第 22卷第 4期 武 汉 测 绘 科 技 大 学 学 报 Vol. 22 No. 4
1997年 12月 Journal of Wuhan Technical University of Surveying and Mapping Dec. 1997

# 空间数据发掘和知识发现的框架*

邸凯昌 李德仁 李德毅

(武汉测绘科技大学信息工程学院,武汉市珞喻路 39号,430070)

摘 要 探讨和提出了空间数据发掘和知识发现的理论技术框架,包括空间数据发掘和知识发现的定义与描述、理论框架、从空间数据库中可发现的知识类型及其应用、数据发掘与知识发现方法、空间知识发现系统的结构及开发方法等,最后探讨了空间数据发掘和知识发现的发展方向。

关键词 空间数据发掘(SDM);从数据库发现知识(KDD);空间数据库;地理信息系统

分类号 TP311

数据发掘(Data Mining,简称 DM),或称从数据库中发现知识(Knowledge Discovery from Databases,简称 KDD),定义为"从数据库中发现隐含的、先前不知道的、潜在有用的信息"。[1,2] KDD侧重于目的和结果,DM侧重于处理过程和方法,研究者们经常把它们等同起来,或放在一起使用。DM和 KDD的定义还有一些不同的表达形式,但其本质是一样的,即从数据库中提取隐含的、感兴趣的、高水平的模式。

随着计算机信息处理技术的进步,数据和数据库急剧膨胀,而数据库中隐藏的丰富的知识远远没有得到充分的发掘和利用,数据库急剧增长与人们对数据库处理和理解的困难之间形成了强烈的反差。DM和 KDD技术就是在这种状况下应运而生的,也是人工智能、机器学习技术发展的结果,其目的是为数据库理解与应用提供自动化、智能化的手段。尽管这项技术刚刚起步,但已显示了诱人的前景。同时,它有着相当大的难度,是一项极具挑战性的课题。

空间数据库是一类重要的、特殊的数据库,地理信息系统(GIS)是空间数据库发展的主体,另外还有图像数据库、CAD数据库等。GIS中含有大量的空间和属性数据,有着比一般关系数据库和事务数据库更加丰富和复杂的语义信息,隐藏着丰富的知识。空间数据发掘和知识发现技术,一方面可使 GIS查询和分析技术提高到发现知识的新阶段,另一方面从中发现的知识可构成知识库用于建立智能化的 GIS系统,同时也将促进 3S的智能化集成。

## 1 空间数据发掘和知识发现的定义与描述

空间数据发掘(Spatial Data Mining,简称 SDM),或称从空间数据库中发现知识(Knowledge Discovery from Spatial Databases),是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其它一些隐含在数据库中的普遍的数据特征[3,8]。

从数据库中发现知识是一个在数据库中进行机器学习的过程。李[5]将此问题抽象为一个五元组 $\{T,D,C,L,K\}$,其中 $T$ 表示某种学习任务,$D$ 表示存储在数据库中的大量数据,$C$ 表示一组有助于发现特定知识的基本概念和背景知识,$L$ 是指用来形成各种发现的语言,$K$ 是通过学习发现的知识。Han[3]提出了 KDD的三要素:与任务相关的数据、学习要求(包括要学习的知识类型、所需阈值、知识表达方式等)和背景知识(以概念树的形式给出)。上述两种对 KDD问题的描述本质上是一致的,它们同样适用于对空间数据发掘和知识发现的描述。

知识发现和数据发掘的目的是把大量的原始的数据转换成有价值的知识,用于描述过去的趋势和预测未来的趋势,它可以看成是决策支持过程。同数据库管理系统检索和查询出的信息相比,KDD系统发掘出的知识是隐含、精练和高水平的,数据、信息、知识构成了金字塔结构,如图 1[2]所示。

