DOI:10.13203/j.whugis2002.04.014
第27卷第4期 武汉大学学报·信息科学版 Vol.27 No.4
2002年8月 Geomatics and Information Science of Wuhan University Aug. 2002

文章编号:1000-050X(2002)04-0397-06 文献标识码:A

知识的综合发现:理论、概念及应用

沙宗尧[1] 边馥苓[1] 陈江平[1]

(1 武汉大学遥感信息工程学院,武汉市珞喻路129号,430079)

摘 要:提出了知识的综合发现思想,重点以空间对象关联中的相邻关系与空间特征属性为知识综合发现的研究对象,对相关问题进行了讨论,并提出了一个高效的知识综合发现算法。实例结果表明,本算法是高效的,发现的知识是有效、可理解的。

关键词:知识的综合发现;空间关联;空间综合信息表;知识发现算法

中图法分类号:P208;TP18

在当前空间知识挖掘领域,存在着两种倾向:①发展对交易型数据库即属性数据库的规则挖掘,忽略了现实世界80%以上的信息都是空间相关的事实;②片面地追求空间数据挖掘,而忽视了空间数据与属性数据是相辅相成地对事物特征进行更深入描述的事实。本文研究了空间数据的空间关联(相邻)规则和属性信息综合挖掘的相关理论、概念、挖掘算法及其应用,为充分利用现有的数据资源提供有效途径。

1 基本理论及相关概念

1.1 空间知识表达系统

借鉴知识表达系统的定义[3],定义5元组$S=(U, C, D, V, f)$,其中$U=\{u_1, u_2, \cdots, u_n\}$是对象的有限集合,$A=C\cup U$为属性集合,$C=\{a_1, a_2, \cdots, a_m\}$是条件属性的有限集(在条件属性中包含了空间制约条件),$D=\{d_1, d_2, \cdots, d_x\}$是决策属性的有限集,$V$是属性$C\cup U$所构成的域,即$V=\bigcup_{p\in A}V_p$,$V_p$是属性$p$的域,$f$为一个信息函数,即$f: U\times A\rightarrow V$。$S$即为空间知识表达系统的形式化定义。

空间知识发现是从空间数据库中发现隐含的、为人们感兴趣的模式[4]。由于空间数据库中不仅存储着空间对象的属性数据、几何数据,而且存储着空间对象之间的空间关系(拓扑关系、度量关系、方位关系等),在属性制约中增加了空间约束条件,因此,相对于事务数据库,空间知识发现具有更大的挑战性。目前空间知识发现的研究主要集中在数据挖掘基本理论、优化算法及应用上。

1.2 相关概念

1) 空间目标分类。

2) 空间关系。空间关系是GIS理论研究的一个重要领域,GIS最终功能将体现在空间分析上,空间关系是空间分析的基础[4]。

3) 知识的综合发现。

4) 综合知识(comprehensive knowledge)。

5) 空间综合信息表(spatial union information table,SUIT)。空间信息应当是空间图形信息、拓扑信息和属性信息的总称,空间综合信息表定义为包含了空间图形信息、拓扑信息以及属性信息的数据表[5]。该表的结构分为两个部分:前一部分为非结构的空间关系信息(spatial relations,SR),记录空间实体对象的分类及其关系;后一部分为属性信息(AI),为空间实体的属性集。SUIT可形式化地表示为五元组$SUIT(T, SR, SRV, AI, AIV)$,其中$T$为目标空间的全集,$SRV$和$AIV$分别为空间对象间的空间关系相对测度值(或表示模式)和特征属性值,空间关系相对测度值表示方法与讨论的具体空间关系有关。通过对SUIT的处理分析,就可以发掘或获取研究对象可能存在的完整信息(包括空间信息泛化、空间关联规则、空间分类与聚类等)。但是在实际研究中,为了特定的目的或因为条件的限制,不可能列

