LIDU JISUAN
ZAISHUJU WAJUEZHONG DE
YINGYONG YANJIU

粒度计算在数据挖掘中的应用研究

张霞◎著

中国物资出版社

粒度计算在数据挖掘中的应用研究

张 霞 著

中国物资出版社

图书在版编目（CIP）数据

粒度计算在数据挖掘中的应用研究/张霞著．—北京：中国物资出版社，2011.7

ISBN 978-7-5047-3646-8

Ⅰ.①粒… Ⅱ.①张… Ⅲ.①数据采集—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2011）第 091396 号

策划编辑	王玉霞	**责任印制**	方朋远
责任编辑	王玉霞	**责任校对**	孙会香　杨小静

出版发行	中国物资出版社		
社　　址	北京市丰台区南四环西路 188 号 5 区 20 楼	**邮政编码**	100070
电　　话	010-52227568（发行部）	010-52227588 转 307（总编室）	
	010-68589540（读者服务部）	010-52227588 转 305（质检部）	
网　　址	http://www.clph.cn		
经　　销	新华书店		
印　　刷	三河市西华印务有限公司		
书　　号	ISBN 978-7-5047-3646-8/TP·0072		
开　　本	710mm×1000mm　1/16	**版　　次**	2011 年 7 月第 1 版
印　　张	10	**印　　次**	2011 年 7 月第 1 次印刷
字　　数	170 千字	**定　　价**	25.00 元

前　言

知识发现和数据挖掘经常被人们用来提取感兴趣的信息和模式，当面临海量的、高维的、动态的、模糊的或不完备的实际应用数据时，采用传统的数据挖掘方法得到的挖掘结果常常不理想，因此研究新的数据挖掘方法是具有重要意义的。

粒度计算是信息处理的一种新的概念和计算范式，覆盖了所有和粒度相关的理论、方法、技术和工具的信息，主要用于描述和处理不确定的、模糊的、不完整的和海量的信息以及提供一种基于粒和粒间关系的问题求解方法。作为一个新兴领域的研究，国内外相关研究人员提供了一些粒度计算的理论基础，并且为检验知识发现过程的有效性、解决实际问题提供了一条可行途径。本书主要针对粒度计算在数据挖掘中的应用进行了深入研究。

本书主要研究内容如下：

传统文本聚类的研究大部分都是基于硬聚类的，即一个文本只能分到一个类中。然而随着信息的膨胀及学科之间的交叉渗透，文本软聚类方法正逐渐成为文本挖掘中的一个研究热点。本书对模糊粒度计算在文本软聚类中的应用进行了深入的研究，提出了一种基于模糊粒度计算的聚类方法，并且利用该聚类方法对 K-means 算法进行了优化。基于模糊粒度计算的聚类是通过归一化的距离函数将聚类问题映射到距离空间，调节粒度产生对文本集合 D 的动态聚类划分。动态聚类既可以作为一个单独的聚类结果，也可以作为其他算法的一个预处理步骤。K-means 算法是一种经典的聚类算法，速度快、消耗资源小，但是算法对初始聚类中心点敏感，容易陷入局部最小值。本书把基于模糊粒度计算的聚类方法作为 K-means 算法的预处理步骤，结果证明，这种预处理有效地消除了 K-means 算法的初始值敏感问题，优化了 K-means 算法。

文本分类是文本挖掘的一个重要内容，其中分类器的设计最为关键。

相对于KNN、SVM和Bayes等基于统计的分类器，当语料训练集不太充分或者类间特征词交叉比较多时，基于规则的分类器更有效。本书提出一种基于粒网络生成分类规则的文本分类方法，将形式概念分析和粒运算结合，利用逻辑决策语言描述概念的两个方面：内涵和外延。概念的内涵用一个语言公式来表示；概念的外延表示为符合这个公式的一个对象集合，即一个粒。通过粒的相关运算，从最大粒提取较小粒，直到产生属于同一个预定义类的粒，即最小粒。如此实现一个自顶向下的粒搜索过程，从而建立粒网络、得到分类规则并且实现文本的分类。

文本情感分类是目前信息处理的一个热点，目前国内情感分类的研究大部分是侧重于分词和特征的选取对情感倾向的影响，对分类算法的研究也大多是基于统计的分类方法。本书利用粒网络生成规则的方法，对文本情感分类进行初步研究，并通过实验证明基于粒网络生成分类规则的文本分类方法的有效性。

