空间数据挖掘研究综述*)

胡彩平　秦小麟
(南京航空航天大学信息科学与技术学院　南京 210016)

摘　要　信息化的发展使得更多的空间数据被使用，因此获取空间知识也就越来越重要和有意义，并使得空间数据挖掘成为一个很有前途的研究领域。本文系统概括了空间分类和预测、空间聚类、空间孤立点和空间关联规则4类空间数据挖掘方法及其进展，最后探讨了空间数据挖掘的未来发展方向。

关键词　空间数据挖掘，空间分类和预测，空间聚类，空间孤立点，空间关联规则

A Survey of Spatial Data Mining Research

HU Cai-Ping　QIN Xiao-Lin
(College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016)

Abstract　More and more spatial data are used with the development of the information, therefore, obtaining the spatial knowledge becomes more and more important and meaningful, this makes spatial data mining become a promising research filed. In this paper, the proceedings of four methods used in spatial data mining, namely spatial classification and prediction, spatial clustering, spatial outlier, spatial association rules are systematically summarized. Finally, the future directions of spatial data mining are discussed.

Keywords　Spatial data mining, Spatial classification and prediction, Spatial clustering, Spatial outlier, Spatial association rules

1　引言

空间数据挖掘(Spatial Data Mining，SDM)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等[1]。空间数据挖掘需要综合数据挖掘(Data Mining，DM)与空间数据库技术，可用于对空间数据的理解，空间关系和空间与非空间数据间关系的发现、空间知识库的构造、空间数据库的重组和空间查询的优化等[2]。空间数据挖掘在地理信息系统、遥感、图像数据库探测、医学图像处理、导航、交通控制、环境研究以及许多使用空间数据的领域中有广泛的应用。

空间数据与传统数据相比，具有自己独特的特点。表现在两个方面：一是空间数据都与某一对象(地点)相关，空间数据中除包含以文字、字符为特征的非空间信息(属性信息)外，还含有以拓扑关系、距离关系、方向关系为特征的空间信息；二是空间数据具有空间自相关性[3]，即每一个事物都与其它事物相关，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。这就使得空间数据挖掘比传统数据挖掘更为困难，因此研发高效的空间数据挖掘技术是当前空间数据挖掘面临的主要挑战。

1989年8月美国底特律市召开的第一届国际联合人工智能学术会议上，从事数据库、人工智能、数理统计和可视化等技术的学者们，首次提出从数据库中发现知识(Knowledge Discovery in Database，KDD)，标志着数据挖掘技术的诞生。空间数据挖掘在数据挖掘技术发展与海量空间数据积累的推动下，国内外都开展了积极的研究。加拿大西蒙法拉色大学计算机科学系的韩家炜教授领导的研究小组，较早对空间数据挖掘进行系统全面的研究[4~10]。1994年，在加拿大渥太华举行的GIS国际学术会议上，我国学者李德仁教授提出了从GIS数据库中发现知识的概念，系统分析了空间知识发现的特点和方法，认为它能够把GIS有限的数据变成无限的知识，并进一步用于精练和更新GIS数据，使GIS成为智能化的信息系统[11]。经过10多年的发展，国内外对空间数据挖掘已取得了一些成果，但空间数据挖掘的研究无论是在理论研究还是相关软件原型的研制等方面都还处于发展的初期。

2　空间数据挖掘方法

空间数据挖掘是计算机技术、数据库应用技术和管理决策支持技术等发展到一定阶段多学科交叉的新兴边缘学科，汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果[11]。一般地，空间数据挖掘可以分成4类挖掘方法[12]：空间分类和预测、空间聚类、空间孤立点和空间关联规则。

2.1　空间分类和预测

空间分类是指分析空间对象导出与一定空间特征有关的分类模式[12]，如地区、高速公路或河流的领域；预测是根据某空间维找出变化趋势。分类和预测具有很大的相似性，在数据挖掘界广泛接受的观点是[2]：用预测法预测类标号为分类，用预测法预测连续值为预测。Ester等人[13]最早提出了一种空间对象分类方法，该方法采用ID3算法，并使用邻域图的概念，分类标准基于分类对象的非空间属性以及描述分类对象与其邻近位置相关对象间空间关系的属性、谓词和函数。该

*)航空科学基金项目(02F52033)资助；江苏省高技术研究计划项目(BG2004005)资助。胡彩平　博士研究生，研究方向：空间数据挖掘等；秦小麟　教授，博士生导师，研究方向：空间数据库、空间数据挖掘和安全数据库等。

方法的缺点是没有分析邻近对象非空间属性的聚合值。另外,该算法也没有进行相关性分析,可能会生成低质量的决策树。而且,该算法没有考虑非空间和空间属性值中可能存在的概念层次。

