空间数据挖掘技术简介

王孟娅，姜孟冯

1 引言

近年来，由于卫星、雷达、传感器等高新技术的迅猛发展，空间数据的数量、大小和复杂性都正在急速的增加。如何有效的利用这些空间数据、使其合理地与非空间数据结合，得出人们想要的有价值的信息，是“空间数据挖掘”所要面对的主要课题。

举一个简单的例子：传统的数据库管理系统(Database Management System，DBMS)可以很轻松地回答“列出人口在 500 万以上的城市”这样的问题；但却很难回答诸如“列出长江周围方圆 50 公里以内的城市”。这主要是因为空间数据在表示上很难用传统的单一数值类型记录来刻画，人们需要新技术来解决这一难题。

2 基础知识

2.1 空间数据库管理系统(Spatial Database Management System, SDBMS)

Shekhar 与 Chawla 所著《Spatial Databases: A Tour》[1]中给出了 SDBMS 的定义：

1) 一个 SDBMS 是一个软件模块，它利用一个底层数据管理系统(如 ORDBMS、OODBMS)
2) SDBMS 支持多种空间数据模型、响应的空间抽象数据类型(ADT)以及一种能够调用这些 ADT 的查询语言
3) SDBMS 支持空间索引、高效的空间操作算法以及用于查询优化的特定领域规则。

2.2 地理信息系统(Geographic Information System, GIS)

地理信息系统(GIS)提供了一套地理数据(以地球表面作为基本参照框架的空间数据)可视化的机制，以及诸如“搜索”、“定位分析”、“分布”、“度量”等常用的空间数据分析处理操作。通常，GIS 是不考虑 I/O 成本的，它假定所有数据都在主存里，而 SDBMS 的主要功能却是海量数据的按条件查询，所以有了以下这种结合方案：以 GIS 作为 SDBMS 的前端，在 GIS 对空间数据进行分析之前、先通过 SDBMS 访问这些数据，便可以大大提高 GIS 的效率和生产率。

这样就确立了一个“空间应用(GIS)←空间数据库管理系统(SDBMS)←传统数据库管理系统(DBMS)”的三层体系结构，实际上现在的商业应用程序也是这么实现的。Oracle 公司的 Oracle Spatial 以 SDO_GEOMERTRY 对象的形式把一个地理对象存储在一个常规表的字段里；而 ESRI 公司的 Arc Spatial

Data Engine 更是一个名副其实的中间件。现在常见的 GIS 应用软件有 AutoCAD® Map 3D、ArcGIS、MapInfo 等。

2.3 空间数据挖掘(Spatial Data Mining，SDM)

数据挖掘(又称 Knowledge discovery from data，KDD)，本意就是从数据中发现有意义的信息模式(interesting pattern)的过程[2]。而空间数据挖掘(SDM)一词最早由 Roddick 与 Spiliopoulou 两人在 1999 年提出，认为 SDM 可以被简单地考虑为多维环境下的时空数据挖掘(temporal data mining) [3]。到了 2003 年，Shekhar 与 Chawla 提出，SDM 是发现隐藏在空间数据库中有意义的、潜在有用的信息模式(pattern)的过程 [1][4]。国际上有代表性的通用 SDM 系统有：加拿大 Simon Fraser 大学的 GeoMiner[5]，德国 Knowledge Discovery Team 以 Java 写成的 Descartes[6]和美国 ESRI 公司的产品 ArcGIS[7]。

3 空间数据模型

3.1 空间数据概念模型—场(field)模型与对象(object)模型

空间数据概念模型有两大类：场(field)模型与对象(object)模型，以下例进行简单说明：假设我们需要建立一块地域的空间数据模型，该地域只有草地、水泥地和柏油路三种地形且互不重叠，如**图 1** 所示。

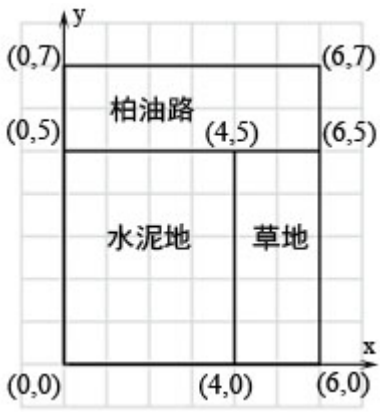

图 1 空间数据模型例子

区域 ID	主要地形	区域/边界
F1	柏油路	(0,5)(6,5)(6,7)(0,7)
F2	水泥地	(0,0)(4,0)(4,5)(0,5)
F3	草地	(4,0)(6,0)(6,5)(4,5)