收稿日期:1997-05-12。邸凯昌,男,30岁,高级工程师,博士生,现从事数据发掘与知识发现、3S集成理论等研究。

* 国家自然科学基金重点资助项目,编号 49631050

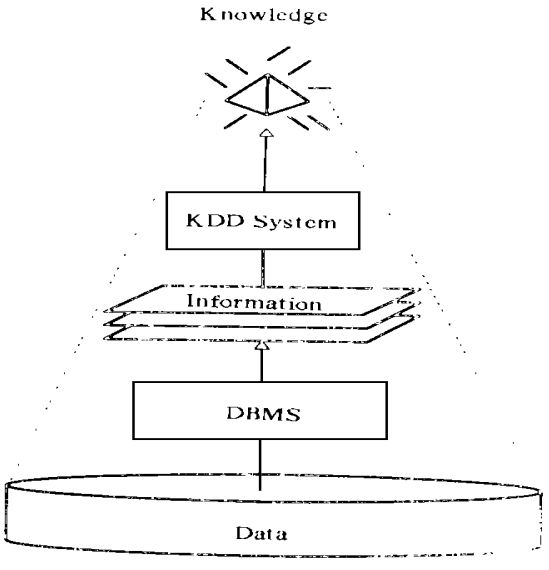


图 1　数据 - 信息 - 知识金字塔

Fig. 1　Data- Information- Knowledge Pyramid

## 2　空间数据发掘的理论框架

研究者们从不同的角度研究 KDD,因而提出了不同的理论框架。有的学者提出用证据理论(Evidence Theory)、Rough集理论(Rough Sets Theory)等作为 KDD的理论框架,李德毅[6]提出了以发现状态空间理论(Discovery State Space Theory)作为 KDD的总体框架,这里作一基本介绍。

发现状态空间是一个三维立体空间,是发现系统实施多种算法的运作空间。在一个二维的平面基底——知识基上逐步抽象,关系数据库可以抽象地看成一个二维通用大表,纵向为属性(Attributes 或 Fileds),横向为元组(Tuples 或 Records)。根据知识发现任务,在原始的数据库经过查询、选择(或抽样)、统计和压缩等数据聚焦处理,形成宏元组(Macro Tuples),是发现状态空间的基底,也可以认为是初始的知识模板。在发现状态空间进行的多种知识汇集和发现操作分成 3个方向: 属性方向(Attribute Oriented),即面向属性的操作,是对属性之间关系的认识和发现活动。宏元组方向(Macro Tuples Oriented),即面向宏元组的操作,是对各种宏元组之间一致性和差异性的认识和发现活动。以上 2个方向的操作都是对特定知识模板的操作。模板方向(Template Oriented),即面向知识模板的操作,是从微观到宏观的发现知识的操作。由一块知识模板上升到抽象级别更高的另一块模板,是提高知识抽象度的操作,是以归纳为核心的知识发现活动。

针对空间数据的特点,我们在三维发现状态空间的基础上增加一维——尺度(scale)维,形成空间数据发掘的四维发现状态空间。在尺度维上,表达了空间数据由细到粗多比例尺或多分辨率的几何变换过程。尺度越小(即比例尺越大),对空间目标表达得越精细、越微观;尺度越大(即比例尺越小),对空间目标表达得越概括、越宏观。例如,在大比例尺数据库中的单个房屋是面状目标,在小比例尺数据库中变为点状目标;在大比例尺数据库线状目标中的细小弯曲在小比例尺数据库中被综合掉。面向尺度的(Scale Oriented)操作,是对空间数据由细到粗的计算、变换、概括、综合过程,地图制图学中的制图综合技术就是典型的面向尺度的操作。

## 3　从空间数据库可发现的知识及应用

由于 GIS数据库是空间数据库的主要类型,并且从 GIS数据库中可发现的知识类型及知识发现方法可以涵盖其它类型的空间数据库,在下文中,我们把 GIS数据库与空间数据库等同起来,并认为从 GIS数据库中发现知识(简称 KDGD)与 SDM有相同的内涵。

从 GIS数据库可以发现的主要知识类型有[3,7~9]:

1)普遍的几何知识(General Geometric Knowledge)

是指某类目标的数量、大小、形态特征等的普遍的几何特征。计算和统计空间目标几何特征量的最小值、最大值、均值、方差、众数等,还可统计出特征量的直方图。在足够样本的情况下,直方图数据可转换为先验概率使用。在此基础上,可根据背景知识归纳出高水平的普遍几何知识。

2)空间分布规律(Spatial Distribution Regularities)

是指目标在地理空间的分布规律,分成在垂直向、水平向以及垂直向和水平向的联合分布规律。垂直向分布即地物沿高程带的分布,如植被沿高程带分布规律、植被沿坡度坡向分布规律等;水平向分布指地物在平面区域的分布规律,如不同区域农作物的差异、公用设施的城乡差异等;垂直向和水平向的联合分布即不同的区域中地物沿高程分布规律。

