第30卷第3期 地球科学与环境学报 Vol. 30No. 3
2008年9月 Journal of Earth Sciences and Environment Sept. 2008

空间数据挖掘的方法进展及其问题分析

胡圣武,李鲲鹏
(河南理工大学 测绘与国土信息工程学院,河南 焦作 454000)

摘要:采用归纳和总结的方法,研究了每种空间数据挖掘方法的特点和使用范围,指出了目前空间数据挖掘方法的局限性,就目前空间数据挖掘存在的问题进行了深入研究。认为空间数据挖掘是一个新兴的而富有前景的研究领域,目前只是取得了一定的初步成果,仍有大量的理论与方法需要深入研究。最后就空间数据挖掘的发展方向进行了研究和归纳。

关键词:空间数据;挖掘;方法;综述
中图分类号:P931.2 文献标志码:A 文章编号:1672-6561(2008)03-0311-08

Development and Problems in Spatial Data Mining Method

HU Sheng-wu, LI Kun-peng
(*School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo* 454000, *Henan, China*)

Abstract: By adopting the methods of inducing and summarizing, this paper studies the characteristic and by using range of every kind method for spatial data mining, points out the limitation and faults at present, probes into the problems of spatial data mining now. It is thought that the spatial data mining is a very young and good foreground research field at present although some elementary achievements have been obtained and there are many problems in theory and method that needs to be studied. The development of spatial data mining in the future is given.

Key words: spatial data; mining; method; review

0 引言

近年来,由于空间信息技术领域内对地观测技术特别是遥感技术和测绘技术、数据库技术、网络技术的快速发展以及观测台站建设的普及和不断完善,包括资源、环境、灾害等在内的各种空间数据指数级数增长,获得的空间数据越来越多。

由于技术和方法多种因素的影响,出现所谓"数据丰富,知识贫乏"的现象,即人类拥有大量的空间数据,但感觉到空间数据的缺乏。如何从大量的空间数据中发现人类所要的知识,就显得非常重要,这就是空间数据挖掘,是目前研究的一个热点和难点。

1 空间数据挖掘定义

空间数据挖掘与一般数据挖掘既有联系又有区别。对于数据挖掘(Data Mining,简称 DM)与从数据库中发现知识(Discovery for Spatial Database,简称 KDD)这两个概念,经常让人混淆。有些学者将 DM 作为 KDD 的一个核心环节,认为 KDD 过程除了包括数据挖掘外还包括数据准备和发现结果解释评估等诸多环节。有的则认为二者本质相同,DM 只是经常用于统计、数据分析和信息系统等工程领域,而 KDD 多用于人工智能和机器学习等领域。还有人认为二者难以分离,应作为一个整体使用,即数据挖掘和知识发现(DMKD)才较为

收稿日期:2007-10-20
基金项目:国家自然科学基金项目(40474003)
作者简介:胡圣武(1970-),男,湖南津市人,副教授,工学博士,从事GIS基础理论和图像处理研究。E-mail: hushengwuzhu@163.com

适宜。但术语“数据挖掘”比“在数据库中发现知识”和“数据挖掘和知识发现”形式更简洁，因此广为流行。

空间数据挖掘(Spatial Data Mining，简称SDM)或称从空间数据库中进行知识发现(Knowledge Discovery from Spatial Database)，定义为：在空间数据库和数据仓库的基础上，综合利用统计学方法、模式识别技术、人工智能方法、神经网络技术、粗集、模糊数学、机器学习、专家系统、可视化技术和其他相关的信息技术作为手段，从大量的空间数据、管理数据、经营数据或遥感数据中析取出可信的、新颖的、感兴趣的、隐藏的、事先未知的、潜在有用的和最终可理解的知识，从而揭示出蕴含在空间数据背后客观世界的本质规律、内在联系和发展趋势，实现知识的自动或半自动获取，为管理和经营决策提供依据[1]。

2 空间数据挖掘研究现状

自 1989 年第十一届人工智能国际联合会议首次提出数据挖掘的概念，国外各类组织在数据挖掘领域展开了大量的研究工作。迄今为止，对关系数据库和事务数据库中数据挖掘的研究已经取得了不少进展[2]，代表性的工作有：用面向属性的归纳方法在关系数据库中发现特征规则和区分规则[3]，在事务数据库中发现关联规则[4]，从大型数据库中发现多层次关联规则[5]，基于多层关联进行分类[6]，基于距离的和基于密度的聚类分析的优化[7]等。许多著名的数据库和数据仓库供应商、统计分析软件开发商、相关的人员和研究所等纷纷投入研究和开发力量，相继开发出一些数据挖掘商用系统和原型系统[8]。