图 2 对象的观点，每个对象有唯一的标识符、主要地形和一块多边形区域。区域边界由坐标指定

$$f(x,y)=\begin{cases}\text{"柏油路"}, & 0\le x\le 6; 5<y\le 7\\ \text{"水泥地"}, & 0\le x\le 4; 0\le y\le 5\\ \text{"草地"}, & 4<x\le 6; 0\le y\le 5\end{cases}$$

图 3 场的观点，这时地域中的每个点被映射为主要地形所对应的值

图 2 与**图 3** 分别代表了对象模型与场模型的基本表示，不难看出，对象模型比较适用于有固定形状的空间实体，例如湖泊、道路、城市等；而场模型，场模型比较适用于无固定形状的概念，例如云区、火灾、洪水等。

3.2 空间数据逻辑模型与开放地理数据互操作规范 OGIS

空间数据的逻辑模型有很多，主流是面向对象数据模型，与概念模型中的对象模型相对应。在对象模型中，区域的概念尤其重要，它必须以一组数据类型确定出一个多边形，并且保证几何操作在这组数据类型上是封闭的(closure)。开放地理信息系统协会 OGC (Open GIS Consortium)在 1999 年提出了 OGIS 标准(Open Geodata Interoperability Specification)，规定几何信息分为四类：点(point)、曲线(curve)、面(surface)和几何体集合(geometry collection)[8]。真正的线对象，例如世界地图上的河流，用线串(linestring)表示，linestring 是 curve 的一个子类。而常用来表示区域的面对象，例如世界地图上的国家，用多边形(polygon)表示，polygon 是 surface 的一个子类。

3.3 空间数据的空间关系

空间关系是指地理空间实体之间相互作用的关系。空间关系主要有：

1) 拓扑空间关系：用来描述实体间的相邻、连通、包含和相交等关系；
2) 顺序空间关系：用于描述实体在地理空间上的排列顺序，如实体之间前后、上下、左右和东、南、西、北等方位关系；
3) 度量空间关系：用于描述空间实体之间的距离远近等关系。

利用 Egenhofer 提出的 9 交模型[9]，可以得出 8 个拓扑关系：相离(disjoint)、相接(meet)、交叠(overlap)、相等(equal)、包含(contain)、在内部(inside)、覆盖(cover)、被覆盖(covered by)。而在 SQL3/SQL99 中,这些关系得到了语言标准上的支持及确切化:Equal、Disjoint、Intersect、Touch、Cross、Within、Contains、Overlap，另外还添加了大量常用计算操作如 Distance、Buffer、Intersection、Union、Difference、SymmDiff 等。

如此我们便可以回答一开始的问题了，假设我们有两张表如下：

```
CREATE TABLE City ( Name varchar(30), Pop integer, Shape Point);
CREATE TABLE River ( Name varchar(30), Origin varchar(30), Length Number, Shape
LineString);
```

则“列出长江周围方圆 50 公里以内的城市”可对应以下查询：

```
Select Ci.name
FROM City Ci, River R
WHERE OverLap(Ci.Shape,Buffer(R.Shape,50)) = 1 AND R.Name = ‘长江’
```

这其中，Shape 是一个抽象的空间数据类型，代表空间对象，城市被指定为点，而河流则被指定为线串。OverLap(x,y)=1 表示对象 x 与对象 y 两个几何体的内部存在非空交集。Buffer(x,N)返回以 x 对象为中心，N 作为尺寸得到的几何区域。很多 GIS 应用都采用 Buffer 运算，包括洪水泛滥区管理以及城市与乡村划分法则。

4.空间数据挖掘

4.1 空间数据挖掘过程

相对于空间数据库的管理系统而言，空间数据挖掘(SDM)面对的是更为复杂的“问题”。如图 4 所示，数据挖掘过程需要领域专家与数据挖掘分析员进行密切交互，挖掘出一组假设(模式)。这些结果可以用统计工具进行严格验证，可用 GIS 可视化表现出来。最后，分析员可以解释这些模式，指定并推荐合适的方案。