实际生活中，不完备信息系统是大量存在的，信息系统中遗漏值的存在大大增加了信息系统的不确定性，无法产生更多潜在的有价值规则，因此，不完备信息系统的数据挖掘研究具有现实的意义。本书提出一种基于信息粒度的遗漏值补齐方法GRCC。首先，根据定义的信息粒度选择信息粒度最大的列作为第一层粒，将第一层粒中含有遗漏项样本的相容类的属性值作为遗漏项补齐的选择项；其次，根据MDL准则，选择这些选择项中使MDL最小的属性值填充遗漏值；最后，将补齐的第一层粒作为完备信息，计算信息粒度，再选择信息粒度最大的列作为第二层粒，再补齐，依次完成整个不完备信息表的补齐。经过实验，GRCC算法补齐的信息表比其他补齐方法补齐的信息表产生更多高可信度和高支持度的规则，降低了信息表的不确定性。

作　者

2011年5月

目 录

1 绪 论

1.1 课题背景和意义

在 Internet 浪潮的冲击下，人们面临着数据爆炸的挑战，如何从浩如烟海的数据中找到内在规律，挖掘这些激增数据背后隐藏的重要信息已成为当前高科技领域研究的一项重要内容。当面临海量的、高维的、动态的、模糊的或不完备的实际应用数据时，传统的数据挖掘方法可能从数据中得不到理想的知识，需要研究新的数据挖掘方法。

粒度计算是信息处理的一种新概念和计算范式，覆盖了所有和粒度相关的理论、方法、技术和工具的信息，主要用于描述和处理不确定的、模糊的、不完整的和海量的信息以及提供一种基于粒和粒间关系的问题求解方法。粒度计算的基本思想是在问题求解中粒的使用，它通过粒对现实问题的抽象、粒之间的关系、粒的分解和合成以及粒或者粒度空间之间的转换来描述和解决问题。作为一个新兴领域的研究，国内外研究人员提供了一些粒度计算的理论基础，并且为检验知识发现过程的有效性、解决实际问题提供了一种可行的方法。

从人工智能的角度看，粒度计算的研究对复杂人工系统的设计和实现有着深远的影响。这是因为人工智能自 1956 年问世以来，尽管在知识表示、推理方法、机器学习和专家系统等领域取得了一些进展，但是在面向复杂智能系统的设计以及模糊的、不精确的、不完整的和海量信息的处理上，人工智能的传统理论和方法遇到了前所未有的困难。Pawlak 认为人的智能表现为对事物（事件、行为、感知等）的分类（Classification）能力。张钹等的看法是，他们认为“人类智能的公认特点，就是人们能从极不相同的粒度（Granularity）上观察和分析同一问题。人们不仅能在不同粒度的世界上进行问题求解，而且能够很快地从一个粒度世界跳到另一个

粒度世界，往返自如，毫无困难。这种处理不同世界的能力，正是人类问题求解的强有力的表现”。如何根据问题求解的需要选择合适的粒度，不同粒度世界之间又如何转换等问题对机器智能的开发意义重大。

从优化论的角度来看，随着信息科学的迅猛发展，现代系统越来越复杂，知识的表示越来越多样，并且信息系统中常常包含着海量、不完整、模糊及不精确的数据及对象，常常使得传统的数学优化方法显得无能为力。探寻能够有效表达和处理模糊的、不完整的、不精确的以及海量的信息，且具有智能特征的优化方法成为优化理论的发展趋势。粒度计算的理论与方法在观念上突破了传统优化思想的束缚，不再以数学上的精确解为目标，不要求目标函数和约束函数的连续性与凸性，并且对计算中数据的不确定性也有很强的适应能力，计算速度也快，这些优点使粒度计算具有更广泛的应用前景。所以，粒度计算理论的研究对推动优化领域的发展极其重要。

从问题求解的角度看，随着问题越来越复杂，所需的问题求解系统也越来越复杂。对于非常复杂的问题而言，将问题划分为一系列更容易管理和更小的子任务，可以降低全局计算代价。在这里，粒度计算能够作为处理上述任务的有效工具。

总的来说，粒度计算的研究目的有两点：一是为了寻找对问题的一种较好的近似解决方案而非最佳的精确方案，得到对问题的简化，实现鲁棒性，降低求解费用及实现对现实问题更好的求解；二是通过合适粒的选择，和粒之间、粒度层次之间以及粒度空间之间的交互使得原本无法解决或者解决困难的问题得到解决。信息科学的发展需求是研究粒度计算的最根本动机。

