统传输网络体系结构亟待升级，安全和生产动态信息的传输缺乏宽带、快速、可靠的通讯平台。二是监控系统配套传感器种类不全、技术性能差。三是煤矿安全生产综合监控系统开放性差，智能分析决策技术水平低。由于没有统一的通信协议和接入技术，生产监控、安全监控和各类灾害监测子系统设备之间不能互联互通，数据、语音、图像不能有效集成。监控系统只能对原始数据进行简单的转换、存储、显示和打印，对有限的数据资源“挖掘”和分析深度不够，灾害隐患判别和应急救援决策的信息量不足，不能及时发现重大灾害隐患。四是矿井重大灾害预测预报技术实用性和准确性不高，不能有效指导安全生产。因此，煤矿安全生产监测监控系统向宽带、快速的“以太网＋现场总线”技术方向发展，已成为不争的事实。现场总线控制系统具有全开放性、全数字化、双向传输、互可操作性与互用性、现场设备智能化、功能自诊断、系统结构高度分散、对现场环境的适应强、节省布线及控制室空间等优点。从国际发展趋势看，现场总线是工业控制系统的新型通讯标准，是低成本自动化系统技术。现场总线技术的采用将带来工业控制系统技术的革命。现场总线技术可以促进现场设备的智能化、控制功能的分散化、控制系统的开放化，符合工业控制系统领域的技术发展趋势。安全生产监测监控系统的发展也将适应这一趋势。

1.1.2 煤矿安全监测监控数据的复杂性

随着国家对煤矿安全生产的重视和关注，煤矿企业十分重视对煤炭生产安全的监测监控，先后构建了多种形式的监测监控系统。这些监测监控系统在日常运行过程中产生了大量文本、图纸、数值、声音、图像等形式的监测监控数据资源，其中蕴含着丰富的、与安全生产相关联的知识资源。这些数据资源具有数量巨大、来源不同、数据的格式和存取形式不同和数据之间耦合度不同等特点。如何从这些复杂数据中挖掘与安全生产和经营管理密切相关的知识，对指导安全生产和提高经济效益十分重要。

近年来，煤矿企业已经陆续建立起各种安全生产管理监控系统，对井下安全和工况进行监测监控，及时记录井下工作状况，实时有效地将这些数据提供给有关领导及技术人员，对

保证煤矿的安全和正常生产起到了重要作用。但是，由于生产厂商和系统建设时期不同，各个系统之间没有统一的通信协议和接入技术，系统之间的数据结构差异很大，呈现出多源性和异构性；同时，由于煤矿开采的对象——煤矿床是分布于三维地理空间、并随着开采进程不断变化着的地质实体，而矿山安全生产的一切过程都离不开三维空间，无论煤层、构造等地质实体，还是纵横交错的井下巷道系统和各种监测监控信息都具有空间属性。不仅如此，煤矿生产活动始终处于一种随时间动态变化的复杂系统之中，如果反映其实际状态的各种数据得不到有效集成，就只能形成彼此隔离的“信息孤岛”，其中蕴含的丰富知识得不到有效地综合利用，更谈不上有效地为煤矿安全提供决策依据。

针对煤矿企业生产的特点，通过对国内外相关领域技术的分析比较，我们发现由于各种煤矿开采过程中获得的监测监控数据在时间、空间和属性方面的差异很大，数据对象结构具有很大的复杂性。煤矿数据源中的数据对象之间往往具有多层关联结构，数据对象的成分可以缺失、多次重复、有序出现或者无序出现，甚至在较高层次上是有序的，在较低的层次上却是无序的（即有序成分中包含着无序的子成分），或反之（即无序成分中包含着有序的子成分），这样就构成了有序成分和无序成分间的复杂关联组合，这些特点构成了煤矿安全监测数据对象结构的关联复杂性。面对庞大复杂的数据源，对于煤矿综合数据处理而言，不是数据源的不足，而是从这些复杂数据源中提取更丰富、更有用和更可靠知识的能力受到现有技术手段的限制。首先，不可能重复低水平的、以人工作业方式为主的常规数据处理方法；其次，各种单一的数据库分析处理手段获取的数据，在时间和空间分辨率等方面存在明显的局限性和差异性，导致其应用分析能力有限，仅仅利用一种数据难以满足要求；特别是由于煤矿生产是一个分散性大、区域特色浓、环境干扰因素繁多的复杂大系统，对数据分析的目标具有多样性，而复杂目标无论在表达上还是在处理上均与领域知识有关，并且在多样性目标下对数据集合的分析，目前还没有现成的且满足可计算条件的一般性理论与方法。所以当前的研究存在一些共性的问题，如数据与模型的分离问题，即对复杂数据进行分析后生成模型，随着数据的变化，模型相应地发生变化，但是，现有系统将模型和数据分离，导致数据的变化不能及时反映到模型上。