收稿日期:2002-04-08。

项目来源:国家测绘局测绘科技发展基金资助项目(99009、20007);湖北省自然科学基金资助项目(99J171)。

出所有因素,因而所有的研究都基于一定的简化。设SUIT′为实际待研究的问题域,T'、SR'、AI'分别为目标空间的子集、目标空间关系(相邻)测度值的子集以及属性子集,即 $T' \subseteq T$, $SR' \subseteq SR$, $AI' \subseteq AI$。如图 1 所示,空间实体对象的子集为 $T'=\{A_1, B_1, C_1, D_1, A_2\}$,特征对象的分类子集(特征类)为 $SR'=\{A, B, C, D\}$, $A_1 \in A$, $A_2 \in A$, $B_1 \in B$, $C_1 \in C$, $D_1 \in D$。假设特征属性集为 $AI'=\{$Area(特征面积), Peri(特征周长)$\}$,图中的SUIT′可以用表 1 表示,其中空间关系测度值是经过归一化处理的测度值,其数值的相对大小表示了该空间特征与其他特征类的相邻关系程度。数值"1"指空间特征与其本身特征类的绝对相邻或该多边形被特征类多边形所包围;"0"表示空间特征与该特征类不相邻或相邻度极小,以至可以忽略;其他数值表示相邻度介于两者之间。

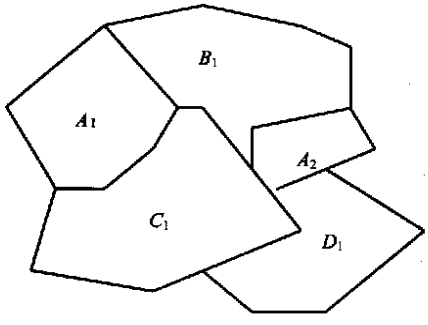

图 1 空间实体对象关系示意

Fig. 1 Relationship Between Spatial Objects

表 1 空间实体对象的空间联合信息表

Tab. 1 SUIT for Spatial Objects

T	A	B	C	D	Area	Peri
A_1	1	0.5	0.4	0	1.95	2.74
B	0.8	1	0.3	0	1.90	2.88
C	0.9	0.3	1	0.5	2.13	3.11
D	0.5	0	0.7	1	1.89	2.69
A_2	1	0.3	0.5	0.5	0.84	1.39

2 空间关联关系

2.1 关联规则及空间关联规则

关联规则可以用数学模型加以描述[2],但关联规则是数据表现的宏观方式,具有统计学意义;关联关系指的是微观实体间的一种特定关系,不具有全局的统计意义。空间关联规则的研究是对自然世界空间信息深层次的发掘,它可以形式化地表示为:对于空间实体类型 A、B 以及实体全集 U, R 为空间关系谓词,如果 A 与 B 具有关联关系 R,则 ARB。例如,如果河流的两旁 5km 范围内分布了 80%的农田,则农田与河流是空间关联的,空间数据库与事务数据库的实体关联关系的发现有很大的差别。首先,空间关联关系具有更大的隐蔽性。空间关联关系一般隐藏在空间数据库中,与事务数据库中的关联关系相比,空间关联关系在空间数据库中一般并没有被显式地表示,而需要数据挖掘人员在深入理解空间数据的语义和空间数据组织的前提下对数据进行预处理,因而规则发现的实施人员必须清楚空间数据结构。其次,空间关联关系具有模糊性质。事务数据库中的两个数据项间的关联关系是 0-1 关系,即不关联或关联,不存在中间状态;而空间数据库中,在对空间实体进行分类之后,对于属于某类的一个空间实体,与之具有关联关系的多个不同的空间实体中有部分归属于同一个特征类,这些归属于同一特征类的部分特征共同决定该空间实体的关联关系,在这种情况下,空间关联关系模糊化了。