顾及决策树邻近对象的非空间属性的聚合值,基于分类对象的非空间属性,描述被分类对象和邻近特征的空间关系的属性、谓词和函数,Koperski 和 Han 提出了空间数据的两步决策分类法[9]。在查找样本对象的粗略描述后,利用机器学习的 Relief 算法提取空间谓词,合并空间谓词和非空间谓词为分类决策知识。但是基于决策树的分类算法不适合处理带有不完整信息的问题。空间数据分类标准中包含数据间的空间关系,从某个训练数据集来讲,空间属性极有可能缺失。如果输入数据出现了不一致、噪声等情况,决策树算法可能会造成误分,就会严重影响决策树算法的预测准确度,因而采用决策树空间分类算法不能很好地体现地理空间关系对于分类的影响。石云等人提出的基于 Rough Set 的空间数据分类方法[14],采用 Rough Set 方法进行空间对象分类,能够较好地反映空间和非空间数据之间的关系,为利用邻近区域中基于非空间属性的聚合值来对空间对象进行分类提供了可行性,较好地解决了上述问题。

文[15]提出了用空间自相关模型和马尔科夫随机域两种方法进行空间分类和预测,并且从理论和实验两个方面对这两种方法进行了比较。文[16,17]提出了一个框架,使用图相似度去预测沼泽地中鸟巢的位置。文[18]针对径向基函数不适合用于空间数据预测的特点,提出了分别在输入层、中间层和输出层中加入空间信息来进行空间数据预测。

2.2 空间聚类

聚类分析是数据挖掘领域中的一项重要的研究课题。所谓聚类[2],就是根据相似性对数据对象进行分组,发现空间数据的分布特征,使得每一个聚类中的数据有非常高的相似性,而不同聚类中的数据尽可能不同。迄今为止,人们已经提出了许多聚类算法。一般地,主要的聚类算法可以分为 5 类[2]:

(1)划分方法(Partitioning Method)。给定一个 n 个对象或元组的数据库,给定要构建的划分的数目 $k(k<n)$,划分方法首先创建一个初始划分。然后采用一种迭代重定位技术,尝试通过对象在划分间移动来改进划分。一个划分方法构建数据的 k 个划分,每个划分表示一个聚类。典型的划分算法如 k-means 算法、k-medoids 算法和 CLARANS 算法等。

k-means 算法[19]以 k 为参数,把 n 个对象分为 k 个聚类,以使聚类内具有较高的相似度,而聚类间的相似度较低。相似度的计算根据一个聚类的平均值(被看作聚类的重心)来进行。但 k-means 算法对孤立点是敏感的。k-medoids[20]算法不采用聚类中对象的平均值作为参照点,选用聚类中位置最中心的对象,即中心点。仍然是基于最小化所有对象与其参照点之间的相异度之和的原则来执行的。

CLARANS 算法[21]由 Ng 和 Han 提出,其聚类过程可以表示为查找一个图,图中的每个节点都是潜在的解决方案。在替换一个中心点后获得的聚类称为当前聚类的邻居。随意测试的邻居的数目由参数 maxneighbor 限制。如果找到一个更好的邻居,将中心点移至邻居节点,重新开始上述过程,否则在当前的聚类中生成一个局部最优。找到一个局部最优后,再任意选择一个新的节点,重新寻找新的局部最优。局部最优的数目被参数 numlocal 限制。可以看到,CLARANS 算法并不搜索遍所有的求解空间,也不限制在任何具体的采样中。CLARANS 算法每次迭代的计算复杂度与对象的数量基本呈线性关系。基于 CLARANS 算法的空间数据聚类算法也有两种:空间支配算法和非空间支配算法。CLARANS 方法的缺点是要求欲聚类的对象必须预先都调入内存里,这对非常大的空间数据库是不合理的。

(2)层次的方法(Hierarchical Method)。层次的方法对给定数据对象集合进行层次的分解,它分为凝聚层次聚类与分裂层次聚类。凝聚层次聚类是自底向上的策略,首先将每一个对象作为一个聚类,然后合并它们,直到满足某个条件;分裂层次聚类正好相反,首先把所有的对象看作一个聚类,然后逐渐细分成越来越小的聚类,直至达到某个终结条件为止。著名的层次方法有 BIRCH 算法和 CURE 算法等。

Zhang 等人提出了平衡迭代消减聚类算法 BIRCH[22],它是一种较为灵活的增量式聚类方法,能根据内存的配置大小自动调整程序对内存的需要。它有两个重要概念:聚类特征(Clustering Feature)和聚类特征树(CF-tree),它们用于概括聚类描述。聚类特征(CF)是一个三元组,给出对象子聚类的信息的汇总描述。给定某个子聚类中有 N 个 d 维的点或对象$\{o_i\}$,则这个子聚类的聚类特征可表示为 $CF=(N,\vec{LS},SS)$。其中,N 是对象的个数,$\vec{LS}$是 N 个对象的线性和,即 $\vec{LS}=\sum_{i=1}^{N}\vec{O}_i$,它代表了这个子聚类的重心;$SS$ 是 N 个对象的平方和,即 $SS=\sum_{i=1}^{N}\vec{O}_i^2$,它代表了这个子聚类的直径大小,$SS$ 越小,这个子聚类聚得越紧。聚类特征树是一个满足两个条件的平衡树。两个条件分别是:分枝因子和子聚类直径的限制。分枝因子规定了树的每个节点的子女的最多个数;而子聚类直径体现了对一子聚类的直径大小的限制,即聚类特征的 SS 不能太大,否则不能聚为一类。非叶子节点上存储了它的子女的聚类特征的和,因此该节点总结了其子女的信息。