空间数据挖掘的研究比起一般的数据挖掘要晚，但近几年已引起广泛的兴趣，加拿大西蒙弗雷泽大学、德国慕尼黑大学以及美国、澳大利亚等国家的许多大学和研究所，都有空间数据挖掘的成果报道[9-10]。这些研究者大多具有计算机科学背景，他们一般把空间数据挖掘作为数据挖掘的一个应用领域，研究的重点是提高原有数据挖掘算法在空间数据上的执行效率[11]，Lu W 等提出了面向属性归纳的基于概化的空间数据挖掘方法[12]，Ng R T 等提出一种基于聚类结果的描述性空间分析方法[13]，Koperski K 等研究了有关空间数据立方体的设计和构造问题[14]。Koperski K 等提出了一种逐步求精的空间关联规则和挖掘方法[15]，Knorr 和 Xu 以及 Ester、Frommelt 给出了空间分析和趋势分析的方法，Koperski、Han 和 Stefanovic 提出了一种空间数据分类的方法。

与国外相比，国内对数据挖掘的研究稍晚，还没有形成整体的力量。1993 年国家自然科学基金首次支持该领域的研究项目。目前，国内许多科研单位和高等院校竞相开展数据挖掘理论及其应用研究[16-19]，如清华大学、武汉大学、南京大学、北京工业大学、中科院计算机技术研究所等，在金融、电信、电力、环境监测、网络、文本处理等领域作了积极研究和探讨。在空间数据挖掘领域，武汉大学李德仁在 1994 年就提出了从 GIS 数据库发现知识的建议，而后与李德毅、邸凯昌、王新洲、王树良等开展了空间数据挖掘的理论、方法与应用研究[20]。单春、张清浦等提出了建立地理因子库进行空间数据挖掘的构想。周成虎和张健挺提出了基于信息熵的地学空间数据挖掘模型[21]。

总的来说，数据挖掘当前相当于数据库技术在 20 世纪 70 年代所处的地位，尤其空间数据挖掘现在基本处于起步阶段，尽管国内外提出了许多理论和方法，但大多是其数据挖掘领域的扩展版本，理论是否可行，方法是否有效，需要人们用实践去证明，遗憾的是，至今为止，空间数据挖掘领域尚无一成形的实用系统。

3 空间数据挖掘方法

空间数据挖掘研究之初，主要是在空间邻接关系的基础上进行空间数据挖掘，认识空间关系是空间实体之间由于空间位置和形状不同而造成的相互之间的各种联系。Easter M 等将空间关系作为一般属性值[22]，存储于关系数据库，充分利用关系数据库管理系统的高效存储及访问机制，并引入邻接图、邻接路径的几个基本操作。在包含邻接关系的数据表基础上，进行空间关联规则发现、空间特征描述、空间聚类分析、空间分类分析、空间趋势检测等典型的空间数据挖掘。

基于邻接关系的空间数据挖掘方法需要计算对象之间的邻接关系，并将它们存储于表中，对于一些基本的挖掘任务表现较高的挖掘效率，作为空间数据挖掘的一种基本策略。目前，空间数据挖掘研究更多的是集中于从空间数据的复杂性特点出发，提出了相应的解决办法。

3.1 处理海量数据的挖掘方法

对于海量数据, 解决算法效率主要方法有: 一是改进原有算法的结构以降低运算的复杂度。二是采用新的运算策略或改变算法运行环境。

3.2 解决空间非线性关系的挖掘方法

在解决空间非线性关系问题上, 主要采用以神经网络为代表的智能计算。神经网络作为模拟复杂系统非线性关系的一种模型, 按照其内部神经元连接的拓扑结构、学习规则以及传递函数的类型等标准可以分为若干种类, 较常见的有前向网络(BP)、相互联接型网络(Hopfield)、自组织映射网络(SOM)和径向基函数网络(RBF)等。

由于神经网络非常适用于非线性复杂关系, 并且在处理复杂问题时不需了解网络内部所发生的结构变化, 因而被广泛应用于空间数据挖掘和知识发现, 并构造不同的网络模型分别实现了空间聚类、分类、关联、回归、模式识别等多种算法。

近年来, 神经网络学习算法中还引入了各种进化算法作为优化策略, 一种是模拟退火算法(Simulated Annealing, 简称 SA), 另一种为遗传算法(Genetic Algorithm, 简称 GA)。

3.3 应用支撑向量机处理高维空间数据挖掘

支撑向量机(Support Vector Machine, 简称 SVM)是统计学习理论中的通用学习算法, 由 Vapnik V 于 1995 年提出, 主要思想是在高维空间内利用线性函数的对偶核, 通过内积空间的向量运算来处理线性不可分数据[23]。支撑向量机的主要优点在于该模型利用优化对偶理论使高维特征空间中的模型参数估计易于计算, 并且运算的复杂度与问题的维数关系不大。支撑向量机模型在学习效率、解决过度拟合问题和全局最优化等方面都表现出优于神经网络的良好性质[24]。