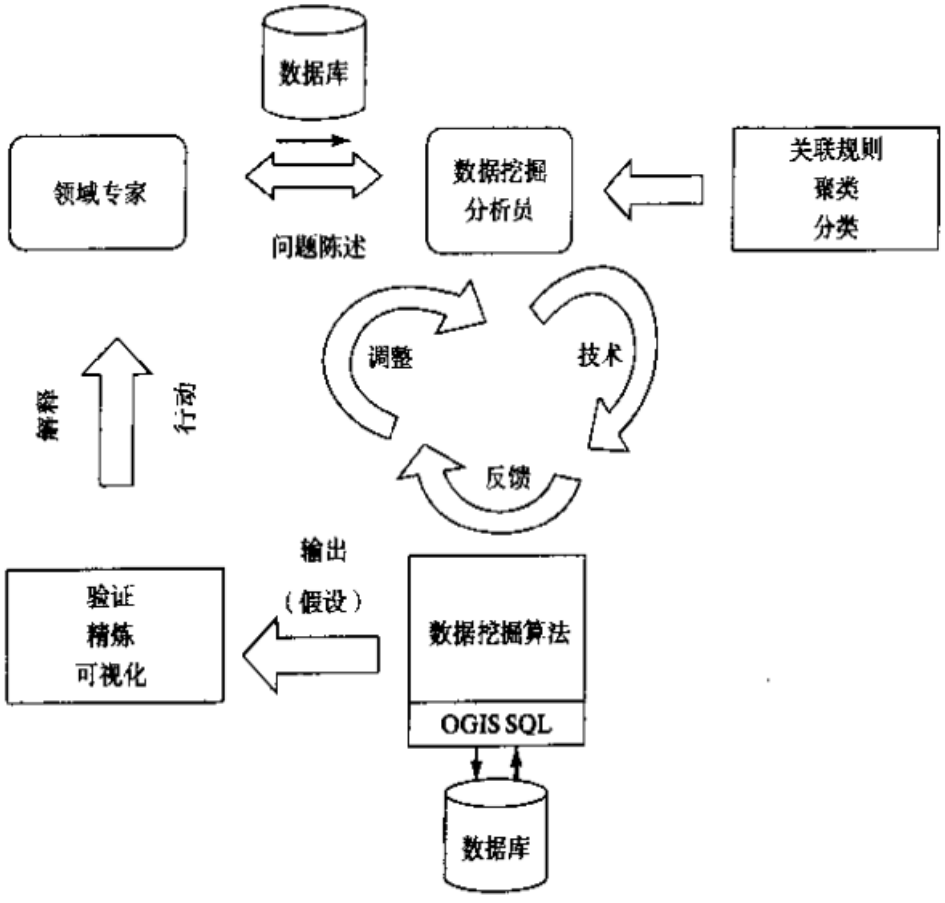

图 4 数据挖掘过程

4.2 空间数据挖掘难点

空间数据挖掘和传统的数据挖掘存在明显差异，主要体现在下述四个方面[4]：

1) 空间数据的复杂性

主要是指空间数据对象模型及空间数据对象之间关系的复杂化。

2) 统计学基础

空间数据的样本往往是自相关的,这导致空间数据样本的独立性假设往往不能成立。

3) 可发现知识类型

与数据挖掘可发现广义型知识、分类型知识、关联型知识和预测型知识相对应，空间数据挖掘发现的知识主要有四种：1、空间聚类和分类；、空间离群点挖掘；3、空间关联规则；4、预测。

4) 算法过程

空间数据挖掘的算法必须考虑空间对象的访问方式与数据结构，效率问题更加严峻。算法结果通常包含太多图形信息，难以用文字表达，因此可视化问题更为重要。

4.3 空间自相关性(spatial autocorrelation)

地理学第一定律：每一个事物都与其他事物相关，但邻近事物间的相关性比距离较远的事物间的相关性要大得多[10]。空间自相关性是空间采样变量经常表现出的一种性质。例如，两个相互邻近的位置的降雨量非常接近，又如气压随空间的变化而渐变。在统计学中，空间自相关性可以用 Moran’s I 系数度量。空间统计学家经常使用空间自相关性的局部度量来追踪空间依赖性如何在同一空间层的不同区域发生变化。局部自相关性在不同位置有较大变化预示着空间异质的存在。

5 常用空间数据挖掘技术

5.1 分类(classification)

简单地讲，分类就是找到一个函数：$f:D->L$.其中，f 的域 D 是属性数据的空间，L 是标号的集合。分类问题的目标是根据给定的有限子集 $Train \subset D \times L$ 来确定合适的函数 f.