CURE 算法[23]选择基于质心和基于代表对象方法之间的中间策略。它不用单个质心或对象来代表一个子聚类,而是选择数据空间中固定数目的具有代表性的对象。

(3)基于密度的方法。其主要思想是:只要邻近区域的密度(对象的数目)超过某个阈值,就继续聚类。代表性算法有 DBSCAN 算法、OPTICS 算法、GDBSCAN 算法、DBRS 算法和 DENCLUE 算法等。

DBSCAN 算法[24]是第一个基于密度的空间聚类算法,用来发现带有噪声的空间数据库中任意形状的聚类。该算法的效率较高,但算法执行前需输入阈值参数。为了解决这个难题,提出了 OPTICS 算法[25]。它没有显示地产生一个子聚类的集合,它为自动和交互的聚类分析计算一个聚类次序,这个次序代表了数据的基于密度的聚类结构。它包含的信息等同于从一个宽广的参数设置范围所获得的基于密度的聚类。GDBSCAN 算法[26]是 DBSCAN 算法的推广,它不仅能对点进行聚类,也能对线或多边形进行聚类。但 DBSCAN 算法仅能对点进行聚类,在现实生活中,要聚类的空间对象都用点来抽象有时不可行。DBSCAN 算法进行聚类的过程就是一个不断执行区域查询的过程,聚类过程的大部分时间都用在区域查询操作上。DBSCAN 算法对核心对象的邻域中包含的所有对象都执行区域查询来扩展聚类,显然没有必要。DBRS 算法[27]仅选择核心对象邻域中的部分代表对象,而不是像 DBSCAN 算法那样选择所有对象,作为种子对象用于聚类的扩展。这样就可以减少区域查询的次数,提高查询效率。另

外，DBSCAN算法没有考虑非空间属性，DBRS算法不仅考虑了空间属性，也考虑了非空间属性。DENCLUE算法[28]是一个基于一组密度分布函数的聚类算法。

(4)基于网格的方法。这种方法采用一个多分辨率的网格数据结构将空间分成有限数目的单元，聚类操作在单元内进行。这种方法处理的时间独立于对象的数目，处理速度快。著名的有STING算法、WaveCluster算法和CLIQUE算法等。

STING算法[29]是一种基于网格的多分辨率聚类技术，它将空间区域划分为矩形单元。针对不同级别的分辨率，通常存在多个级别的矩形单元，这些单元形成了一个层次结构：高层的每个单元被划分为多个低一层的单元。关于每个网格单元属性的统计信息(例如平均值、最大值和最小值)被预先计算和存储。WaveCluster算法[30]是一种多分辨率的聚类算法，它首先通过在数据空间上强加一个多维网格结构来汇总数据，然后采用一种小波变换原特征空间，在变换后的空间中找到密集区域。CLIQUE算法[31]综合了基于密度和基于网格的聚类方法，它对于大型数据库中的高维数据的聚类非常有效。

(5)基于模型的方法。这种方法为每个聚类假设一个数学模型，试图为数据查找合适的数学模型。主要有两类方法：统计的方法与神经网络方法。统计的方法如AutoClass[32]等；神经网络方法如Rumelhart和Zipser提出的竞争学习法[33]等。

但以上的这些算法并没有考虑现实空间中可能存在的约束，如河流、湖泊、山脉等障碍物，以及桥梁、隧道等连接设施。由于障碍物的存在，有些聚类必须进一步划分；又由于连接设施的存在，有些聚类必须合并。

Tung等人最早提出了一种在空间数据挖掘中实行空间聚类时，处理河流、高速公路等阻隔的COD_CLARANS算法[10]。COD_CLARANS算法是在CLARANS算法基础上的改进，主要的思想是用两个点间阻隔距离(obstructed distance)代替欧氏距离。AUTOCLUST＋算法[34]是基于Voronoi图和Delaunay三角剖分基础上的空间障碍聚类算法，它是AUTOCLUST算法[35]的改进版，其优点是不需要用户输入参数。Zaïane O. R.等人第一次提出了能够处理障碍物和连接设施的空间聚类DBCLuC算法[36]，它来源于DBSCAN算法。Wang Xin等人提出了基于DBRS算法的空间障碍聚类DBRS＋算法[37]，该算法也能够处理障碍物和连接设施，而且在这4个能够处理空间约束的聚类算法中它的效率最高。