与单源数据相比，煤矿安全监测监控数据具有冗余性、互补性和关联性等一系列特点。冗余性是指各种数据对目标的表示、描述或解析结果具有重复性，对冗余数据的处理分析，可以去除一定的误差和不确定性，从而提高识别率和精确度；互补性是指数据来自不同自由度且相互独立的系统，互补数据的处理能够提高最终结果的可信度；关联性是指不同系统在观测和处理数据时对其他数据有依赖关系，关联性的应用可提高系统整体协调性能。因此，把这些复杂数据加以分析和利用，获得生产过程中重要对象的重要知识，可以为煤矿灾害防治提供科学依据。

煤矿安全监测监控数据是对目标的各种属性或特征以及背景或环境数据给出的定量表示，它们之间不仅常常存在不同程度的关联性或互补性，而且由于矿山开采环境的复杂性、传感器或观测者本身的局限性、信息获取技术或方法的不完善性，这些复杂数据通常表现出诸多不确定性，如随机性、模糊性、不完整性和冗余性等。面向煤矿安全监测监控数据的知识学习方法研究就是针对煤矿安全监测监控数据的特点，在数据层面和特征层面同时进行分析，挖掘数据与数据之间、数据与特征之间以及特征与特征之间的关联性，并基于此进行

模式识别，可以看出其是一种综合性的模式识别。

对于安全监测监控复杂数据知识挖掘，首先应该解决复杂数据源的融合问题。目前众多研究者采用不同的方法，大多表现为针对多传感器或多采样对象采用各种智能的数据融合算法[6,7]。国内外许多学者对智能数据融合进行了研究，例如，对于数据融合的智能化算法，权太范和孙庆伟以神经网络和模糊推理理论为基础建立数据融合模型，使数据融合智能化，且具有自适应性和学习性[8]；李晨和韩崇昭等针对多目标跟踪的数据融合问题，通过引入目标的方向性信息，提出了一种基于广义联合概率事件分割组合的新关联算法，使最终的估计值更准确，关联精度得到进一步提高[9]；敬忠良等把模糊神经网络和D—S证据理论相结合，提出了一种智能特征信息融合结构[10]；王婷杰[11]和童树鸿[12]等在模糊理论基础上，建立了具有智能推理特征的数据融合模型，着重处理了融合数据的不一致和冲突问题；此外，也有学者提出了利用知识库系统和模糊理论建立智能融合模型，用于不确定信息处理和智能决策系统[13]。

与此同时，针对煤矿具体情况，如地下开采过程中经常会发生煤与瓦斯突出、冲击矿压、顶板冒落等动力灾害，赵俊三教授根据地理空间信息数据整合的原则与方法，结合具体应用分析信息共享，提出了地理空间数据整合与更新的实现方法[14]；夏士雄教授也对矿山多源数据融合模型及关键技术进行了一定的分析[5]。虽然许多专家利用推理网络和人工智能等相关技术针对矿山的实际情况进行了分析和研究[15~18]，但是大量相关文献表明智能信息处理和知识挖掘的结合并不完善，依然存在许多问题需要进一步研究，尤其是在煤矿生产相对复杂的环境下，知识发现方法等一些关键技术基本上处于起步阶段。

综上所述，以煤矿安全生产和灾害防治理论为指导，以煤矿各种安全监测监控复杂数据为数据源，通过对现有煤矿安全生产过程中各种复杂数据的统计整理，及时有效地引入相近学科中的新理论和新方法，通过知识学习，从中提取关键信息，不但可以为煤矿安全生产和决策指挥提供直观、可靠的知识，而且可以为煤矿安全生产建设决策支持保障系统做好基础工作，也为数字矿山基础理论研究提供了一定的技术支持。