在解决空间数据的分类、特征识别、图像压缩等方面取得了一定进展[25]。从支撑向量机产生的背景和应用效果来看, 它特别适合处理高维、复杂的目标识别问题, 将在遥感影像理解特别是对复杂地学信息的识别等方面表现出广阔的应用前景。

3.4 基于信息熵的空间数据挖掘方法

信息熵作为衡量信息量的指标之一, 也被引用到空间数据挖掘研究领域。信息熵之所以被广泛应用, 主要原因是信息熵采用简单的方式定义系统的复杂性, 并具有明确的物理含义, 即信息熵是在平均意义上表征信息源的总体特征。在空间数据挖掘的各方面研究中, 信息熵的作用在机器学习领域中得到充分展示。目前机器学习没有一个统一的定义, 一般被定义为它是系统可以自我改进、更新的过程。机器学习按照是否需要先验知识和学习样本可分为有监督机器学习和无监督机器学习。机器学习也被分为规则提取算法、决策树算法以及两种算法的综合。文献[21] 利用地学信息熵作为空间数据分割的标准, 对上述空间数据进行分割, 从而将空间属性与熵标准判定有机结合在一起。

3.5 应用尺度空间概念的空间数据挖掘方法

Witkin P[26] 和 Koenderink J J[27] 最先提出了尺度空间的概念, 并将之应用于图像结构表达。关于空间数据的尺度特征, 文献[28] 认为空间数据是包含了尺度维的四维状态空间, 尺度维反映的是空间数据由细到粗多比例尺或多分辨率的几何变换过程。要刻画空间数据的尺度性, 关键是建立一种空间数据分辨率由细到粗的序列。构造尺度空间的基本原理就是将空间数据集投影到不同分辨率的空间内, 并挖掘尺度空间下的知识。这一过程可以利用视觉想象进行类比, 在最高分辨率下, 空间中的每一点可视为一个小光点, 整个的空间数据就成为完整的一幅图像, 当逐渐远离这幅图像时, 小光点变得模糊, 进而融合为小光斑, 当图像进一步模糊时, 多个小光斑又融合为大光斑, 这一过程不断重复下去直至所有的数据点都融为一个大光斑, 就此完成了尺度空间的构造[29]。

建立尺度空间的方法多种多样, 基本的有小波滤波、高斯平滑、高斯导数滤波等。尺度空间概念的应用领域包括空间特征识别和空间尺度聚类等方面。在空间特征识别方面, Andrew D J[30] 等借助向量场算子理论对尺度空间内实物图像的边界和对称性进行了识别。在空间尺度方面, 张讲社等[29] 借鉴视觉的基本理论, 将尺度空间应用到聚类算法中, 并确定聚类结果的有效性, 从而减少在聚类过程中人为的干预。

3.6 基于模糊集和粗集理论的空间数据挖掘方法

对于空间关系的不确定性, 通常采用模糊集理论加以描述。模糊集理论的优势在于利用隶属函数刻画空间关系的不确定性, 用对象部分归属代替整个对象归属的概率。模糊集的思想已渗透到空间数据知识发现的各种方法中, 如模糊聚类与分类、模糊神经网络、模糊专家系统等。

隶属函数虽然对不确定关系进行了成功刻画,

打破了非此即彼的传统概念，但隶属度的确定仍然需要借助先验知识，必然导致结果的多解性。

Pawlak 提出的粗集理论利用了模糊概念的优点，克服了隶属函数的不足，成为研究模糊现象的又一有力工具。粗集方法不需要先验假设，而是利用集合论中的上近似和下近似来刻画集合，当个体 A 属于集合 X 的下近似时，A 肯定属于集合 X；而当 A 不属于集合 X 的上近似时，则 A 肯定不属于集合 X；如果 A 属于 X 的上近似而不属于 X 的下近似，则 A 有可能属于集合 X。

在目前的空间数据库应用中，概括数据的手段主要是执行 Zoom in、Zoom out 操作，但要实现真正意义上的数据概括，仍然比较困难，必须发展一种抽象和浓缩数据的算法，算法在执行过程中还须保证数据的质量。粗集理论凭借不需要定量化和不确定性优势实现这一点。例如，文献[31] 应用模糊集理论在模糊图像子集中抽取模糊边界图形对象的空间关系。Beaubouef T 等讨论了不确定边界问题的描述方法，并将粗糙集利用不可辨识性关系和邻近区域，实现空间关系规则等挖掘任务[32]。