2.2 空间相邻关系的关联性测度

空间关系的内容广泛,一般来说,空间实体的类型和数量甚多,为此,笔者以空间相邻关系为重点来研究知识综合发现中的关联性测度问题。

按照 Gold 的定义,如果两个空间目标具有公共 Voronoi 边,则认为它们具有空间相邻关系[7]。但是该定义仅表达了空间相邻的定性关系,为了定量地表示空间相邻关系,用一种简单的方法来近似地定量表示面状实体的空间邻近关系,从而反映空间关联(相邻)规则的强度指标。如图 2 所示,多边形 1 与多边形 2 具有相邻线 AB。定义相邻度 N_q 为空间相邻关系的测度标识,对于简单多边形,N_q 具有与邻接线长度成正比,而与特征间的重心距成反比的性质。图 2 所示的多边形的重心点分别为 O_1、O_2,相邻线 AB 的长度为 l_{AB},则 $N_q = l_{AB} / lo_1o_2$, lo_1o_2表示 O_1O_2 的欧式距离,且 $lo_1o_2 \neq 0$。若 $lo_1o_2 = 0$(例如多边形包含时),则认为是绝对相邻的,N_q 取研究目标相邻测度的最大值。当与某一特定空间特征相邻的多个特征属于同一分类 X 时,则该特征与相邻的同类特征空间相邻关系测度值取和,作为与 X 类的空间相邻测度值。图 2 所示的 3 个空间实体多边形 1、多边形 2、多边形 3,如果与多边形 2 相邻的特征多边形 1 和多边形 3 属于同一分类集 X,则它们合二为一,但多边形 2 与 X 类特征的相邻关系测度值为$\sum l / lo_{2-P}$, O_2-P 表示多边形 2 的重心点到相邻特征重心点的欧式距离。当两类特征

的相邻测度值达到给定的阈值时, 则认为研究区该特征是空间相邻的。以空间特征为基本单位, 计算其在特征分类集内与每一类分类特征的空间相邻测度值, 形成空间相邻测度矩阵, 即空间相邻测度值经过归一处理, 附加上空间特征的属性信息, 形成通过关联关系的测度, 就可以定量地对空间关联规则进行分析, 同时也可以推广到空间邻近关系(缓冲区分析), 形成类似于距某特征缓冲范围内某一特定特征的邻近度(Close-To)问题。

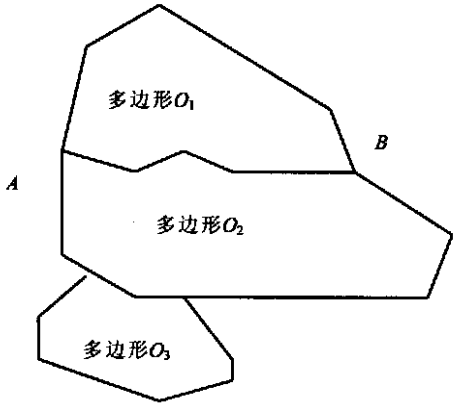

图 2　面状空间实体的相邻关系测度

Fig. 2　Neighboring Index Value for Polygon

2.3　关联规则及空间关联规则发现算法

关联规则的挖掘步骤为: 1)首先找出一个支持度大于给定值的频繁数据项集, 这是算法的关键。2)用频繁数据项集产生规则。对给定的全集 U, 如果其非空子集 $A \subseteq U$, 有 $\sup(A)/\sup(U) > \text{Confidence}$, 其中 $\sup(X)$、Confidence 分别表示支持度与信任度, 则产生形式为 $A \Rightarrow U - A$ 的规则。

经典的关联规则挖掘算法主要有 Apriori 和 DHP(direct hashing and pruning)等, 它们都属于数据库遍历类算法, 除此以外, Brins 等人提出了动态项目集计数 DIC 算法, 文献[8~11] 分别提出了“通用关联规则”、“多层次关联规则”、“定量关联规则”、“周期关联规则”等, 但是所有这些关联规则的挖掘都要多次扫描数据库, 造成算法的 I/O 开销过大, 特别是对大型数据库进行知识发现时, 算法的效率很低。

以上提出的各种算法主要针对事务型数据, 对于空间关联发现, 原则上也可以直接或经过少量的修正来应用, 但要在对空间数据进行转换以建立适合于算法的数据模型的基础上。即便如此, 以上算法的效率仍比较低下, 为此, 笔者提出了空间关联规则发现的回滚式算法。