1.2 课题研究内容

（1）聚类

聚类分析是数据挖掘中一个重要研究领域，它既可以作为一个单独的工具，也可以作为其他数据挖掘分析算法的一个预处理步骤。快速和高质量的聚类技术可以将大量信息组织成少数有意义的簇，从而实现有效的信息检索。

从信息粒度的角度看，会发现聚类操作是在一个统一的粒度下进行计算的，《聚类/分类中的粒度原理》中提出了聚类中的粒度理论，如果改变聚类的相似度函数，则聚类的处理对象将变换到另一粒度空间中，这样就可以先根据需要改变当前聚类所在的粒度空间，然后通过粒度粗细的变化来改进聚类的结果。由于粒度的变换可以将复杂的问题简单化，即可将对象简化成若干个保留重要特征和性能的点，以便于分析，因此聚类可以进行粒度空间的变换粒度（Granu Larity）。

（2）分类

分类是数据挖掘中的一个重要目标和任务。其目的是构造一个分类模型（称做分类器），该模型能把数据库中的数据项映射到给定类别中。

在各种分类方法中，决策树算法和粒度思想具有更直观的联系。在由训练样本所生成的决策树中，一致分类问题中每个叶节点具有唯一类属，称为相容信息粒，而每个非叶节点具有多个类属，称为不相容信息粒。其中，处于同一层次的节点在同一粒度层次，节点越靠近根节点，则其粒度层次越“粗”，反之越“细”。

（3）不完备信息处理

波兰学者 Pawlak 在 20 世纪 80 年代提出了粗糙集理论。粗糙集理论的研究，已经经历了二十多年的时间，取得了很多成果，也建立了一套较为完善的粗糙集理论体系。当前，粗糙集理论已是处理模糊、不精确和不完备问题的重要数学工具。它在机器学习、知识获取、决策分类、数据库的知识发现等众多应用领域取得了不少成果，也成为粒计算研究的主要方向之一。

1.3 主要创新点

本书由上述三项研究内容入手，分别对粒度计算在聚类、分类、不完备信息系统数据处理三个方面进行了研究：

（1）基于模糊粒度计算的文本聚类

传统的文本聚类研究大部分都是基于硬聚类的，即一个文本只能分到一个类中。然而随着信息的膨胀及学科之间的交叉渗透，文本日益呈现出多样性和大量性，一个给定的文本往往可能属于多个类，所以我们需要一

种更客观的文本分类描述方法，由此基于模糊聚类技术的文本软聚类方法正逐渐成为文本挖掘中一个研究的热点。本章内容对模糊粒度计算在文本聚类中的应用进行了深刻的研究。根据模糊粒度计算的理论基础，我们得出可以通过归一化的距离函数 $d(d_i, d_j)$，它唯一确定一个模糊等价关系 $\tilde{R}$，进而通过不同的粒度 d_λ 产生对文本集合 D 的动态聚类划分。

K-means 聚类是一个被广泛应用的算法，具有计算速度快、消耗资源小的优点；缺点在于算法对初始聚类中心点敏感，聚类的结果很可能陷入局部极值，无法找到全局最优解。针对这个问题，我们通过模糊粒度计算得到一组初始聚类中心。通过实验比较证明，基于模糊粒度计算优化初始聚类中心的方法有效地消除了 K-means 算法对初始值的敏感，并且在准确率和迭代时间上都远远优越于传统的 K-means 算法。

（2）基于粒网络生成规则的文本分类

基于统计的文本分类器虽然具有操作简单、高精度的优点，但在语料不充足、特征词交叉比较多时，精度会大大下降。

本书在决策树分类 ID3 算法的基础上进行扩充首先研究了粒度计算在文本分类中的应用，将形式概念分析和粒运算结合，提出一种通过建立粒网络生成规则的方法进行分类。粒网络的建立是一个自顶向下的粒搜索过程，从最大的粒提取较小粒，直到产生属于同一个预定义类的粒，即最小粒。最小粒就是论域中可定义的合取粒。最小粒的族就形成了一个由可定义合取粒组成的覆盖。粒网络的建立得到分类规则，从而实现文本分类。