1.2 知识发现

21世纪将是知识经济占主导地位的世纪，一个拥有持续创新能力和大量高素质人才资源的国家，将具备发展知识经济的巨大潜力[20]。

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、查询和统计等功能，但是，无法发现数据中存在的关系和规则，并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知识的手段，导致了“数据爆炸但知识贫乏”的现象。因此，产生了数据挖掘。

数据挖掘是从存放在数据库、数据仓库和其他信息库中的大量数据中挖掘有趣知识的过程[19]。数据挖掘其实是一个逐渐演变的过程，电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习成为人们关心的焦点。随着神经网络技术的形

成和发展，人们的注意力转向知识工程。它不同于机器学习那样给计算机输入范例，让它生成规则，而是直接给计算机输入已被代码化的规则，计算机则通过使用这些规则来解决某些问题。20世纪80年代人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。在20世纪80年代末出现一个新的术语，它就是数据库中的知识发现(Knowledge Discovery in Database，简称KDD)。

这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现的知识都是相对的，是有特定前提和约束条件且面向特定领域的，同时还要能够易于理解，最好能用自然语言表达所发现的结果。

1.2.1 KDD的产生

20世纪以来，计算机与信息科学技术的迅速发展，特别是在数据库技术、人工智能、机器学习以及计算机硬件等方面所取得的令人吃惊的飞速进步，大大推动了商用数据库与信息产业的发展。随之，海量数据的收集、存储成为可能，随着人们获得的数据量的不断增多，人们的注意力越发成为一种宝贵的资源，人们常常埋没于大量的数据中，却又无法得到足量的有用知识。“数据丰富，信息贫乏”成为这一状况的真实写照。特别是在严重依赖数据分析的领域中，而对海量的数据集合和多样化格式的原始数据源，无论是从时间意义上还是从空间意义上，传统的数据分析方法已经很难满足人们对隐藏在数据背后的内在联系和信息的需求。正是在这种情况下，产生了对能更加有效进行数据分析的理论和方法的需求。人们希望有一种新的技术和工具来帮助其完成对大量原始数据的分析任务，进而获得所需的有用知识。在这种客观需要的推动下，KDD应运而生，进而迅速发展，成为数据库研究、开发和应用最活跃的分支之一[21]。

许多人也把数据挖掘(Data Mining，简称DM)视为知识发现的另一个常用术语，这是因为数据挖掘是知识发现过程中最重要的一个步骤，但一般认为它是知识发现的一个处理阶段，是KDD处理过程中的一个关键步骤。知识发现一般在科研领域使用较多，在工程应用领域多称之为数据挖掘，通常可以不加区分地混用[22]。

综上所述，知识发现与数据挖掘(DM)是人工智能、机器学习(Machine Learning)与数据库技术相结合的产物。机器学习是用计算机模拟人类学习的一门科学，始于20世纪60年代末，真正发展是在20世纪70年代，由于当时在专家系统开发中存在知识获取的“瓶颈”现象，所以采用机器学习来完成知识的自动获取。1980年，在美国召开了第一届国际机器学习研讨会；1984年，《机器学习》杂志问世。我国很快跟上了国际步伐，于1987年召开了第一届全国机器学习研讨会。

数据库中知识发现(KDD)一词首次出现在1989年8月在美国底特律召开的第十一届国际联合人工智能学术会议专题讨论会上[20]。1989～1994年举行了四次KDD专题讨论会。随着参加会议人数的增多，以及KDD在学术界和工业界的影响越来越大，国际KDD组委会于1995年把专题讨论会更名为国际会议，1995年在加拿大蒙特利尔市召开了第一届知识发现与数据挖掘的国际学术会议，1998年建立了新的学术组织ACM－SIGKDD，即ACM下的数据库中的知识发现专业组。2008年，ACM－SIGKDD组织了第十四届知识发现与数据挖掘的国际学术会议(KDD－2008)，专题杂志 *Data Mining and Knowledge Dis-*