3)空间关联规则(Spatial Association Rules)

是指空间目标间相邻、相连、共生、包含等空

间关联规则。例如,村落与道路相连,道路与河流的交叉处是桥梁等。

4)空间聚类规则(Spatial Clustering Rules)

空间聚类规则,或空间分类规则,是指特征相近的空间目标聚类成上一级类的规则,可用于GIS的空间概括和综合。例如,将距离很近的散布的居民点聚类成居民区。

5)空间特征规则(Spatial Characteristic Rules)

是指某类或几类空间目标的几何的和属性的普遍特征,即对共性的描述。普遍的几何知识属于空间特征规则的一类,由于它在遥感影像解译中的作用十分重要,所以分离出来单独作为一类知识。

6)空间区分规则(Spatial Discriminate Rules)

指两类或多类目标间几何的或属性的不同特征,即可以区分不同类目标的特征。

7)空间演变规则(Spatial Evolution Rules)

若GIS数据库是时空数据库或GIS数据库中存有同一地区多个时间数据的快照(Snapshot),则可以发现空间演变规则。空间演变规则是指空间目标依时间的变化规则,即哪些地区易变,哪些地区不易变,哪些目标易变及怎么变,哪些目标固定不变。

8)面向对象的知识(Object Oriented Knowledge)

是指某类复杂对象的子类构成及其普遍特征的知识。

可用的知识表达方法有:特征表、谓词逻辑、产生式规则、语义网络、面向对象的表达方法、可视化表达方法等,应根据不同的应用选用不同的表达方法。各种表达方法之间也可以相互转换。

从GIS数据库发现的知识,可有下面两大方面的应用:(1)GIS智能化分析。SDM获取的知识同现有GIS分析工具获取的信息相比更加概括、精练,并可发现现有GIS分析工具无法获取的隐含的模式和规律,因此SDM本身就是GIS智能化分析工具,也是构成GIS专家系统和决策支持系统的重要工具。(2)在遥感影像解译中的应用。用于遥感影像解译中的约束、辅助、引导,解决同谱异物、同物异谱问题,减少分类识别的疑义度,提高解译的可靠性、精度和速度。SDM是建立遥感影像理解专家系统知识获取的重要技术手段和工具,遥感影像解译的结果又可更新GIS数据库。因此,SDM技术将会促进遥感与GIS的智能化集成。

## 4 空间数据发掘与知识发现方法

DM和KDD是多学科和多种技术交叉综合的新领域,它综合了机器学习、数据库、专家系统、模式识别、统计、管理信息系统、基于知识的系统、可视化等领域的有关技术,因而数据发掘与知识发现方法是丰富多彩的。针对空间数据库的特点,我们总结和提出下列可采用的空间数据发掘与知识发现方法[3,7-9]:

1)统计方法

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。

2)归纳方法

即对数据进行概括和综合,归纳出高层次的模式或特征。归纳法一般需要背景知识,常以概念树的形式给出。在GIS数据库中,可有属性概念树和空间关系概念树两类。背景知识由用户提供,在有些情况下也可以作为知识发现任务的一部分自动获取。

3)聚类方法

聚类分析方法按一定的距离或相似性测度将数据分成一系列相互区分的组,它与归纳法不同之处在于不需要背景知识而直接发现一些有意义的结构与模式。经典统计学中的聚类分析方法对属性数据库中的大数据量存在速度慢、效率低的问题,对图形数据库应发展空间聚类方法

4)空间分析方法

空间分析方法可采用拓扑结构分析、空间缓冲区及距离分析、叠置分析等方法,旨在发现目标在空间上的相连、相邻和共生等关联关系。

5)探测性的数据分析

探测性的数据分析,简称EDA,采用动态统计图形和动态链接窗口技术将数据及其统计特征显示出来,可发现数据中非直观的数据特征及异常数据[13,14]。EDA与空间分析(Spatial Analysis)相结合,构成探测性的空间分析(简称ESA)。EDA和ESA技术在知识发现中用于选取感兴趣的数据子集,即数据聚焦(Data Focusing),并可初步发现隐含在数据中的某些特征和规律。