除了利用模糊集和粗集处理空间数据的不确定以外，邸凯昌等[33] 提出云模型，该模型将模糊性与随机不确定性有机结合，从另一角度解决了模糊集理论中隶属函数的固有缺陷。

3.7 针对高维数据的挖掘算法

要解决高维数据的挖掘问题，必须先了解高维数据的性质。与低维空间性质对比，高维空间所表现出的性质与我们日常接触的三维空间性质完全不同[34]。对高维数据进行挖掘的思路一般有两种：一种降维的方法，是将高维数据通过线性变换投影到低维空间，然后再采用一般的挖掘算法；另一种就是采用适合处理高维数据的算法直接进行信息提取。

3.8 处理缺值数据的挖掘方法

空间数据缺值研究的过程主要可以分为两步：首先利用各种统计手段模拟出缺失值，然后再利用包含已知观测值和缺失值的全集进行统计分布函数参数估计。产生缺失值的方法包括均值转嫁、有补充的随机替代、无补充的随机替代、时间序列转换、完全均值转换等，这些方法将缺值的补充与分布函数的参数估计作为两个独立部分看待。而最大值期望算法则将二者有机联系起来，在处理非完整数据集时对其最大似然函数应用了迭代方法，从而在模拟缺失样本的同时估计出分布函数的未知参数。此外，Kumar J K[35] 还提出在空间信息不完整的情况下，采用模糊神经网络对空间分布进行预测。要进一步了解不同分布空间数据的缺值研究可参阅文献[36]。

3.9 图像分析和模式识别方法

空间数据库（数据仓库）中含有大量的图形图像数据，一些图像分析和模式识别方法可直接用于挖掘数据和发现知识，或作为其他挖掘方法的预处理方法。用于图像分析和模式识别的方法主要有决策树方法、图论方法、数学形态学方法、神经元网络、空间离群数据方法等。

3.10 计算几何分析方法

1975 年，Shamos 和 Hoey 利用计算机有效地计算平面点集 Voronoi 图，并发表了一篇著名论文，从此计算几何诞生了，现在 Voronoi 图是计算几何中一个被广泛研究的课题，并取得了辉煌的成果，使得计算几何成为理论计算机科学领域中一个新的极有生命力的领域，并且计算几何中的研究成果已在计算机图形学、统计分析、化学、模式识别、空间数据库以及其他许多领域得到了广泛应用。计算几何研究的典型问题包括几何基元、几何查找和几何优化等。

空间数据挖掘领域中的空间拓扑关系、数据的多尺度表达、空间同位、自动综合、空间聚类、空间目标的势力范围、公共设施的选址、最短路径等问题都可以利用 Voronoi 图进行解决。

3.11 统计分析方法

统计方法一直是分析空间数据的常用方法，有着较强的理论基础，拥有大量的算法，可有效地处理数字型数据。这类方法有时需要数据满足统计不相关假设，但很多情况下这种假设在空间数据库中难以满足。另外，统计方法难以处理字符型数据。应用统计方法需要有领域知识和统计知识，一般由具有统计经验的领域专家来完成。

以变差函数和 Kriging 方法为代表的地学统计方法是地学领域特有的统计分析方法，由于考虑了空间数据的相关性，地学统计在空间数据统计和预测方面比传统统计学方法更加合理有效，因而在空间数据挖掘中也可以充分发挥作用[37]。

3.12 空间分析方法

空间分析能力是 GIS 的关键技术，是 GIS 系统区别于一般数字制图系统的主要标志之一。目前

常用的GIS 系统的空间分析功能有综合属性数据分析、拓扑分析、缓冲区分析、密度分析、距离分析、叠置分析、网络分析、地形分析、趋势面分析、预测分析等, 应用这些方法可以交互式地发现目标在空间上的相连、相邻和共生等关联关系以及目标之间的最短路径、最优路径等辅助决策的知识。空间分析往往是应用领域知识产生新的空间数据, 所以常作为预处理和特征提取方法与其他数据发掘方法结合起来从空间数据库发现知识。

当然, 空间数据挖掘方法还有归纳学习方法、神经网络方法、决策树方法、证据理论等。上述每一种方法都有一定的适用范围。在实际应用中, 为了发现某类知识, 常常要综合运用这些方法。空间数据挖掘方法还要与常规的数据库技术充分结合。例如, 在时空数据库中挖掘空间演变规则时, 可利用GIS 的叠置分析等方法首先提取出变化了的数据, 再综合统计方法和归纳方法得到空间演变规则。总之, 空间数据挖掘利用的技术越多, 得出的结果精确性就越高, 因此, 多种方法的集成也是空间数据挖掘的一个有前途的发展方向。此外, 空间数据挖掘除了发展和完善自己的理论和方法, 还要充分借鉴和吸取数据挖掘和知识发现、数据库、机器学习、人工智能、数理统计、可视化、地理信息系统、遥感、图形图像学、医疗、分子生物学等学科领域成熟的理论和方法。