2.3 空间孤立点

孤立点就是在数据集中与其它数据点表现不一致的对象，或者大大地偏离其它数据点以致于怀疑它是由不同机制生成的对象。现有的发现孤立点的方法大多建立在统计学的基础上，大致可以分为4类[38]：基于分布的、基于深度的、基于距离的和基于密度的。

(1)基于分布的方法在统计领域较为常见。人们用各种统计模型来测试，把偏离这些模型的对象当作孤立点[39]。但是，大多数分布模型只能直接应用于单变量的特征空间，难以应用于多维空间。而且，这种模型要求预先知道数据的分布，但这种知识往往难以获得，这就要进行耗时的测试来决定。

(2)基于深度的方法[40]是一种基于计算几何的方法，它计算不同层面的k-凸面来查找孤立点，凸面的外层被认为是孤立点。这个算法对二维和三维空间上的数据比较有效，但对四维及四维以上的数据，处理效率比较低。

(3)Knorr与Ng[41]提出了一个基于距离的孤立点概念。在数据库中$p\%$的对象与某对象的距离超过d，这个对象为孤立点。文[42]对基于距离的孤立点的概念进行了扩展，根据k-最近邻距离对孤立点进行排序，并给出了一种有效的方法，计算排在最前面的n个孤立点。

(4)Breunig等[43]提出了局部孤立因子的概念LOF，这是一种基于密度的方法。通过数据空间的所有维度来计算对象的距离，进而计算对象的可达密度，最后通过局部的孤立度来判断孤立点。

上述孤立点查找方法不是专门针对空间数据的，它们没有区分空间维与非空间维，忽略了空间信息。若直接应用在空间数据中，其结果可能导致发现的孤立点没有实际意义，或产生错误的判断。空间孤立点是指空间领域中非空间属性与其他对象明显不同的空间对象。空间领域可以基于空间属性(如位置)使用空间关系(如距离或邻接)进行定义。检测空间孤立点在地理信息系统和空间数据库的很多应用中非常有用，这些应用领域包括交通、生态、公共安全、公众健康、气候和基于位置的服务等。

下面我们来讨论专门针对空间孤立点的查找方法。在相关的空间统计文献中提供了两种测试，即图形测试和定量测试。图形测试是用可视化的方法来查找空间孤立点，如变差云图[44]与Moran散点图[45]；定量方法提供了一个精确测试，将空间孤立点与其它数据区别开来，如散点图[46]。这些方法的缺点是：①只适合一维变量，不适合多维空间；②变量云要求大量的后处理，因而造成查找效率低；③基于图形的方法没有确定的标准来区分孤立点等等。Shekhar等[47~49]提出了一个新的空间孤立点查找方法，用空间属性来定义空间领域，用非空间属性来识别空间孤立点。这种方法的主要思想是：假设空间对象o的非空间属性$f(o)$服从正态分布，定义空间统计量$Z_s(o)=\left|\frac{S(o)-\mu_s}{\sigma_s}\right|$。对于非空间属性为$f(o)$的每个空间对象$o$，$S(o)$是空间对象$o$的非空间属性值与$o$的领域对象的平均非空间属性值之差，即$S(o)=f(o)-\frac{1}{N(o)}\sum_{p\in N(o)}f(p)$，$\mu_s$是$S(o)$的平均值，$\sigma_s$是所有空间对象的$S(o)$的标准差，$N(o)$是$o$的空间邻域。如果$\left|\frac{S(o)-\mu_s}{\sigma_s}\right|>\theta$($\theta$的选择和制定的置信度有关)，则空间对象$o$为空间孤立点。但该方法只能查找单属性的空间孤立点。Chang-Tien Lu等人[50]将此方法进行了拓展，可以查找多属性的空间孤立点。但这些方法的缺点是使用统计测试仅适合查找全局孤立点，不适合局部孤立点，但有时候局部孤立点更有意义。Sanjay Chawla和Pei Sun[51]提出了一个查找局部空间孤立点的算法并且给每个空间对象o定义了空间孤立点因子$SLOM(o)$的概念，该因子定义了空间对象的孤立程度。同时，一个空间对象的孤立程度与它邻域中的空间对象有关，这体现了“局部”的概念。

2.4 空间关联规则

Agrawal等人引入关联规则的概念，是为了挖掘大型的事务型数据库。Koperski等人将这个概念扩展至空间数据库[7]，提出了一种在空间数据库中挖掘强空间关联规则的算法，并给出了两步式的空间优化技术。空间关联规则是根据空间谓词而不是根据项来定义的。一个空间关联规则可表示为[12]：

$P_1 \wedge P_2 \wedge \cdots \wedge P_n \rightarrow Q_1 \wedge Q_2 \wedge \cdots \wedge Q_m (c\%, s\%, i\%)$