6) Rough集方法

Rough集理论 ( Rough Sets Theory)是波兰华沙大学 Z. Pawlak教授在 1982年提出的一种智能数据决策分析工具[10],被广泛研究并应用于不精确、不确定、不完全的信息的分类分析和知识获取[11]。

Rough集理论为 GIS的属性分析和知识发现开辟了一条新途径,可用于 GIS数据库属性表的一致性分析、属性的重要性、属性依赖、属性表简化、最小决策和分类算法生成等[9]。

Rough集方法与其它知识发现方法相结合,可以在 GIS数据库中数据不确定情况下获取多种知识。例如,在经过统计和归纳从原始数据得到普遍化数据的基础上, Rough集用于普遍化数据的进一步简化和最小决策算法生成,使得在保持普遍化数据内涵的前提下最大限度地精练知识。

7)云理论

这是由李德毅博士提出的用于处理不确定性的一种新理论,由云模型 (cloud model)、不确定性推理 ( reasoning under uncertainty)和云变换 (cloud transform)三大支柱构成。云理论将模糊性和随机性结合起来,解决了作为模糊集理论基石的隶属函数概念的固有缺陷,为 KDD中定量与定性相结合的处理方法奠定了基础[12]。

8)图像分析和模式识别

空间数据库中含有大量的图形图像数据,一些行之有效的图像分析和模式识别方法可直接用于发现知识,或作为其它知识发现方法的预处理手段。

另外,决策树 ( Decision Tree)、神经网络 ( Neural Network)、证据理论 ( Evidence Theory)、模糊集 ( Fuzzy Sets)理论、遗传算法 (Genetic Algorithms)等也可用于空间数据发掘和知识发现。

当然,这些方法不是孤立应用的,为了发现某类知识,常常要综合应用这些方法。知识发现方法还要与常规的数据库技术充分结合。例如,在时空数据库中发掘空间演变规则时,可利用空间数据库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。又如,我们把面向属性的归纳方法 ( Attribute Oriented Induction)与探测性的数据分析和 Rough集方法结合起来,构成探测性的归纳学习方法 ( Exploratory Inductive Learning),可用于发现空间特征规则、空间区分规则、普遍的几何知识、空间演变规则、空间分布规律等多种知识。

## 5　空间知识发现系统的结构及开发方法

借鉴有关专家提出的 KDD系统的结构[1,3,4],我们提出一种空间知识发现系统的结构,见图 2。图中单线箭头方向为控制流,实心箭头方向为信息流。从图中可以看出,知识发现同空间数据库管理是密切联系的,用户发出知识发现命令,知识发现模块触发空间数据库管理模块从空间数据库中获取感兴趣的数据,或称为与任务相关的数据,知识发现模块根据知识发现要求和领域知识从与任务相关的数据中发现知识,发现的知识提供给用户应用或加入到领域知识库中,用于新的知识发现过程。一般说来,知识发现要交互地反复进行才能得到最终满意的结果,所以,在启动知识发现模块之前,用户往往直接通过空间数据库管理模块交互地选取感兴趣的数据,用户看到可视化的查询和检索结果后,逐步细化感兴趣的数据,然后再开始知识发现过程。

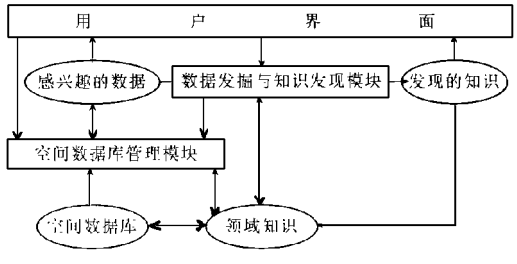


图 2　空间知识发现系统的结构

Fig. 2　An Architecture of Spatial Knowledge Discovery System

在开发知识发现系统时,有两个重要的问题需要考虑和作出选择: (1)自发地发现还是根据用户的命令发现。自发地发现会得到大量不感兴趣的知识,而且效率很低,根据用户命令执行则发现的效率高、速度快,结果符合要求。一般应采用交互的方式,对于专用的知识发现系统可采用自发的方式。(2) KDD系统如何管理数据库,即 KDD系统本身具有 DBMS功能还是与外部 DBMS系统相连。KDD系统本身具有 DBMS的功能,系统整体运行效率高,缺点是软件开发工作量大,软件不易更新。KDD系统与外部 DBMS系统结合使用,整体效率稍低,但开发工作量小,通用性好,易于及时吸收最新的数据库新技术成果。由于 GIS系统本身比较复杂,在开发 SDM工具时应在 GIS系统之上进行二次开发。