4　空间数据挖掘存在的问题

空间数据挖掘已经成为数据库和信息决策领域一个重要研究方向, 虽然取得一定进展, 但它仍然极具吸引力和挑战性, 还存在很多问题需要研究。

4.1　基于理论与挖掘算法研究

经过近年来的研究, 空间数据挖掘继承和发展相关基础学科(如机器学习、统计学等)的已有成果方面, 并探索出独具特色的理论体系。但是, 这决不意味着空间挖掘理论已经完善, 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。

4.2　应用研究

空间数据挖掘应用与实现也是目前研究的热点之一, 主要集中于多算法集成、挖掘系统中的人机交互技术和可视化技术、空间数据挖掘系统与地理信息系统、遥感解译专家系统、空间决策支持系统等的集成, 与特定应用目标的融合, 如客户关系模式、电子商务等。

4.3　目前空间数据挖掘还没有一个实用的系统

目前空间数据挖掘系统比较多, 但每一种系统都是针对一定的用途开发的, 还没有一个比较流行的、实用的系统存在。

4.4　在不确定性下进行数据挖掘

空间数据含有随机不确定性和模糊性[38-39], 但目前的空间数据挖掘方法对空间数据的不确定性处理还存在一些问题。有的方法就没有考虑空间数据的不确定性; 有的方法考虑了随机不确定性; 有的方法考虑空间数据的模糊性。还没有一种方法能较好地考虑空间数据随机不确定性又考虑空间数据模糊性。

4.5　空间数据挖掘质量评价

空间数据挖掘的知识很多, 但挖掘的程度和效益如何等这些问题目前还没有进行研究, 实际上是空间数据挖掘质量评价的问题, 只有解决了这个问题, 才能开发出更好的空间数据挖掘系统和探索出更加完善的空间数据挖掘方法和算法。

4.6　空间对象

空间对象包含空间属性和非空间属性, 尽管有的空间属性经过处理可以转化为一般的属性要素进行分析, 如距离、方向, 但是大部分的空间属性如空间关系并不适合于这种简化处理。如何有效地将空间分析融入数据挖掘过程, 是空间数据挖掘的最大难点, 也是空间数据挖掘区别于其他数据挖掘的最显著特点。

4.7　空间数据库类型的多样性问题

许多空间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、复杂的空间数据对象、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, 局域网和广域网上更是连接了许多空间数据源并形成了巨大的、分布式的、分层的和异构的空间数据库。从不同格式或非格式地具有不同数据语义的空间数据源而来的空间数据集, 对空间数据挖掘提出了新的挑战, 所以一个强有力的空间数据挖掘系统应该能够有效地处理这些复杂的数据类型, 然而, 空间数据种类繁多以及空间数据挖掘的不同目标, 使得采用一个空间数据挖掘系统处理所有类型的空间数据是

不可能的。

4.8 性能问题

空间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。空间数据挖掘算法的效率和可伸缩性是指为了有效地从空间数据库中的大量数据中抽取有用的知识,知识发现算法是有效的和可伸缩的。也就是说,一个空间数据挖掘算法在大型空间数据库中的运行时间必须是可预计的和接受的。许多现有的空间数据挖掘算法往往适合于常驻内存的、小数据集的空间数据挖掘,而大型空间数据库中存放了 TB 级的数据,所有的空间数据无法同时导入内存,所以从空间数据库的观点,有效性和可伸缩性是实现空间数据挖掘系统的关键问题。

4.9 空间数据不断改变的问题

许多实际空间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。为了随时获得一个与空间数据相关的有效模式,需要以一定的时间间隔不断重复同样的空间数据分析。

4.10 空间数据挖掘结果的有用、确定、可表示性

空间数据挖掘算法可能会发现数以千计的模式,其中有些模式是错误的,对于给定用户,许多模式未必是感兴趣的,因此,如何提供给用户有用、确定、可表示性的知识是一个需要研究的课题。

4.11 空间数据挖掘方法和用户交互问题

空间数据挖掘方法反映了所挖掘的知识类型、多粒度下挖掘知识的能力、领域知识的使用、特定的挖掘等。

由于不同的用户可能对不同类型的知识感兴趣,空间数据系统应该覆盖范围很广的数据分析和知识发现任务,在相同的空间数据上发现不同的知识,有必要提供交互式手段,开发不同的空间数据挖掘技术。