其中，P_i 和 Q_j 中至少有一个是空间谓词，$c\%$、$s\%$和 $i\%$分别是规则的可信度、支持度和兴趣度。空间谓词有 3 种形式：表示拓扑关系的谓词，如相交、覆盖等；表示空间方向的谓词，如东、西、左、右等；表示距离的谓词，如接近、远离等。Lizhen Wang 等人[52]提出了利用划分来挖掘多层空间关联规则的算法，并且引入了等价划分树的概念，使得挖掘多层空间关联规则既容易又有效。L. K. Sharma 等人[53]提出了挖掘多层空间正负关联规则的算法。Annalisa Appice 等人[54]利用在归纳逻辑规划(Inductive Logic Programming，ILP)领域中的表达和推理技术，提出了在人口普查数据库中挖掘空间关联规则的算法。刘君强等人[55]设计了一种基于前缀树的单层布尔型关联规则挖掘算法(FPT-Generate)，优点是不需要反复扫描数据库，不产生候选模式集。由于技术手段的限制或人为的误差，空间数据中存在大量的不确定性和模糊性，研究基于不确定性和模糊性的空间数据挖掘技术就显得尤为迫切。针对空间数据的这个特点，人们又提出了一些挖掘空间关联规则的算法。Clementin 等人[56]提出了在宽边界的空间实体中挖掘多层次的空间关联规则的算法。刘大有等人[57]以定性空间推理的 RCC 理论为基础，结合模糊逻辑，提出了一种面向空间数据库的近似区域空间关系模型，在此基础上给出了多层空间关联规则的挖掘算法 QSRSAR。Esen Kacar 等人[58]提出了挖掘空间模糊关联规则的方法。

S. Shekhar 等人[59]提出了基于空间相关的同位模式，它把事务概念泛化，以包括邻域集合，并且将关联规则的概念泛化为同位规则，在获取同位模式时，很好地考虑了空间相关性。文[60]在文[59]的基础上，提出了不需要支持度剪枝，挖掘可信空间同位规则的方法。文[61]提出了在扩展的空间对象(例如线和多边形)上挖掘空间同位规则的框架。Jin Soung Yoo 在对文[59]中的方法挖掘空间同位规则的时间效率上进行了改进，首先提出了基于半连接操作挖掘空间同位规则的算法[62]，然后又提出了基于更少连接操作的空间同位规则挖掘方法[63]，该方法首先要确定每一个空间特征的星形领域，在星形领域的基础上只要用很少的连接操作就可以挖掘出空间同位规则，运算速度有了很大的提高。空间同位规则与空间关联规则既有联系又有区别，联系表现在它们都可以在地理空间中产生 if-then 规则，区别表现在空间关联规则是发现与一个特定的空间特征相关的空间特征的子集，表现为星形模式；而空间同位规则没有一个特定的空间特征，表现为团形模式[64]。

另外，空间统计学(Spatial Statistics)、神经网络(Neural Network)、证据理论(Evidence Theory)、模糊集(Fuzzy Sets)、粗糙集(Rough Sets)和遗传算法(Genetic Algorithm)等这些传统的数据挖掘方法经过修改，加入空间信息也可用于空间数据挖掘。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。

3 空间数据挖掘的未来发展方向

空间数据挖掘是一个非常年轻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。主要表现在：

(1) 如何扩展传统的数据挖掘技术进行空间数据挖掘

经过 10 多年的发展，数据挖掘领域取得了大量的研究成果。但因为空间数据与传统数据的区别，传统数据挖掘技术不能直接应用于空间数据挖掘，必须对传统数据挖掘技术进行修改，可以通过加入空间信息来实现。例如 EM 算法[65]是一个十分著名的划分聚类算法，但它并不能直接用于空间聚类。通过对它的修改，在它里面加入空间惩罚因子，引入 NEM 算法[66]，就可以进行空间聚类。

(2) 基于不确定性和模糊性的空间数据挖掘

由于技术手段的限制或人为的误差，空间数据中存在大量的不确定性和模糊性，研究基于不确定性和模糊性的空间数据挖掘技术就显得尤为迫切。空间统计学、证据理论、模糊集、粗糙集和云理论等方法是处理不确定性的很好方法，把这些方法应用于空间数据挖掘领域有待进一步拓展。

(3) 加入时间维的时空数据挖掘

空间数据挖掘实际上是用静止的观点来看待空间现象。然而，空间现象是随着时间的改变而改变的。因此，时空数据挖掘才能使我们更好地理解空间现象。可以这么说，空间数据挖掘的必然趋势就是时空数据挖掘。

(4) 有约束条件的空间聚类技术

现有大量的聚类算法在聚类时并没有考虑现实空间中可能存在的约束，如河流、湖泊、山脉等障碍物，以及桥梁、隧道等连接设施。研究有约束条件的空间聚类技术更有现实意义。

(5) 栅格矢量一体化数据挖掘

空间数据结构是空间信息管理系统的基础。空间数据库中的数据结构主要有两种：基于栅格的数据结构和矢量的数据结构。研究栅格矢量一体化数据挖掘方法，就能够很方便地在这两种数据结构中进行数据挖掘。