根据对上述两个问题的考虑,提出下列开发

空间知识发现系统的建议。用通用 GIS的二次开发工具及 Visual Basic或 Visual C++ 在 Windows及 Windows95环境下开发,采用 ODBC标准及 OLE DLL编程技术提高软件的通用性和开放性。支持常用的标准数据格式。SDM系统可单独使用,也可作为插件式(Plug In)软件附着在 GIS系统之上使用,或者 SDM系统本身就是未来智能化 GIS系统的有机组成部分。知识发现算法可自动地执行,又要有较强的人机交互能力。用户可定义感兴趣的数据子集,提供背景知识,给定阈值,选择知识表达方式等。若不提供所需参数,则自动地按缺省参数执行。

## 6 空间数据发掘发展方向探讨

在 SDM的理论和方法方面,重要的研究方向有[8,9]:背景知识概念树的自动生成、不确定性情况下的数据发掘、递增式数据发掘、栅格矢量一体化数据发掘、多分辨率及多层次数据发掘、并行数据发掘、新算法和高效率算法的研究、空间数据发掘查询语言、规则的可视化表达等等。在 SDM系统的实现方面,要研究多算法的集成、SDM系统中的人机交互技术和可视化技术、SDM系统与地理信息系统、遥感解译专家系统、空间决策支持系统的集成等。

### 参 考 文 献


1 Frawley W, Piatesky-Shapiro G, Matheus C. Knowledge Discovery in Databases: An Overview. In: Piatesky-Shapiro G, Frawley W. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

2 Piatesky-Shapiro G. An Overview of Knowledge Discovery in Databases: Recent Progress and Challenges. In: Ziarko W. Rough Sets, Fuzzy Sets and Knowledge Discovery. Springer-Verlag, 1994.

3 Han J. Data Mining Techniques. ACM-SIGMOD 96 Conf. Tutorial, 1996

4 Matheus C, Chan P K, Piatesky-Shapiro G. System for Knowledge Discovery in Databases. IEEE Trans. on Knowledge and Data Engineering, 1993, 5(6)

5 李德毅.归纳学习:从数据库中发现知识.沈阳:第十届全国数据库学术会议,1992

6 李德毅.发现状态空间理论.小型微型计算机系统,1994, 15(11)

7 李德仁,程 涛.从 GIS数据库中发现知识.测绘学报,1995, 24(1)

8 邸凯昌,李德仁. KDD技术及其在 GIS中的应用与扩展.北京:中国 GIS协会第二届年会,1996

9 Koperski K, Adhihary J, Han J. Spatial Data Mining: Progress and Challenges. SIGMOD 96 Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'96). Canada: Montreal, 1996.

10 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991.

11 Ziarko W. Rough Sets, Fuzzy Sets and Knowledge Discovery. Springer-Verlag, 1994.

12 Li Deyi. Knowledge Representation in KDD Based on Linguistic Atoms. Singapore: 1st Pacific-Asia Conf on KDD&DM, 1997

13 Haslett J. SPIDER- An Interactive Statistical Tool for the Analysis of Spatially Distributed Data. Int. J. GIS, 1990, 4(3)

14 Batty M, Xie Y. Modelling Inside GIS: Part 1. Model Structures, Exploratory Spatial Data Analysis and Aggregation. Int. J. GIS, 1994, 8(3)


## A Framework of Spatial Data Mining and Knowledge Discovery


*Di Kaichang Li Deren Li Deyi*

(School of Information Engineering, WTUSM, 39 Luoyu Road, Wuhan, China, 430070)



**Abstract** A framework of spatial data mining and knowledge discovery is proposed. The concept and theoretical model of spatial data mining, knowledge types, methods suitable for knowledge discovery in spatial databases, the architecture and development strategy of the spatial knowledge discovery system, etc., are presented and described in the framework. The future directions of spatial data mining and knowledge discovery are discussed at last.



**Key words** spatial data mining(SDM); knowledge discovery form databases(KDD); spatial database; geographic information system(GIS)