4.12 与空间数据库的无缝集成

空间关系查询语言允许用户提出特定的空间数据检索和查询要求,但并不适合于空间数据挖掘的数据查询,为此需要开发高级的空间数据挖掘查询语言,使用户通过分析任务的相关数据集、领域知识、所挖掘的数据类型、所发现的模式必须满足的条件和约束,来描述特定的空间数据挖掘任务。空间数据挖掘的查询语言与空间数据库和空间数据仓库查询语言的集成和无缝链接,可以提高空间数据挖掘的有效性和灵活性。

4.13 不同技术的集成

大多数空间数据挖掘系统采用一种技术或有限的几种技术执行空间数据分析任务,由于众所周知的原因,空间数据挖掘中没有所谓最好的技术。问题的不同,空间数据挖掘目标的不同,将导致不同的空间数据挖掘方法和技术。所以,开发和研究基于多种不同技术集成的空间数据挖掘系统是未来的研究方向。

4.14 私有性与空间数据挖掘问题

知识发现可能导致对于私有权的入侵,研究采取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空间数据时,数据安全性将受到严重威胁。这时空间数据保护和空间数据挖掘可能会造成一些矛盾的结果。

4.15 空间数据挖掘的智能化

目前空间数据挖掘已经用到了人工神经网络等智能算法,但现有的空间数据挖掘系统的智能化程度比较低,还需要进一步提高。

4.16 面向对象的空间数据库的知识挖掘

目前在实际中应用的空间数据挖掘方法都假定空间数据库中采用的是扩展的关系模型,而关系型数据库并不能很好地处理空间数据,面向对象(Object Oriented,简称 OO)模型比传统的关系模型或扩展关系模型更适合处理空间数据。因此,在空间数据挖掘中开发 OO 技术是一个具有极大潜力的领域。

空间数据挖掘存在的问题还很多,如空间数据格式的转换问题、海量空间数据更新的问题等,在此就不一一列出,可参阅文献[40~46]。

5 结语

空间数据挖掘是一个新兴的而富有前景的研究领域,目前只是初步取得了一定的成果,仍有大量的理论与方法需要深入研究。其中主要包括多源空间数据的清理、基于空间不确定性(位置、属性、时间等)的数据挖掘、递增式数据挖掘、栅格矢量一体化数据挖掘、多分辨率及多层次数据挖掘、并行数据挖掘、新算法和高效算法的研究、空间数据挖掘查询语言、遥感图像数据库的数据挖掘、多媒体空间数据库的知识发现、网络空间数据的挖掘等方向。在开发实现空间数据挖掘的计算机软件系统时,还要研究多源空间数据的集成、多算法的

集成、存储空间和计算效率的降低、人机交互技术、可视化技术、空间数据挖掘系统与 GIS、空间数据仓库、空间决策支持系统和遥感解译专家系统的集成等问题。

此外, 空间数据挖掘除了发展和完善自己的理论和方法, 也要充分借鉴和汲取数据挖掘和知识发现、数据库、机器学习、人工智能、数理统计、可视化、遥感、图形图像学、医疗、分子生物学等学科领域成熟的理论方法。

参考文献:

[1] 李德仁, 王树良, 史文中, 等. 论空间数据挖掘和知识发现[J]. 武汉大学学报: 信息科学版, 2001, 26(6): 491-499.

[2] Chen M S, Han J, Yu P S. Data Mining: An Overview from Database Perspective[J]. IEEE Transaction on Knowledge and Data Engineering, 1996, 26(3): 235-246.

[3] Han J, Cai Y, Cercone N. Knowledge Discovery in Databases: an Attribute Oriented Approach[C] //In Proc of the 18th Conf on Very Large Data Bases. Very Large Data Bases. Vancouver: the 18th Conf on Very Large Data Bases, 1992: 547-559.

[4] Cheung D W, Han J W, Ng V T, et al. Maintenance of Discovered Association Rules in Large Databases: An Dncremental Updating Technique[C] //In Proc of the 12th Int Conf on Data Engineering. Data Engineering. New Orleans: the Institute of Electrical and Electronics Engineers Inc, 1996: 106-114.

[5] Han J, Fu Y. Discovery of Multiple-level Association Rules from Large Databases[C] // In Proc of the 21st Int'l Conf on Very Large Databases. Very Large Databases. Zurich: the 21st Int'l Conf on Very Large Databases, 1995: 420-431.

[6] Li W, Han J W, Pei J. CMAR: Accurate and Efficient Classification Based on Multiple Class-association Rules[C] //In Proc of the IEEE Int Conf on Data Mining. Data Mining. San Jose: the IEEE Int Conf on Data Mining, 2001: 369.