(6) 线形或多边形聚类技术

目前空间聚类问题的解决方案尚局限在对点对象的聚类，该问题的未来方向是处理可扩展的对象的聚类，如线形或多边形聚类。

(7) 空间数据挖掘查询语言

面向空间数据挖掘的查询语言还很不完善，研究更有效、更方便的空间数据挖掘的查询语言也是一项很迫切的任务。

(8) 知识的可视化表示

理解所发现知识的最有效的方式是进行图形可视化。可视化仍是一个不成熟的领域，有待进一步研究。

(9) 面向对象(Object-Oriented，OO)的空间数据库中的知识挖掘

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，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。

参 考 文 献

1 Lu W, Han J, et al. Discovery of general knowledge in large spatial databases. In: Proc. Far East Workshop on Geographic Information Systems. Singapore, 1993. 275～289

2 Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明，孟小峰，等译. 北京：机械工业出版社，2001

3 Cressie N. In statistics for spatial data. Wiley-Interscience, 1993

4 Koperski K, Han J, Adhikary J. Mining Knowledge in Geographical Data[J]. IEEE Transaction on Knowledge and Data Engineering, 1993, 10: 903～913

5 Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining. In: The 20th Very Large Databases Conference, Santiago, Chile, 1994

6 Ester M, Kriegel H P, et al. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: Advances in Spatial Databases, Proc. of 4th Symp SSD'95, Berlin: Springer-Verlag, 1995. 67～82

7 Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases. In: Proceedings of the 4th International Symposium on Large Spatial Databases(SSD95), Maine, 1995. 47～66

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

9 Koperski K, Han J, et al. An efficient two-step method for classification of spatial data. In: Proc. Int'l Sympon Spatial Data Handling SDH'98, Vancouver, BC, Canada, 1998

10 Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles. IEEE Transactions on Data Engineering, 2001, 11: 359～369

11 Li D R, Cheng T. KDG-Knowledge Discovery from GIS. In: Proceedings of the Canadian Conference on GIS, Ottawa, 1994

12 Shashi S, Chawla S. 空间数据库[M]. 谢昆青, 马修军, 杨冬青, 等译. 北京: 机械工业出版社, 2004

13 Ester M, Kriegel H P, Sander J. Spatial Data Mining: a Database Approach. In: Scholl M V, ed. Proceedings of the 5th International Symposium on Spatial Databases (SSD. 97). Berlin: Springer_Verlag, 1997

14 石云, 孙玉芳, 左春. 基于 RoughSet 的空间数据分类方法. 软件学报, 2000, 11(5): 673～678

15 Shekhar S, Schrater P, Vatsavai R, et al. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. IEEE Transactions on Multimedia (special issue on Multimedia Databases), 2002

16 Chawla S, Shekhar S, Wu W. Predicting Locations Using Map Similarity (PLUMS): A Framework for Spatial Data Mining. In: Proc. of the 6th International Conference on Knowledge Discovery and Data Mining, Boston, MA, 2000

17 Chawla S, Shekhar S, Wu Weili, et al. Modeling Spatial Dependencies for Mining Geospatial Data. In: 1st SIAM International Conference on Data Mining, 2001

18 Hu Tianming, Sung Sam Yuan. Data Fusion in Radial Basis Function Networks for Spatial Regression. Neural Processing Letters, 2005, 21(2): 81～93

19 MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symposium in Mathematics, Univ of California, Berkeley, USA, 1967

20 Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to cluster Analysis. New York: John Wiley & Sons, 1990

21 Ng R, Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Trans Knowledge & Data Engineering, 2002, 14(5): 1003～1016

22 Zhang T, et al. BIRCH: An efficient data clustering method for very large databases. In: Proc. of ACM-SIGMOD Int'l Conf. on Management of Data. ACM, NewYork, 1996. 103～114

23 Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm ofr large databases. In: Proc. 1998 ACM-SIGMOD Int Conf. Management of Data (SIGMOD'98), Seattle, Washington, 1998. 73～84

24 Ester M, Kriegel H, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. of 2nd KDD, Portland, 1996. 226～231

25 Ankerst M, Breunig M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure. In: Proc. 1999 ACM-SIGMOD Int Conf Management of Data (DIGMOD'99), Philadelphia, PA, 1999. 49～60

26 Sander J, Ester M, Kriegel H -P, et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. Data Mining and Knowledge Discovery, 1998, 2(2): 169～194

27 Wang Xin, Hamilton H J. DBRS: A Density-Based Spatial Clustering Method with Random Sampling. In: Proc. of the 7th PAKDD, Seoul, Korea, 2003. 563～575

28 Hinneburg A, Keim D A. An efficient approach to clustering in large multimedia databases with noise. In: Proc. 1998 Int Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 1998

29 Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining. In: Proc. 1997 Int Conf. Very Large Data Bases (VLDB'97), Athens, Greece, 1997. 186～195

30 Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In: the 24th International Conference on Very Large Data Bases, New York City, 1998. 428～439

