空间数据挖掘与GIS集成及应用研究

毛克彪[1]，覃志豪[1]，李 昕[2]，李海涛[1]

（1. 南京大学国际地球系统科学研究所，江苏 南京 210093；2. 南京大学计算机系，江苏 南京 210093）

摘 要：阐明空间数据挖掘与GIS集成的优越性，分析空间数据挖掘与关系数据库系统的区别，介绍面向对象技术对空间数据挖掘和空间数据挖掘的常用算法，在此基础上介绍地理信息系统与空间数据挖掘工具及应用。

关键词：空间数据挖掘；GIS；面向对象；关系数据库系统

中图分类号：P 209 **文献标识码**：A **文章编号**：1672－5867(2004)01－0014－04

The Integration and Application of Spatial Data Mining and GIS

MAO Ke－biao[1], QIN Zhi－hao[1], LI Xin[2], LI Hai－tao[1]

(1. International Institute for Earth System, Nanjing University, Nanjing 210093, China; 2. Department of Computer Science, Nanjing Universty, Nanjing 210093, China)

Abstract: With the development of the GIS database, the data are so large and come from different sources that the data can not be dealt with by human brains. Therefore, it becomes increasely important to find a method that can automatic, quickly, efficiently, get more information from the GIS database. The technology of Data Mining meets the need. Therefore, the technology of Data Mining is introduced into GIS. This paper describes the difference between the spatial data mining and the RDMS and some methods of spatial data mining. Furthermore, a system of integration of spatial data mining and GIS and its application are introduced.

Key words: spatial data mining; GIS; object－oriented; RDMS

0 引 言

近年来，大量的空间数据从遥感、地理信息系统等多种应用中得到，传统的信息和知识抽取方法受到极大的挑战，迫切要求引入数据挖掘思想，以便更好地分析复杂的空间现象和空间对象。但是，从大量的空间数据中发现知识不同于一般的数据挖掘。关系型及事务型数据挖掘算法的一个重要前提是假定数据是独立的，而在空间数据库中一个对象可能会受其邻近若干个对象的影响。空间数据带有拓扑及距离信息，通常由复杂的多维空间索引结构组织，并通过空间数据存取方法存取，其间需要空间推理、几何计算和空间知识表示等技术。这些特性使得空间知识的发现极具挑战性。空间数据挖掘指的是从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中的其它模式，用于理解空间数据、发现空间和非空间数据间的关系、构建空间知识库、查询优化、空间数据库数据重组，以简单精确的方式描述通用特征等等[1]。

1 空间数据挖掘与关系数据库系统的区别

空间数据挖掘的算法高度依赖于邻近关系的处理。因为在一个特定的算法里，许多邻近目标需要考虑，这样就提供了一个邻近关系的概念，这些概念的实现能够允许空间数据挖掘算法和空间数据库管理系统高度集成。普通的数据挖掘算法可被一些基本的数据操作支持。

关系数据库系统和空间数据库系统的主要区别在于在空间数据库系统中选取的目标的近邻属性会对它有影响，因此邻近点需要被考虑。在空间数据挖掘算法中使用空间目标的绝对定位和扩展定义空间的邻近关系，即拓扑关系、距离和方位[2][3]。

拓扑关系是指在拓扑转换下各个对象间的关系。在这里，拓扑关系包括分离、相遇、交迭、相等、

收稿日期：2003－08－25

基金项目：国家重点基础研究发展规划(973)项目(2001CB309404)；海外青年学者合作研究基金(40128001)和教育部科学技术重点项目(2001)

覆盖、被覆盖、包含和被包含等。简单说来,两个对象间的拓扑关系可以这样理解:对象 A 和 B 的相邻关系在拓扑图中用连接 A、B 的线条表示(如图 1 所示)。不论两个对象间"接触"的面积有多大(比较 A、C 之间和 B、D 之间),只要它们相邻,一概用连接这两个对象的线条表示。

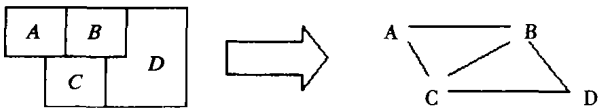

图 1　拓扑结构示意图

Fig. 1　Topological structure sketch

在关系型数据库里,两个元组之间的"距离"关系是人为的概念上的定义,即任取两个属性 X 和 Y,则元组 A 和 B 间的距离通常定义为 $f(A,B) = (Ax - Bx)2 + (Ay - By)2$。但在空间数据库里,距离函数 $f(A,B)$ 是有实际意义的。例如 $f(A,B) = 100$ 可能指 AB 两地的距离为 100 km。可以定义基于距离函数 $f(A,B)$[4] 的距离关系 $r, r \in \{<, >, =\}$。ArB 当且仅当 $f(A,B)\ r\ K$,其中 K 为某个阈值。