3　知识的综合发现实现算法

3.1　一般过程描述

空间-属性综合挖掘可以分为 3 个子过程: ①首先找出一个支持度大于给定值的大数据项集; ②用这个大数据项集在给定的信任度下产生规则; ③知识的综合发现的规则集形成。其中①和②实质上是空间关联规则的发现, I/O 操作对象为 SUIT 中的 T, SR, SRV。而③是在发现的关联规则的基础上联合特征属性发现综合知识的过程, 该步骤可以利用前述的任一知识发现方法实现, 利用最小规则集生成算法完成[14], 完成知识的泛化过程。下面仅给出①的实现算法。

3.2　算法实现描述

用“位”段来设置关联标识, 采用回滚式的挖掘算法层层递进。设可能形成的项目集的最大维数为 m, 数据记录集的总记录数为 n, 要完成步骤①, 如果仅进行定性的大数据项集的发现, 只要一次扫描数据表即可, 并且对系统的内存要求不高, 如果要定量地发现空间关联规则, 则还需要访问一次数据库, 获取空间关联(相邻关系)规则的测度值。算法实施的基础是已经生成了 SUIT 数据表, 对数据表的扫描即为对 SUIT 表的扫描。基于“位”的回滚算法的实现过程如下。

1) 定义主列表二维数组 $N(p_1 \times p_2)$与加和一维数组 $A(k \times p_3)$, 其中 $p_1 = n$, $p_2 = \text{MOD}[(\sum C_m^i + 7)/8]$ $(i = 1, 2, \cdots, m)$, 等式 p_2 是对 m 个可能的项目集从 1 到 m 的组合和取模并按每字节包含 8 位字段求字节数, k 为数据类型为 long 型的字长(一般为 4 个字节), 取值 4, $p_3 = \sum C_m^i (i = 1, 2, \cdots, m)$。假设可能形成的项目集的最大维数为 10, 则主列表占用的空间为 128 个字长、p_1 个记录集的二维数组, 其中前 m 个 bit 为数据表数据初始化读入区, 该区记录的数据称为输入属性, 其他 bit 为递进算法的数据寄存区, 寄存区数据称为评价属性。每个 bit 代表的项目集的次序是严格排序的, 排序的规则为按一定的字母顺序排序, 如按前述的空间实体分类集 T, 排序为 A、B、C、D、E、AB、AC、…、$ABCDE$。加和一维数组中的元素用于计算和表示项目集的支持度, 该支持度与预定支持度比较, 如果小于给定的支持度, 则该项目集被淘汰, 且在向形成高维项目集回滚时, 该项目集不被考虑, 这是因为当低维项目集不是频繁项目集时, 包含该低维项目集的高维项目集不可能是频繁项目集。为方便起见,

初始化时数组的所有元素设置为 0。

2) 初始化数据读入区。扫描 SUIT 表,读取其内容来初始化数据读入区,当 SUIT 表的元素值不为 0 时,初始化数据读入区对应"位"段内容填 1,反之填 0。

3) 步骤 2)完成后,按列进行求和,并填写加和一维数组对应的列。当填写的值小于给定的最小支持度时,该列被放弃,即不是频繁项目集,回滚求更高维频繁项目集时将不予考虑。

4) 回滚求高维频繁项目集。按步骤 3)的结果,选取加和数组列表示的支持度大于给定支持度的列进行两两组合回滚"与"运算,其结果填写对应的回滚运算可能形成的二维频繁项目集数据寄存区,同样在完成运算之后填写加和数组对应的列,当填写的值小于给定的最小支持度时,该列被放弃,回滚求更高维频繁项目集时,该列也将不予考虑。当进行第 $k(k\geqslant 3)$次回滚时,将第 $k-1$ 次和初始化数据读入区进行"与"运算,如此回滚,最终加和数组中元素值大于预定的最小支持度的项目集,即为所求的所有频繁项目集。由于分类集进行了严格的排序,所以"与"运算的次数可以大大地缩小,在将第 $k-1$ 次的排序数据寄存区与初始化数据读入区的列进行组合时,仅需按一个方向即可,以空间分类集 $T=\{A, B, C, D, E\}$ 为例,如果一维频繁项目集为$\{A\}$、$\{C\}$、$\{D\}$、$\{E\}$,二维频繁项目集为$\{AC\}$、$\{CD\}$和$\{CE\}$,则生成三维可能的频繁项目集时只要进行 AC-D、AC-E、CD-E 三次比较即可。