covery 自 1997 年起由 Kluwers 出版社出版。此外还有一些国际和地区性数据挖掘会议，例如，“知识发现与数据挖掘太平洋亚洲会议”(PAKDD)、“数据库中的知识发现原理与实践欧洲会议”(PKDD)、“数据仓库与知识发现国际会议”(DaWaK)、“ACM－SIGMOD 数据管理国际会议”(SIGMOD)、“超大型数据库国际会议”(VLDB)、“ACM－SIGMOD－SIGART 数据库原理研讨会”(PODS)、“数据工程国际会议”(ICDT)、“扩展数据库技术国际会议”(EDBT)、“数据库理论国际会议”(ICDT)、“信息与知识管理国际会议”(CIKM)、“数据库与专家系统应用国际研讨会”(DEXA)、“数据库系统高级应用国际会议”(DASFAA)、“人工智能国际联合会议”(IJCAI)和“美国人工智能学会会议”(AAAI)等等[19]。表 1-1 列出了历届 KDD 会议的时间、会议名称和会议地址，由此可以看出 KDD 的研究状况和重要程度。

表 1-1　　历届 KDD 国际学术会议一览表

时　间	会议名称	会议地址
1989-6	Workshop on KDD	Detroit, Michigan, USA
1991-7	Workshop on KDD	Anaheim, California, USA
1993-7	Workshop on KDD	Washington, USA
1995-8	KDD－1995	Montreal, Canada
1996-8	KDD－1996	Portland, Oregon, USA
1997-2	PAKDD－1997	Singapore
1997-8	KDD－1997	California, USA
1998-4	PAKDD－1998	Melboume, Australia
1998-8	KDD－1998	New York, USA
1999-4	PAKDD－1999	Beijing, China
1999-8	KDD－1999	San Diego, CA, USA
2000-4	PAKDD－2000	Kyoto, Japan
2000-8	KDD－2000	Boston, MA, USA
2001-4	PAKDD－2001	Hongkong, China
2001-8	KDD－2001	San Francisco, CA, USA
2002-4	PAKDD－2002	Taipei, Taiwan, China
2002-7	KDD－2002	Edmonton, Alberta, Canada
2003-4	PAKDD－2003	Seoul, Korea
2003-8	KDD－2003	Washington, DC, USA
2004-5	PAKDD－2004	Sydney, Australia
2004-8	KDD－2004	Seattle, WA, USA
2005-5	PAKDD－2005	Hanoi, Vietnam
2005-8	KDD－2005	Chicago, Illinois, USA
2006-4	PAKDD－2006	Singapore
2006-8	KDD－2006	Philadelphia, PA, USA
2007-5	PAKDD－2007	Nanjing, China
2007-8	KDD－2007	San Jose, CA, USA
2008-8	KDD－2008	Las Vegas, NV, USA

与国外相比，国内对 KDD 的研究稍晚，没有形成整体力量。但是，国内也有相当多的数据挖掘和知识发现方面的研究成果，许多学术会议上都设有专题进行交流。

1993 年，国家自然科学基金首次支持该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中国科学院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究；北京大学也在开展对数据立方体代数的研究；华中理工大学、复旦大学、浙江大学、中国科技大学、中国科学院数学研究所和吉林大学等单位开展了对关联规则开采算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web 数据挖掘。

最近，Gartner Group 的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术的前两位。根据最近 Gartner 的 HPC 研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点”。

因此，可以看出知识发现的研究和应用受到了学术界和实业界越来越多的重视。

1.2.2 KDD 的相关概念

KDD 一词出现于 1989 年，其定义几经变动，1996 年 U. M. Fayyad 给出了目前公认的定义：KDD 是从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程[24,25]。在上述定义中，涉及几个需要进一步解释的概念：“数据集”、“模式”、“过程”、“有效性”、“新颖性”、“潜在有用性”和“最终可理解性”[26]。

(1) 数据集(F)是一组事实(如关系数据库中的记录)，它记录了事物有关方面的原始信息，如学生档案数据、商场销售数据或者银行客户信息等。由于 KDD 处理的数据是从现实世界中得来的，因而并不能保证所有数据都规范，一般需要对数据进行预处理，使之适于知识提取。