方位关系是指找准一个参照对象 A 和一个需定位的对象 B,方位关系的定义如图 2 所示。

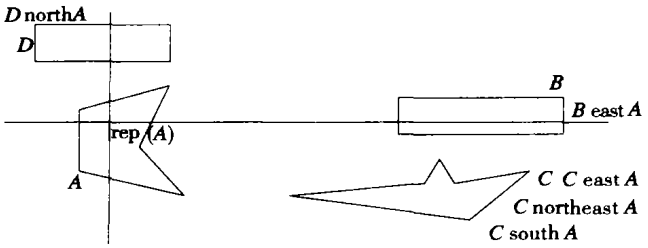

图 2　方位关系示意图

Fig. 2　Bearing relationship sketch

设 rep(A) 是参照对象 A 中的一个特征点[5]:

1) B 在 A 的东北方向, iff $\forall b \in B: b_x \geq \mathrm{rep}(A)_x \wedge b_y \geq \mathrm{rep}(A)_y$,同理,东南,西南,西北方向的确定条件;

2) B 在 A 的北方, iff $\forall b \in B: b_y \geq \mathrm{rep}(A)_y$,同理南、西、东;

3) B 在 A 的某一个方向,对于所有的 A 和 B,此关系都为真。

2　面向对象技术对空间数据挖掘的支持[6]

空间数据库需要空间属性的正确表示。在空间数据查询、约束、算法方面,一个有效的方式是进行空间操作。GIS 的观念、逻辑和物理设计需要使用面向对象、关系数据库和已存在的空间数据应用。基于实体的观念模型,已扩充了面向对象的查询和约束特性,在空间数据库中已被定义,而且被赋予了查询和约束的谓词语言。这种空间查询和操作称为空间关系代数。另外,在面向对象的对象信息数据库中可用模糊逻辑进行空间数据挖掘。美国海军 Stennis 空间中心地图科学实验室已意识到空间数据仓库和数据库与地理信息数据库以及与空间数据挖掘集成的益处。用面向对象的方式建立目标模型,它能够很容易地扩展到包括所有的地理数据类型。用面向对象技术和标准如 CORBA 和 VRML 能够使二维和三维的数据在 Internet 上显示。

3　空间数据挖掘算法

由于空间数据的关联特性,空间数据挖掘算法必须对传统的挖掘算法加以拓展才能更好地分析复杂的空间对象。评价一个空间数据挖掘方法主要看它的挖掘效率、与数据库结合层度、与用户的交互能力,以及发现新知识的能力等。空间数据挖掘的技术有聚类、分类、空间关联等等。

3.1　聚类算法

简单说来,聚类就是要将数据库中的对象根据一定的要求分为若干个有特定意义的对象子集,所得的子集称为簇(在聚类时,要使簇中的对象尽可能相近,而使不同簇的对象尽可能相异)。聚类后可以得到整个数据库中对象的分布模式。这里简单介绍 Kaufman 与 Rousseeuw 提出的 PAM 算法(Partition Around Mediods)和 Raymond T. Ng 与 Jiawei Han 提出的 CLARANS 算法[7]。

3.1.1　PAM 算法

PAM 算法首先假定若干个对象为各个集簇中位于中心位置的对象(称为 Medoid)。然后将其他未选定的对象根据其与各个选定对象的距离加入到最相近的集簇中。

若选定的集簇中心为对象 Oi,某个未选定的对象 Oj 与 Oi 的距离可以定义为 d(Oi,Oj)。比较对象 Oj 与各个集簇中心的距离,若 d(Oi,Oj) 为最小值,则称对象 Oj 属于 Oi 所在的集簇。集簇中心是可以调整的,计算将集簇中心由对象 Oi 改为 Oj 的代价函数 Tij,若代价函数为负值,则将集簇中心由 Oi 转至 Oj;否则不变。

3.1.2　CLARANS 算法

CLARANS 算法是基于局部最优的思想。其聚类过程可以理解为查找一个图,图中的每个节点都是潜在的解决方案。这里的节点是对象的集合,记为 S。则两个节点 $S1$ 和 $S2$ 是邻居,当且仅当集合 $S1$ 和 $S2$ 只有一个元素相异。算法具体参见[7]。CLARANS 算法并不遍历求解空间,也不限制于具

体的采样方式。其迭代次数与参数 max - neighbor 和 num - local 的设定有关(参数的最优设定在此不加详述),计算复杂度与对象的数量基本呈线形关系。基于 CLARANS 算法的空间数据聚类算法又可以细分为空间支配算法和非空间支配算法