5) 空间关联规则的定量挖掘。以上算法对一般的属性数据库进行频繁项目集的挖掘是非常有效的,但当对空间关联规则进行挖掘时,由于 SUIT 表不仅表示了空间是否具有所描述的空间关系,而且也表示了空间关系的测度值,故要得到空间关联的定量信息,还要对 SUIT 表进行第二次扫描,扫描的目的是获取测度值,当测度值的和大于给定的阈值时才认为该规则是定量空间关联的。

通过以上 5 个步骤,就可以生成具有复杂关系的空间关联规则,再用这个数据项集在给定的信任度下产生规则和知识的综合发展的规则集的形成过程,就可以完成知识的综合发现的全过程。以上算法在效率上是十分高的,在笔者的试验中(机器 CPU 为 PⅢ733),当空间对象集达到 1000 以上且空间对象平均相邻特征对象数在 4 个以上时,利用本文提出的回滚式算法与 Berzal 等提出的 TBAR 算法[15] 相比较,时间消耗要少一半以上。

4 应用实例

4.1 基本情况

通过对农业生产统计资料的分析,发现在不同地区种植的同种作物的产量存在较大的差异,其原因可能有两地区自然环境差异(属性信息)和作物空间布局的差异。为此,对两地区分别进行图像分类、空间关联关系发现、规则解释,寻找最终结论。

4.2 数据准备

研究区域是北方某县的两幅遥感图像,分别对应区域 R_1 和 R_2,从中选取农田影像特征比较明显的区域进行研究。首先对遥感影像进行有监督的空间目标分类,将空间实体(地块)分类为种植花生(A)、棉花(B)、玉米(C)、高粱(D)、其他作物类(E)5 类,提取出属于以上 5 类所有的特征对象,然后将影像转换为矢量型。笔者用 Arc/Info 的宏汇编语言(AML),按以上对空间相邻关系测度的定义,编程实现了研究区域分类特征的空间相邻关系的测度值计算。本例中的空间特征即为农田地块,按以上 5 类空间地物特征,将每个地块的所属类别附加在特征顺序编号前,R_1 的空间相邻关系的测度矩阵如表 2(仅选取部分特征),其中 *不参与归一化运算。同理可以得到 R_2 的空间相邻关系的测度矩阵,由于对不同研究区域数据处理方法相同,故以下空间关联规则生成以 R_1 为例说明。对表 2 中的空间相邻测度值进行归一化处理,*用常数 1 代替,表示空间特征的绝对相邻关系,附加上感兴趣的空间特征(地块)的属性值(单位面积的产量,通过遥感估产和实际调查获取,单位:kg/亩),获得了空间综合信息表(表 3)。以上过程中,考虑到各特征的相邻特征实体必须为考察的对象类型,凡是相邻对象

表 2 试验区田块的空间相邻关系测度矩阵

Tab. 2 Spaital Association Matrix in R_1

田块编号	A	B	C	D	E
A0001	*	3.12	2.15	0	1.08
A0002	*	0.94	0	0.26	0
…					
B0001	0	*	0	5.17	1.10
…					
D0001	0	4.98	2.15	*	3.45
…					
E0001	0.59	2.36	0	3.31	*
…					