本书还对文本情感分类建模进行了研究。国内的文本情感分类还处于刚刚起步的阶段，目前国内研究文本情感的研究侧重在分词和特征选取对情感倾向上的影响，对分类算法的研究特别是基于规则分类的方法还很少，本书利用粒网络生成规则的方法，对文本情感文类进行了初步的研究。通过实验的验证，算法是有效的。

（3）基于信息粒度的不完备信息处理

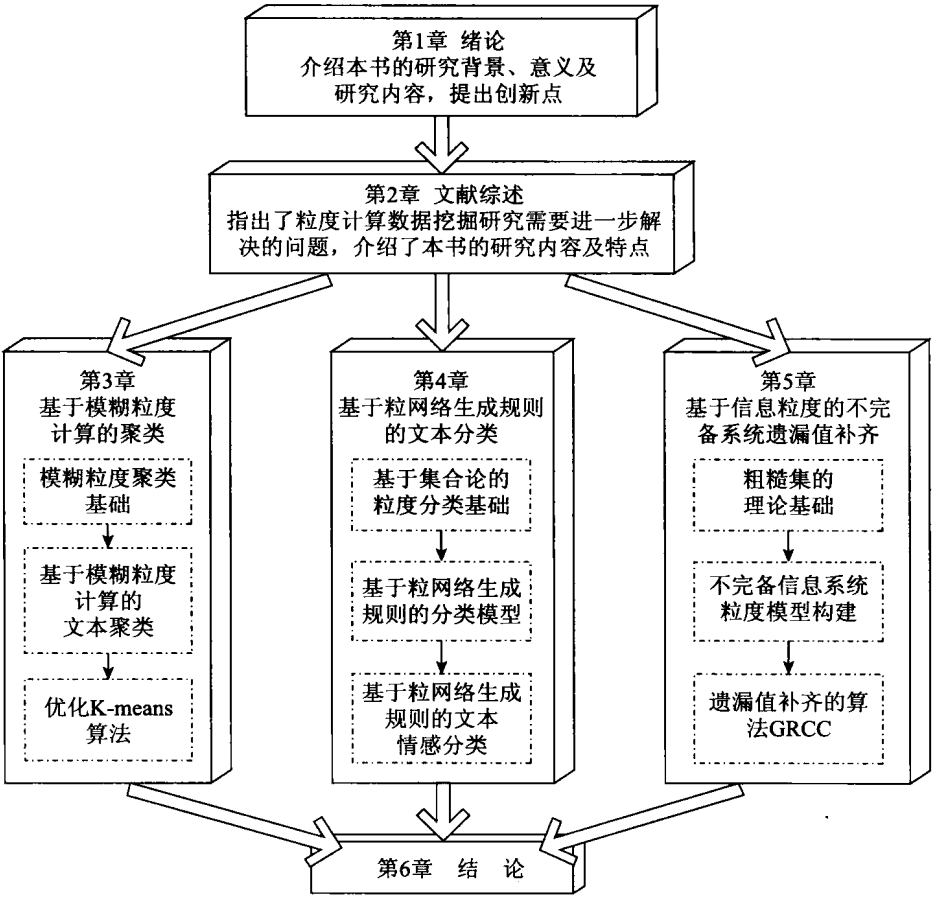
实际生活中，不完备信息系统是大量存在的，如医疗信息系统，不完备信息系统的数据挖掘研究具有现实的意义。遗漏值的存在大大增加了信息表的不确定性，信息表无法产生更多潜在的有价值规则。

一种做法是先将不完备信息表补齐再提取规则。本书提出一种基于信息粒度的遗漏值补齐方法 GRCC，信息粒度越高，不确定性越低。首先，

根据定义的信息粒度选择信息粒度最大的列；其次，由相容类产生遗漏值的属性值范围；最后，利用 MDL 准则确定遗漏项的属性值，如此逐列进行填充直到完成全部信息表的补齐。经过实验，GRCC 算法补齐的信息表比其他补齐方法产生的信息表产生更多高可信度和高支持度的规则，降低了信息表的不确定性。

1.4 本书的逻辑结构

本书的逻辑结构如下图所示。



本书的逻辑结构图

第 1 章：绪论。介绍了本书的研究背景、意义及与相关领域的研究现状。首先，从人工智能、优化论和问题求解的角度介绍了粒度计算对于数据挖掘研究的必要性；其次，介绍了粒度计算在数据挖掘的主要研究领域及现状；最后，介绍了本书的主要创新点和章节安排。