在最大或然率分类中，目标完全指定联合分布 p(D,L)，这通常用贝叶斯定理来完成。决策树分类器(decision-tree classifier)将属性空间 D 划分为区域，然后为每个区域分配一个标号。神经网络(neural network)通过计算具有非线性边界的区域来概化决策树分类器。回归分析(regression analysis)以方程的形式来构建 D 与 L 之间相互作用的模型。

设有线性回归方程 $E[Y|X=x] = f(\alpha + \beta x)$, $X=(X_1、X_2、…、X_n)$，该表达式等价于更常见的 $Y=X\beta+\varepsilon$。则其对应的空间回归(SAR)方程为 $Y=\rho WY+X\beta+\varepsilon$。$W$ 是 X_i 的邻接矩阵。

5.2 关联规则(association rule)

关联规则是形如 X->Y 的模式，用概率论的属于来表述，关联规则 X->Y 是条件概率 P(Y|X)的一种表示。利用统计学原理分别计算关联规则的支持度与置信度，然后即可利用 Apriori 算法[11]找出关联规则。

Apriori 算法是发现关联规则领域的经典算法。该算法将发现关联规则的过程分为两个步骤：第一步通过迭代，检索出事务数据库中的所有频繁项集，即支持度不低于用户设定的阈值的项集；第二步利用频繁项集构造出满足用户最小置信度的规则。

从关联规则的生成扩展到空间关联规则有两种方案，一种是把项 X、Y 换成空间谓词 $P_1 \wedge P_2 \ldots P_n$ 与 $Q_1 \wedge Q_2 \ldots Q_m$，另一种是把事务集邻域化。

5.3 聚类(clustering)

聚类是一个在大型数据库中发现“群”或者“簇”的过程，它基于一种用于确定数据库中每对元组之间的关系的“相似性”。相似的元组放在一个组中，然后再标记这个组。聚类算法非常多，常分为以下四类：

1) **层次的(hierarchical)**：以所有模式作为单一聚类开始，然后连续执行分裂和合并，直到满足某个终止标准。例如：采用层次方法的平衡迭代归约和聚类(BIRCH)、采用代表点的聚类(CURE)以及采用链的健壮聚类(ROCK)。
2) **分区(partitional)**：以每个模式作为单一聚类开始，迭代地重新分配数据点到每个聚类，直到满足某个终止标准。例如： K-medoids 方法、围绕中心点的划分(PAM)、聚类大型应用(CLARA)、基于随机搜索的大型应用聚类(CLARANS)以及期望最大化(EM)
3) **基于密度的(density-based)**：基于某一区域中数据点的密度来尝试发现聚类。例如：带噪声的应用的基于的空间聚类(DBSCAN)和基于密度的聚类(DEBCLUE)。
4) **基于网格的(grid-based)**：首先将聚类空间离散化为有限数量的单元格，然后在离散的空间商执行所要求的操作。包含的点多于特定数量的单元格被认为是密集的，将密集的单元格连接起来形成聚类。例如：统计信息基于网格的方法(STING)、STING+、WaveCluster、BANG 聚类和 CLIQUE。

6 结论

数据挖掘是一个发展迅猛的领域，是数据库管理、统计学、人工智能等领域的交叉学科。数据挖掘提供半自动化的技术来挖掘海量数据中的未知模式。空间数据挖掘是数据挖掘中用于快速分析空间数据的一个很好的研究领域。在研究空间数据挖掘技术的过程中，应充分考虑空间自相关性的影响与空间数据特有的空间对象结构，不断推动数据挖掘技术向前发展。

参考文献

[1] S. Shekhar and S. Chawla, “Spatial Databases: A Tour”, *Prentice Hall, 2003*

[2] J. Han , M. Kamber and J. Pei, “Data Mining: Concepts and Techniques, Third Edition”, *Elsevier, 2011*

[3] J. F. Roddick and M. Spiliopoulou, “A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research”, *SIGKDD Explorations, Vol. 1, Num. 1, pp.34-38, June 1999*

[4] S. Shekhar, P. Zhang, Y. Huang and R. Raju Vatsavai, “Trends in Spatial Data Mining”, *http://www.spatial.cs.umn.edu/paper_ps/dmchap.pdf, 2008*

[5] J. Han, "OLAP Mining: An Integration of OLAP with Data Mining", *Proc. IFIP Conference on Data Semantics (DS-7), Switzerland, 1997*

[6] N. Andrienko and G. Andrienko, Knowledge Discovery Team (KD), “Intelligent visual data analysis service in the Internet”, *The 2nd International Conference on Web Information Systems Engineering, 2001*

[7] *http://www.esri.com/software/arcgis/index.html*

[8] MySQL 5.6 Reference Manual :: 11.17.2.1 The Geometry Class Hierarchy *http://dev.mysql.com/doc/refman/5.6/en/gis-geometry-class-hierarchy.html*

[9] M. Egenhofer, A. Frank, and J. Jackson, “A topological data model for spatial databases” *the Fisrt Symposium SSD '89, Santa. Barbara, California, USA, July 1989*

[10] W. R. Tobler, “*A Computer Movie Simulating Urban Growth in the Detroit Region*”, *Economic Geography, Vol. 46, pp. 234-240, 1970*

[11] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, *International Conference on Very Large Databases, VLDB, 1994*