[7] Ester M, Kriegel H P, Xu X. A Database Interface for Clustering in Large Spatial Databases[C] //In Proc 1st Int Conf on Knowledge Discovery and Data Mining. Knowledge Discovery and Data Mining. Montreal: the 1st Int Conf on Knowledge Discovery and Data Mining, 1995: 94-99.

[8] Feeldersa A, Daniels H, Holsheimer M. Methodological and Pratical Aspects of Data Mining[J]. Information and Management, 2000, 37: 271-281.

[9] Ester M, Formmelt A, Kriegel H P. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support [J]. Data Mining and Knowledge Discovery, 2000, 10(5): 123-129.

[10] Ester M, Gundlach S, Kriegel H P, et al. Database Primitives for Spatial Data Mining[C] //In Proc Int Conf on Databases in Office, Engineering and Science. Databases in Office, Engineering and Science. Freiburg: the Int Conf on Databases in Office, Engineering and Science, 1999: 137-150.

[11] Koperski K, Adhikary J, Han J W. Spatial Data Mining: Progress and Challenges Survey Paper[C] //In Proc ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Data Mining and Knowledge Discovery. Montreal: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996: 1-10.

[12] Lu W, Han J W, Ooi B C. Discovery of General Knowledge in Large Spatial Databases[C] //In Proc of Far East Workshop on Geographic Information Systems. Geographic Information Systems. Singapore: the Far East Workshop on Geographic Information Systems, 1993: 275-289.

[13] Ng R, Han J W. Efficient and Effective Clustering Methods for Spatial Data Mining[C] //In Proc of the 20th Int'l Conf on Very Large Data Bases. Very Large Data Bases. Santiago: the 20th Int'l Conf on Very Large Data Bases, 1994: 144-155.

[14] Koperski K, Han J W. Data Mining Methods for the Analysis of Large Geographic Databases[C] // In Proc of Conf on Geographic Information Systems. Geographic Information Systems. Vancouver: the Conf on Geographic Information Systems, 1996.

[15] Koperski K, Han J W. Discovery of Spatial Association Rules in Geographic Information Database[C] // Egenhofer M J, Herring J R. In Advances in Spatial Database. Berlin: Springer-Verlag, 1995: 47-66.

[16] 陈富赞, 寇继淞, 王以直. 数据挖掘方法的研究[J]. 系统工程与电子技术, 2000(22): 54-58.

[17] 吕安民, 林宗坚, 李成名. 数据挖掘和知识发现的技术方法[J]. 测绘科学, 2000, 25(4): 36-39.

[18] 田金兰, 张素琴, 黄　刚. 用关联规则方法挖掘保险业务数据中的投资风险规则[J]. 清华大学学报: 自然科学版, 2001, 41(3): 15-18.

[19] 杨　武, 陈　庄. 数据库知识发现技术及应用[J]. 重庆工学院学报, 2001, 15(5): 21-24.

[20] 邸凯昌. 空间数据挖掘与知识发现[M]. 武汉: 武汉大学出版社, 2000.

[21] 周成虎, 张健挺. 基于信息熵的地学空间数据挖掘模型[J]. 中国图象图形学报, 1994(4): 621-625.

[22] Ester M, Kriegel H P, Sander J. Spatial Data Mining: A Database Approach[C] //Scholl M, Voisard A. Advances in Spatial Databases. Berlin: Springer-Verlag, 1997: 47-66.

[23] Vapnik V N. The Nature of Statistical Learning Theory[M]. Berlin: Springer-Verlag, 1995.

[24] Cburges C J. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(1): 121-167.

[25] Hermes L, Frieau D, Puzicha J, et al. Support Vector Machines for Land Usage Classification in Landsat TM Imagery

[C] //In Proc of the IEEE 1999 Int Geoscience and Remote Sensing Symposium. International Geoscience and Remote Sensing Symposium. Hanburg: the IEEE 1999 International Geoscience and Remote Sensing Symposium, 1999: 348-350.

[26] Witkin A P. Scale-Space Filtering[C] // In Proc 8th Int Joint Conf Art Intell. Art Intell. Karlsruhe: the 8th Int Joint Conf Art Intell, 1983: 1019-1022.

[27] Koenderink J J. The Structure of Images [J]. Biological Cybernetics, 1984, 50(2): 363-370.

[28] 邸凯昌, 李德仁, 李德毅. 空间数据挖掘和知识发现的框架[J]. 武汉测绘科技大学学报, 1997, 22(4): 328-332.

[29] 张讲社, 徐宗本. 基于视觉系统的聚类: 原理与算法[J]. 工程数学学报, 2000, 17(增刊): 14-20.