第 2 章：文献综述。首先，介绍数据挖掘的基本任务和现有的主要方法；其次，介绍了粒度计算的经典模型和在数据挖掘中的应用现状，指出了粒度计算数据挖掘研究需要进一步解决的问题；最后，介绍了本书的研究内容和特点。

第 3 章：基于模糊粒度计算的聚类。首先推导了通过归一化的距离函数将聚类问题映射到距离空间的模糊粒度理论基础，在此基础上提出基于模糊粒度的文本聚类算法，通过调节粒度实现对文本集合 D 的动态聚类划分。最后与 K-means 算法结合，将模糊粒度计算的聚类结果作为 K-means 算法的优化初始值，有效地消除了 K-means 算法的初始值敏感问题。

第 4 章：基于粒网络生成规则的文本分类。首先，介绍了集合论粒度计算理论基础；其次，定义了几种粒的运算并应用于分类问题，提出一种通过建立粒网络生成规则的方法进行文本分类；最后，对文本情感分类进行了初步的研究，提出了基于粒度计算的文本情感分类模型 GRCSC。

第 5 章：基于信息粒度的不完备系统遗漏值补齐。首先，介绍了粗糙集的理论基础、不完备信息系统的特点和已有的几种空值补齐的方法；其次，定义了信息粒度，从降低信息系统不确定性的角度构建粒度模型，并基于此模型提出遗漏值补齐的算法 GRCC，并对实验结果作出分析。

第 6 章：结论。对本书所做的研究工作进行了总结，指出了本书所需进一步研究的问题和内容。

2 文献综述

2.1 数据挖掘概述

数据挖掘是20世纪90年代中期兴起的一项新技术，是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的新技术。它能开采出潜在的模式，找出最有价值的信息，指导商业行为或辅助科学研究。还有很多和数据挖掘这一术语相近的术语，如从数据库中发现知识、数据分析、数据融合等。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、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。已有的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。

因此，数据挖掘是一门广义的交叉学科，涉及机器学习、模式识别、统计学、数据库、知识获取、知识表达、专家系统、神经网络、模糊数学、遗传算法等多个领域，内涵和研究热点非常广泛。所谓数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、以前未知的、具有潜在的或者现实应用价值的信息和知识的过程。

2.1.1 数据挖掘的步骤

一般地，数据挖掘与知识发现过程包括以下五步，如图2-1所示。

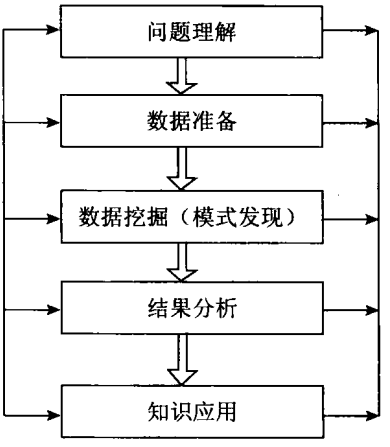


图 2-1　数据挖掘的基本过程

（1）问题理解：在开始数据挖掘之前，最基础的就是理解数据和明确目标；

（2）数据准备：获取原始数据，并从中抽取一定数量的子集，建立数据库，并对数据进行初步整理，清洗不完全的数据，做初步的描述分析；

（3）数据挖掘：针对特定的任务，设计或选择有效的数据挖掘算法并加以实现；

（4）结果分析：对数据挖掘的结果进行解释和评价，转换成为能够最终被用户理解的知识；

（5）知识应用：将分析所得到的知识集成到业务信息系统的组织结构中去。

需要指出的是，以上步骤不是一次完成的，可能其中某些步骤或者全部步骤要反复进行。数据挖掘是一个反复的人机交互过程，是由数据挖掘者根据待发现的任务和自身的背景知识，对数据进行研究、选择、设计合适的挖掘方法，并根据结果对数据进行修改、调整挖掘算法的循环的人机交互过程。所有发现的知识都是相对的，是有特定前提和约束条件，面向特定领域的。

2.1.2　数据挖掘的基本任务与方法

根据挖掘任务可以分为：分类或预测模型发现、数据总结与聚类发

现、关联规则发现、序列模式发现、相似模式发现、混沌模式发现、依赖关系或依赖模型发现、异常和趋势发现等。

根据挖掘对象可以分为：关系数据库、面向对象数据库（Object-Oriented Database）、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库、遗产数据库以及 Web 等对象的挖掘。