3.2 分类算法

分类是将数据库中的对象根据一定的意义划分为若干个子集。它和聚类算法的差别在于聚类算法是根据一定要求将对象聚为一个集合,最后得到的分布模式是聚类之前并未确知的;分类算法是根据已知分布模式的属性要求将数据库对象归入相应的分类中。

Krzysztof Koperski、Jiawei Han 和 Nebojsa Stefanovic[8] 对 Ester 等人提出的空间对象分类算法加以改进,降低了算法的时间复杂度。该算法使用决策树对空间对象进行分类。假定算法的输入部分为:

1)空间数据库,其中包含已分类对象 Oc 和其他的具有非空间属性的空间对象;

2)非空间概念层次的集合;

3)空间挖掘查询要求的详细说明。

算法的输出则是二叉决策树。这个算法的基本思路是先对数据对象进行抽样,在得到的较小的抽样集上进行第一次"挖掘",得到可能的谓词描述,然后对更多的数据对象在已有谓词描述的基础上再次"挖掘",得到最后的决策树,这大大提高了算法的性能。具体算法详见[9]。

3.3 基于空间关联的算法

空间关联是将一个或多个空间对象与其他空间对象相关联。

空间关联规则的形式是 XúY(c%),其中,X、Y 是空间或非空间谓词的集合,c% 为规则的置信度。空间谓词的形式有三种:表示拓扑结构的谓词,表示空间方向的谓词和表示距离的谓词。

Krzysztof Koperski 和 Jiawei Han 利用空间数据的关联特性改进其分类算法,使得它适合于挖掘地理数据中的相关性。

总的说来,空间数据挖掘算法是对一般挖掘算法的特殊化。或者借用一般算法的思想并为之赋予新的含义,或者改进算法使其更适合空间数据挖掘。

4 GIS 与空间数据挖掘工具集成的实例及应用

GIS 和数据挖掘一直是两个分开的学科,直到近来,才被意识到两者结合的巨大潜力。而且由于 GIS 空间数据存储的特殊性,目前对二者集成的研究还只是刚刚起步[9]。下面介绍国外的一个实例:地理信息系统 Descartes 和数据挖掘工具 Kepler 的集成 Descartes 系统提供独特的性能[9],首先是智能地图绘制的支持;其次是空间数据交互可视化分析的频谱功能。

Descartes 能够自动产生地图、表示用户选择的数据支持多种地图显示的交互式操作。支持可视分析的数据转换,支持对已存变量的逻辑查询和算术操作方式驱动的动态计算。Kepler 数据挖掘系统提供易用、灵活和有力的结合各种数据挖掘方式的平台。此开放平台为加入新的方法提供一个通用的 plug - in 接口。kepler 支持整个的数据挖掘过程,包括数据输入工具、格式转换工具、数据库查询、管理,及不同种类数据挖掘结果(trees、rules、groups)的地图表示。

这种集成工具产生新一代的空间数据分析。为了完全支持空间相关数据的分析,在 Kepler 和 Descartes 之间必须建立联接。基本观点是:一个分析可看得见原数据,可看得见表达空间信息的地图和统计图标形式的数据挖掘结果。这种分析能够很容易地发现空间关系和模式。从概念上,这种集成系统有三个连接:

1)从"geography"到"mathematics":当可视化地开发和操作地图时,用户可发现空间现象;他可用数据挖掘的方式对此发现做出判断。

2)从"mathematics"到"geography":数据挖掘产生的结果可视化地在地图上表示出来。

3)两者之间的对话(连接显示):不管是可视还是不可视数据,在同一个时刻,数据保持一致。

这种集成系统是客户 - 服务器结构,服务端是 C++ 实现,客户端是 java,在 windows 和 Unix 下都可使用。

5 空间数据挖掘与 GIS 集成的具体应用

空间数据挖掘与 GIS 集成有着广泛的应用前景,两者之间的集成研究刚刚起步,技术上还不成熟,但它的应用价值已经体现出来。

5.1 空间数据挖掘在市场经济中的应用

WEBGIS 是一个可以在不同的操作系统上提取地理空间及属性数据并能提供分析的一个系统。它在市场经济,特别是电子商务中的应用不仅表现为选址,而且还可以表现为物流、客流分析等。因此基于 WEBGIS 的电子商务数据挖掘是一个重要的研究方向。

5.2 GIS 对数据进行可视化分析

GIS 技术已从大型的图像系统得到了发展。现

在的商业地图软件不仅实用,而且容易掌握。这种软件允许用户综合空间和关系数据库系统层次不受数量限制的数据。地图中的目标有以下形式:区域、点,每个地图目标附着一些描述它的详细的数据,如客户和商业人口统计,历史的客户的购买方式等。