(2) 模式是一个用语言(L)来表示的一个表达式(E)，它可用来描述数据集(F)的某个子集(FE)，E 作为一个模式，要求它比对数据子集的枚举要简单(所用的描述信息量要少)。模式可以看作是知识，它给出了数据的特性或数据之间的关系，是对数据包含的信息更抽象的描述。例如，如果同一信用卡在短时间内被连续使用，则该信用卡可能丢失而被其他人盗用；成绩优秀的学生学习都是非常刻苦的等等。模式的表示方式很多，有时甚至无法用显式的方法进行描述，例如，利用神经网络可以对手写体汉字进行分类；学习结果是神经网络中各个单元之间的连接权值，模式是通过这些连接权值在使用过程中体现出来的。

(3) 过程在 KDD 中通常指多阶段的一个过程，涉及数据准备、模式搜索、知识评价以及反复的修改求精。该过程要求是非平凡的，即要有一定程度的智能性、自动性(仅仅给出所有数据的总和不能算作是一个发现过程)。

(4) 有效性是指发现的模式对于新的数据仍保持有一定的可信度。

(5) 新颖性要求发现的模式应该是新的。

(6) 潜在有用性是指发现的知识在将来有实际效用，如用于决策支持系统里可提高经济效益。

有效性、新颖性、潜在有用性和最终可理解性综合在一起可称之为兴趣性。通过 KDD 从当前数据所发现的模式必须有一定的正确程度和新颖性，否则 KDD 就毫无作用。虽然知识发现可以对已有的知识进行验证，但发现新的知识往往更重要，或者对已有的知识进行拓展以得到更全面、更具有实际意义的知识。发现的知识必须经过实践的检验，并通过在实际应用中发现的问题对学习数据和策略进行修改、重新进行学习，从而得到更精确的知识。一般在使用提取出的知识之前，要使用一些数据进行测试，只有测试结果达到要求才能应用。

(7) 最终可理解性要求发现的模式能被用户理解，目前它主要体现在简洁性上。KDD 的目标就是将数据中隐含的模式提取出来，从而帮助人们更好地了解数据中所包含的信息。但一般知识学习算法得到的模式对普通用户来说很难理解，更不用说使用了。因此，KDD 不仅应该能够将知识提取出来，更应该将发现的知识以直观易用的方式呈现给用户。当然，一个模式是否容易被人理解，这本身就很难衡量，往往需要按照用户能够理解的形式表现出来。

1.2.3 KDD 的基本任务

对于知识发现技术的研究集中于寻求各种问题的解决办法，包括将数据归为不同的种类，刻画一组数据的特征，发现数据项之间的关联和相关性，发现顺序模式及规则数据的相似性[26]。知识发现的基本任务包括以下几个方面：

(1) 数据分类。分类是数据挖掘研究的重要分支之一，是一种有效的数据分析方法。分类的目标是通过分析训练数据集，构造一个分类模型(即分类器)，该模型能够把数据库中的数据记录映射到一个给定的类别上，从而可以用于数据预测。

(2) 数据聚类。当要分析的数据缺乏必要的描述信息，或者根本就无法组织成任何分类模式时，利用聚类函数把一组个体按照相似性归成若干类，这样就可以自动找到类。聚类和分类类似，都是将数据进行分组。但与分类不同的是，聚类中的组不是预先定义的，而是根据实际数据的特征按照数据之间的相似性来定义的。

(3) 衰退和预报。这是一种特殊类型的分类，可以看作是根据过去和当前的数据预测未来的数据状态。通过对用衰减统计技术建模的数字值的预测，学习一种(线性或非线性)功能将数据项映射为一个数字预测变量。

(4) 关联和相关性。它是指发现大规模数据集中项集之间有趣的关联或相关关系。关联规则是指通过对数据库中的数据进行分析，从某一数据对象的信息来推断另一数据对象的信息，寻找出重复出现概率很高的知识模式，常用一个带有置信度因子的参数来描述这种不确定的关系。

(5) 顺序发现。它通常指确定数据组中的顺序模式。当数据的特定类型的关系已被发现时，这些模式同关联和相关性相似。但对关系基于时间序列的数据组，顺序发现和关联就不同了。顺序发现是将数据映射为有关数据组的简练描述的子集或映射为数据库中一组特定用户数据的高度概括的数据。

(6) 描述和辨别。它是指发现一组特征规则，其中的每一条都是显示数据组的特征或者从对比类中区别试验类的概念的命题。

(7) 时间序列分析。其任务是发现属性值的发展趋向，如股票价格指数的金融数据、客