[30] Andrew D J, Edwin R H. Kale Space Vector Fields for Symmetry Detection[J]. Image and Vision Computing, 1999, 17(2): 337-345.

[31] Keller J M, Sztandera L. Spatial Relations Among Fuzzy Subsets of an Image[C] //In 1st Int Symposium on Uncertainty Modeling and Analysis. Uncertainty Modeling and Analysis. Maryland: the 1st Int Symposium on Uncertainty Modeling and Analysis, 1990: 207-211.

[32] Beaubouef T, Petry F E. A Rough Set Foundation for Spatial Data Mining Involving Vagueregions[C] //In Proc of the 2002 IEEE Int Conf on Fuzzy Systems. Fuzzy Systems. Honolulu: the 2002 IEEE Int Conf on Fuzzy Systems, 2002: 767-772.

[33] 邸凯昌, 李德毅, 李德仁. 云理论及其在空间数据发掘和知识发现中的应用[J]. 中国图象图形学报, 1999, 4(11): 930-935.

[34] Jimenez L, Landgrebe D. Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate Data[J]. IEEE Transaction on Systems Man and Cybernetics, 1998, 28(1): 39-54.

[35] Kumar J K. An Application of Spatial Prediction Using a Fuzzy Neural Network[C] //Int Joint Conf on Neural Networks. Neural Networks. Washington DC: the Int Joint Conf on Neural Networks, 1999: 4241-4246.

[36] Barroso P L, Wilton O B, Martin K. Best Linear Unbiased Predictor Mixed Model with Incomplete Data[J]. Communications in Statistics: Theory and Methods, 1998, 27(1): 121-129.

[37] 邸凯昌. 空间数据挖掘和知识发现的理论与方法[D]. 武汉: 武汉测绘科技大学, 1999.

[38] 王新洲, 史文中, 王树良. 模糊空间信息处理[M]. 武汉: 武汉大学出版社, 2004.

[39] 胡圣武. GIS 质量评价与可靠性分析[M]. 北京: 测绘出版社, 2006.

[40] 滕明贵. 空间数据挖掘技术及其应用研究[D]. 合肥: 中国科学技术大学, 2004.

[41] 陈久军. 基于统计学习的图像语义挖掘研究[D]. 杭州: 浙江大学, 2006.

[42] 曾松峰. GIS 中空间数据挖掘初步研究[D]. 南京: 南京大学, 2002.

[43] 焦李成, 刘 芳, 刘 静, 等. 智能数据挖掘与知识发现[M]. 西安: 西安电子科技大学出版社, 2006.

[44] 蓝荣钦. 模糊空间数据挖掘方法及应用研究[D]. 北京: 北京科技大学, 2005.

[45] 胡彩平, 秦小麟. 空间数据挖掘研究综述[J]. 计算机科学, 2007, 34(5): 14-19.

[46] 张 楠, 曲海平, 刘 念, 等. 空间数据挖掘的研究进展[J]. 微处理机, 2007, 20(2): 1-4.

(上接第 304 页)

参考文献:

[1] Rector J W. Cross-well Methods: Where are We, Where are We Going? [J]. Geophysics, 1995, 60(3): 627-630.

[2] 姚忠瑞, 何惺华, 左建军, 等. 多方位 Walk-away VSP 处理方法[J]. 石油物探, 2006, 45(4): 381-384.

[3] 马德堂, 朱光明, 张文波. 双重网格井间地震层析成像技术[J]. 地球科学与环境学报, 2005, 27(4): 83-86.

[4] Li Y P. 3C VSP Tomography Inversion for Subsurface P- and S-wave Elocity Distribution[C] // 75th SEG Annual Meeting. Expanded Abstracts of 75 th SEG Mtg. New Orleans: Society of Exploration Geophysicists, 2006: 2625-2628.

[5] Xin W. 3C-VSP Imaging and Absorption Coefficient Estimation[C] //74th SEG Annual Meeting. Expanded Abstracts of 74 th SEG Mtg. Houston: Society of Exploration Geophysicists, 2005: 2665-2668.

[6] 杜世通. 地震波动力学[M]. 山东东营: 石油大学出版社, 2003.

[7] Thomsen L. Weak Elastic Anisotropy[J]. Geophysics, 1986, 51(10): 1954-1966.

[8] Faria E L, Stoffa P L. Traveltime Computation in Transversely Isotropic Media [J]. Geophysics, 1994, 59 (2): 272-281.

[9] 张文波. 井间地震交错网格高阶差分数值模拟及逆时偏移成像研究[D]. 西安: 长安大学, 2005.

[10] 张文波, 朱光明, 马德堂, 等. 井间地震初至波分析[J]. 地球科学与环境学报, 2006, 28(1): 70-74.