31 Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dmensional data for data mining applications. In: Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), Settle WA, 1998. 94～105

32 Cheeseman P, Stutz J. Bayesian classification (AutoClass): Theory and results. In: Advances in Knowledge Discovery and Data Mining, Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al., eds. Cambridge, MA: AAAI/MIT Press, 1996. 153～180

33 Rumelhart D E, Zipser D. Feature discovery by competitive learning. Cognitive Science, 1985, 9: 75～112

34 Estivill-Castro V, Lee I J. AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles. In: Proc. of Intl Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lyon, France, 2000. 133～146

25 Estivill-Castro V, Lee I J. AUTOCLUST: Automatic Clustering via Boundary Extraction for Mining Massive Point-Data Sets. In: Proc. of the 5th Intl Conf On Geocomputation, 2000. 23～25

36 Zalane O R, Lee C H. Clustering Spatial Data When Facing Physical Constraints. In: Proc. of the IEEE International Conf. on Data Mining, Maebashi City, Japan, 2002. 737～740

37 Wang Xin, Rostoker C, Hamilton H J. Density-based Spatial Clustering in the Presence of Obstacles and Facilitators. In: PKDD, 2004. 446～458

38 魏藜, 宫学庆, 钱卫宁, 等. 高维空间中的离群点发现. 软件学报, 2002, 13(2): 280～290

39 Barnett V, Lewis T. Outliers in Statistical Data. New York: John Wiley and Sons, Inc, 1994

40 Johnson T, Kwok I, Ng R T. Fast computation of 2-dimensional depth contours. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York City, New York, USA, 1998. 224～228

41 Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets. In: the 24th International Conference on Very Large Data Bases, New York, 1998. 392～403

42 Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, 2000. 93～104

43 Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers. In: the ACM SIGMOD International Conference on Management of Data, Dallas, Texas, 2000. 427～438

44 Haslett J, Brandley R, Craig P, et al. Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies. The American Statistician, 1991(45): 234～242

45 Anselin L. Local Indicators of Spatial Association: LISA. Geographical Analysis, 1995, 27(2): 93～115

46 Anselin L. Exploratory Spatial Data Analysis and Geographic Information Systems. In: Painho M, ed. New Tools for Spatial Analysis, 1994. 45～54

47 Shekhar S, Lu C T, Zhang P. A Unified Approach to Detecting Spatial Outliers Geoinformatica, 2003, 7(2): 139～166

48 Shekhar S, Lu C T, Zhang P. Detecting Graph-Based Spatial Outlier: Algorithms and Applications. In: Proc. of the Seventh ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2001

49 Lu C T, Chen D, Kou Y. Algorithms for spatial outlier detection. In: the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourne, Florida, USA, 2003. 596～600

50 Lu CT, Chen DC, Kou YF. Detecting spatial outliers with multiple attributes. In: Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), Sacra-

mento, California, 2003. 122～128
51 Chawla S, Sun P. SLOM: a new measure for local spatial outliers. Knowledge and Information Systems, 2006, 9(4): 412～429
52 Wang Lizhen, Xie Kunqing, Chen Tao, et al. Efficient discovery of multilevel spatial association rules using partitions. Information and Software Technology, 2005, 47: 829～840
53 Sharma L K, Vyas O P, Tiwary U S, et al. A Novel Approach of Multilevel Positive and Negative Association Rule Mining for Spatial Databases, MLDM, 2005. 620～629
54 Malerba D, Esposito F, Lisi F A, et al. Mining spatial association rules in census data: A relational approach. Research in Official Statistics, 2002, 5(1): 19～44
55 刘君强，潘云鹤. 挖掘空间关联规则的前缀树算法设计与实现. 中国图象图形学报, 2003, 8A(4): 476～480
56 Clementini E, Felice P D, Koperski K. Mining Multiple level Spatial Association Rules for Objects with a Broad Boundary. Data and Knowledge Engineering, 2000, 34: 251～270
57 刘大有，王生生，虞强源，等. 基于定性空间推理的多层空间关联规则挖掘算法. 计算机研究与发展, 2004, 41(4): 565～570
58 Kacar E, Cicekli N K. Discovery Fuzzy Spatial Association Rules, Data Mining and Knowledge Discovery: Theory, Tools and Technology IV. In: Dasarathy B V, ed. Proceedings of SPIE, Vol4730, 2002. 94～102
59 Shashi S, Yan Huang. Discovering spatial co-location patterns: A summary of results[A]. In: Proceedings of the 7th International Symposium on Spatial and Temporal Databases, Redondo Beach, CA, 2001. 236～256
60 Huang Y, Xiong H, Shekhar S, et al. Mining Confident Co-location Rules without A Support Threshold. In: Proc. 2003 ACM Symposium on Applied Computing, New York, NY, USA, 2003. 497～501
61 Xiong H, Shekhar S, Huang Y, et al. A Framework for discovering co-location patterns in datasets with extended spatial objects[A]. In: Berry M W, Dayal U, Kamath C, et al., eds. Proceedings of the Fourth SIAM International Conference on Data Mining[C]. Florida, USA, 2004. 78～89
62 Yoo Jin Soung, Shekhar S. A partial join approach for mining co-location patterns[A]. In: Foser D P, Cruz I F, Ronthaler M, eds. 12th ACM International Workshop on Geographic Information Systems[C]. Washington, DC, USA, 2004. 241～249
63 Yoo Jin Soung, Shekhar S, Celik M. A Join-less Approach for Co-location Pattern Mining: A Summary of Results. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), Houston, USA, 2005
64 Zhang X, Mamoulis N, Cheung D, et al. Fast Mining of Spatial Collocations. In: Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004
65 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, 1977, B(39): 1～38
66 Ambroise C, Govaert G. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters, 1998, 19(10): 919～927