根据挖掘方法可以分为：机器学习方法、统计方法、聚类分析方法、探索性分析、神经网络（Neural Network）方法、遗传算法（Genetic Algorithm）、数据库方法、近似推理和不确定性推理方法、基于证据理论和元模式的方法、现代数学分析方法、粗糙集（Rough Set）方法、集成方法等。

2.1.3 数据挖掘常用的知识表示模式与方法

根据数据挖掘所能发现的知识可以分为：广义知识挖掘、关联规则挖掘、类知识挖掘、预测型知识挖掘、特异型知识挖掘等。

1. 广义知识挖掘

广义知识（Generalization）是指描述类别特征的概括性知识。我们知道，在源数据（如数据库）中存放的一般是细节性数据，而人们有时希望能从较高层次的视图上处理或观察这些数据，通过数据进行不同层次上的泛化来寻找数据所蕴涵的概念或逻辑，以适应数据分析的要求。数据挖掘的目的之一就是根据这些数据的微观特性发现有普遍性的、更高层次概念的中观和宏观的知识。因此，这类数据挖掘系统是对数据所蕴涵的概念特征信息、汇总信息和比较信息等的概括、精练和抽象的过程。被挖掘出的广义知识可以结合可视化技术以直观的图表（如饼图、柱状图、曲线图、立方体等）形式展示给用户，也可以作为其他应用（如分类、预测）的基础知识。

（1）概念描述（Concept Description）方法

概念描述本质上就是对某类对象的内涵特征进行概括。概念描述分为特征性（Characterization）描述和区别性（Discrimination）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。

概念描述是广义知识挖掘的重要方法，目前已经得到广泛研究。归纳起来有下面一些有代表性的方法：

①概念归纳（Concept Induction）方法。这种方法来源于机器学习。我们知道，典型的示例学习把样本分成正样本和负样本，学习的结果就是形成覆盖所有正样本但不覆盖任何负样本的概念描述。关于这类学习算法可以在经典的机器学习的教程中找到，这里不再赘述。但是，要把这种思想应用到数据挖掘中要解决两个关键问题。第一，必须扩大样本集的容量和范围。传统的机器学习希望是精练的小样本集，而数据挖掘系统必须忠实于源数据，是面向大容量数据库等存储数据集的。所以，扩大后的样本集可能难于有效地精确实现“覆盖所有正样本但不覆盖任何负样本”的概念归纳目标。要结合概率统计方法，在检验部分正样本或负样本情况下得到概念的描述。因此，最大限度地使用样本进行归纳就是必须解决的关键问题之一。第二，对于数据挖掘系统来说，正样本来自源数据库，而负样本是不可能在源数据库中直接存储的，但是缺乏对比类信息的概念归纳是不可靠的。因此，从源数据库中形成负样本（或区别性信息）以及相关的评价区别的度量方法等也是要解决的另一个重要问题。

②多维数据分析可以看做一种广义知识挖掘的有效方法。数据分析的经常性工作是数据的聚集，诸如计数、求和、平均、最大值等。既然很多聚集函数需经常重复计算，而且这类操作的计算量一般又特别大，因此一种很自然的想法是，把这些汇总的操作结果预先计算并存储起来，以便于高级数据分析使用。最流行的存储汇集数据类的方法是多维数据库（Multi-dimension Database）技术。多维数据库总是提供不同抽象层次上的数据视图。例如，可以存放每周的数据，也可在月底形成月数据，月数据又能形成年数据。关于多维数据模型的操作，已经被很好研究，许多文献可能和数据仓库、OLAP 联系起来。其实，这种模型，特别是它操作的完备性（如上钻、下钻等），可以成为广义知识发现的基础。

③面向数据库的概化方法。数据库，特别是关系型数据库是数据挖掘的主要源数据类型。近年来，在面向数据库的广义知识挖掘方面进行了有针对性的研究。值得一提的是，加拿大 SimonFraser 大学提出的面向属性的概念归约方法。它直接对用户感兴趣的数据视图（用一般的 SQL 查询语言即可获得）进行泛化，而不是像多维数据分析方法那样预先就存储好了泛化数据。原始关系经过泛化操作后得到的是一个泛化关系，它从较高的层次上总结了在低层次上的原始关系。有了泛化关系后，就可以对它进