利用 GIS 可以直观地显示商业人口、聚居人口、某个给定区域的一定范围的平均收入等,可以在地图上清晰地选择出客户可能的竞争选择等。其中人口报告通常用来产生潜在的商店、主要的营销区域、个体客户的详细描述等。但这些信息在 GIS 里不会自动产生有决定性的决策,然而借助数据挖掘技术可以做到。

5.3　数据挖掘提供决策支持

数据挖掘从统计学、人工智能和符号处理进化而来,主要的技术有神经网络、聚类、遗传算法、模糊逻辑、决策树、各种衰退方式、主要的组件分析、因素分析等。这些技术最重要的目标是从大量的历史数据中提取重要的趋势和模式,从而对将来的情况做出预测。空间数据挖掘技术加强 GIS 的力量在于用一种有意义的方式,为综合各种数据系统选择一个丰富的数据结构[6]。采用 GIS 分析技术,把人口统计报告变成真正的市场智能的关键是分析数据,精简大量的数据为单个的预见或评分。这就是预见性数据挖掘的力量所在,这就是要把 GIS 和数据挖掘集成的原因。基于特定的应用,GIS 能够联合客户的历史数据或商店的销售记录及企业的人口统计、商业、运输、市场研究数据,这些数据建立预见性的模型,用来评价有潜力的新的区域和顾客、交叉性买卖、目标市场、客户摆动和其它相似的应用是非常理想的,这是一个非常好的应用研究领域。其具体应用流程见图 3

6　结　语

空间数据挖掘与 GIS 集成的研究刚刚起步,是一个重要的研究课题。二者集成把 GIS 技术提高到发现新知识的阶段,发现的知识可构成知识库用于建立智能化的知识 GIS 系统,将使得系统具有自动学习的功能,使系统自动获取知识,将 GIS 系统建立成真正的智能系统,而且会有效地促进 GPS、RS 与 GIS 等技术的集成。

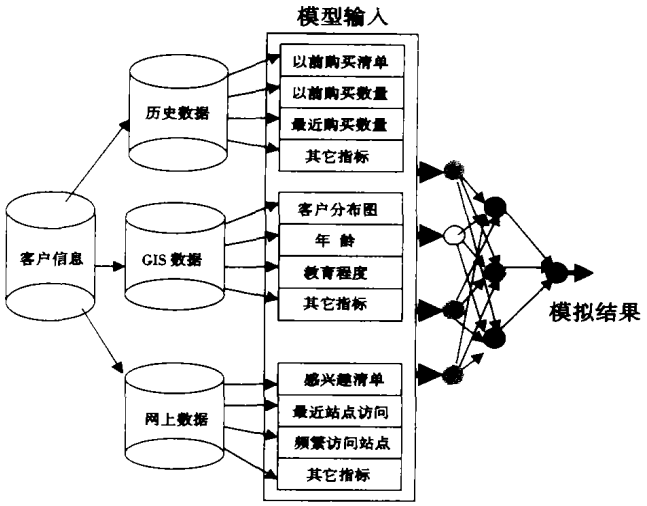

图 3　对不同数据源模型训练示意图

Fig. 3　Model training of different data sources

参考文献:

[1] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques. 北京:高等教育出版社,2000.

[2] 黄杏元、马劲松、汤　勤. 地理信息系统概论[M]. 北京:高等教育出版社,2001.

[3] 毛克彪,覃志豪,李海涛,周若鸿. 基于空间数据仓库的空间数据挖掘研究[J]. 遥感信息,2002.

[4] Martin Ester, Hans - Peter Kriegel, J rg Sander, Knowledge Discovery in Spatial Database, Institute for Computer Science, University of Munich, Oettingenstr, 2000,4,193 216.

[5] Martin Ester, Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support Data Mining and Knowledge Discovery ,2000,,193 - 216.

[6] M. ChungR. Wilson K. Shaw Spatial Data Mining Using Fuzzy Logic In An Object - Oriented Geographical Information Database.

[7] R. Ng, J. Han. Efficient and effective clustering method for spatial data mining. In: Proc of Int'l Conf. VLDB, San Francisco, CA: Morgan Kaufmann, 1994. 144 - 155.

[8] K. Koperski and J. Han: Discovery of Spatial Association Rules in Geographic Information Databases, Proc. 4th Int. Symp. On Large Spatial Databases (SSD 95), Portland, ME, 1995, pp47 - 66.

[9] Natalia Andrienko, Gennady Andrienko, Alexandr Savinov and Dietrich Wettschereck, Descartes and Kepler for Spatial Data Mining.

作者简介:

毛克彪(1977 -),男,硕士研究生,主要从事空间数据挖掘、遥感数字图像信息提取,遥感和 GIS 应用等方面的研究。