(上接第 3 页)

不同部分的开发，但是还没有一个完整的用于语义桌面开发的集成环境。表 1 对这些工具进行了分类讨论。

表 1 语义桌面开发工具列表

类型		现有的工具	语义桌面工具
基础技术	存储	● Jena ● Sesame ● RDF Gateway	storage supports SPARQL and semantic protocols
	搜索	● Lucene ● Desktop Search Tools	semantic search services
	通讯	● Jabber,IM ● email ● P2P networks	semantic messaging and P2P
本体		● DC ● SKOS ● FOAF ● Thesauri ● iCalendar ● PIM	● popular ontologies ● ontology mapping tools ● desktop ontologies
信息交互	本体编辑	● Protege ● IsaViz ● KAON	ontology editors present in all applications
	专用工具	● Tidepool/Storymill ● PhotoStuff ● RSS Readers	Semantic Desktop Applications
	PIM 和工作流	● Microsoft Outlook ● Lotus Notes ● Frodo Taskman	Semantic PIM,Semantic Work-flow

我们将这些工具分为基础技术、本体和信息交互 3 个大的类别，每个类别又根据功能划分为若干小类。我们分别列出了每个类别现有的开发工具和将来具有语义功能的高级开发工具。

小结和展望 Stefan Decker 将语义桌面的发展划分为 3 个阶段。从以往的情形来看，我们已经解决了第一个阶段中的绝大部分问题，现在正处于第二个阶段。我们现在的工作就是将现有的那些成熟的语义 Web、P2P 和社会化网络技术结合到一起。本文仅针对其中语义桌面这一种结合方式。语义桌面作为一个目标工程涉及了许多不同领域的工作。

同时，我们认为，必须建立标准化的应用程序编程接口，并且提供一个背景框架，用于支持可能的服务。我们还介绍了现有软件的用户界面和体系结构，并提取了它们的设计模式，从而给予开发人员更多的启示。由 DFKI 领导的 NEPOMUK 项目即将完成，它将帮助开发和使用语义桌面的专家建立一个交流的平台。这个项目的一部分将成为标准化接口的免费资源，提供给开发人员一些应用程序实例，也可以为终端用户提供一些实用的语义桌面应用程序。

总之，语义桌面可以将语义 Web 与工作于个人计算机上的用户连接起来。它让人们可以随时记录下自己的思想和知识，并且将这些信息与别人分享。

参 考 文 献

1 Berners-Lee, Hendler J, Lassila O. The semantic web: Scientific American, 2001, 89
2 Sauermann L. The gnowsis-using semantic web technologies to build a semantic desktop: [Diploma thesis]. Technical University of Vienna, 2003
3 Decker S, Frank M. The social semantic desktop: WWW 2004 Workshop Application Design. Development and Implementation Issues in the Semantic Web, 2004
4 Sauermann L. The semantic desktop: a basis for personal knowledge management. In: The 5th International Conference on Knowledge Management, 2005. 294～301
5 Sauermann L. Gnowsis adapter framework- Treating structured data sources as virtual RDF graphs. In: Proc. of the ISWC2005, 2005
6 Corp, Microsoft. Information bridge framework. http://msdn.microsoft.com/office/understanding/ibframework/
7 Quan D, Huynh D, Karger D R. Haystack: A platform for authoring end user semantic web applications. In: International Semantic Web Conference, 2003. 738～753
8 AG B. The brainfiler text classification system. http://www.brainbot.de
9 Maus H, Holz H, Bernardi A, Rostanin O. Leveraging passive paper piles to active objects in personal knowledge spaces. In: Proceedings of 3rd Conference Professional Knowledge Management: Experiences and Visions, 2005. 43～46
10 Clark H. Using language. Cambridge University Press, 1996
11 Schwarz S. A context model for personal knowledge management. In: Proc. of the IJCAII'05 Workshop on Modeling and Retrieval of Context, Edinburgh, 2005
12 Chirita P A, Gavriloaie R, Ghita S, Nejdl W, Paiu R. Activity based metadata for semantic desktop search, 2004
13 Barreau D, Nardi B A. Finding and reminding. File organization from the desktop, 1995