为非考察类型(对应 ArcInfo coverage 格式的第 0

#多边形)的实体予以去除。然后在空间综合信息表的基础上, 利用提出的关联规则生成算法再进行下一步的规则生成。

表 3　试验区田块的空间综合信息表

Tab. 3　SUIT for Spatial Objects in R_1

田块编号	*A*	*B*	*C*	*D*	*E*	*Y*
*A*0001	1	0.37	0.26	0	0.13	…
*A*0002	1	0.11	0	003	0	…
…						
*B*0001	0	1	0	0.62	0.13	73.1
…						
*D*0001	0	0.60	0.26	1	0.41	…
…						
*E*0001	0.07	0.28	0	0.40	1	…
…						

4.3　空间关联规则生成

按回滚式算法步骤 1)构造的主列表二维数组的总列数为 4 个字节, 由空间综合信息表来初始化数据读入区(5 个位段), 然后进行回滚运算, 得到研究区域的空间对象相邻关系的规则集。由于仅研究相邻特征的关系, 因而仅需获取 2-频繁项目集。实例中, 预设 $s=70\%$, $c=50\%$。通过挖掘发现, 在支持度 $s=70\%$下, 区域 R_1 中得到的 2-频繁项目集为$\{B, D\}$, 在信任度 $c=50\%$下, 得到了规则 $B \Rightarrow D$, 意为在给定的支持度与信任度下, 种植棉花地块与高粱空间相邻。而在区域 R_2 中却没有发现具有空间关联的规则。

4.4　规则解释与评价

提取出具有空间相邻关联规则的空间实体集, 将研究区域 R_1、R_2 按特征的分类集进行属性数据(如产量)的平均值比较, 可以得出空间关联(相邻关系)规则导致产量的差异。农业领域专家可以据此进一步分析深层次的原因, 从高粱和棉花在种植的空间布局上具有空间关联关系的角度, 可以为他们提供一条有利的线索, 可能因为种植高粱有利于提高棉花的抗病虫害能力, 或者在养分、空气流通等因素上两者相邻会促进棉花的生长, 这在农业生产中也是合理的。获取的知识对农作物布局决策将起到很好的指导作用, 人们可以有意规划出作物生长布局的这种相邻关系, 达到增产的目的。所以发现的知识是有用的、可理解的。

需要明确的是, 例中形成的空间相邻关联在生产上可能并非预先有意创造的, 但是通过空间关联规则的发现, 从无意的生产操作发现了潜在的必然规律, 即如果具有这种空间相邻关系, 则该关系下的与空间决策有关的特征属性值将会显著高于(或低于)不具有该关系下的特征属性值。该情形在生产实践中具有广泛的应用价值, 这就是知识的综合发现的意义所在。

5　结　语

空间知识发现是针对具有空间特性的数据进行隐含模式的发现的, 空间关联规则发现作为空间知识发现的重要内容, 已经受到 GIS 及相关领域研究的重视。研究表明, 空间知识发现中也决不是独立于特征属性的, 本质上应与特征属性联合在一起。由于一般的属性数据库进行知识发现不可能发现空间规则, 因此迫切要求空间、属性特征的联合数据挖掘。空间、属性特征的综合数据挖掘可广泛地应用于环境、资源管理、工农业布局决策等领域。尽管本文提出的知识联合发现主要应用在关联规则与属性数据的联合数据挖掘上, 综合发现的思想在空间数据分类、聚类等其他知识发现也将具有很广泛的前景, 这将是下一阶段进行研究的重点。

参　考　文　献

1　Piatetsky-Shapiro G. Discovery, Analysis, and Presentation of Strong Rules. In: Piatetsky-Shapiro G, Frawlay W J, eds. Knowledge Discovery in Databases. Massechusetts: AAAI/ MIT Press, 1991. 229～238

2　Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. The 1993 ACM-SIGMOD, Washington DC, 1993

3　曾黄麟. 粗集理论及其应用. 重庆: 重庆大学出版社, 1996. 54～65

4　汪　闽, 周成虎. 空间数据挖掘方法的研究进展. 中国地理信息系统协会 2001 年年会, 成都, 2001

5　Lu W, Han J, Oci B C. Discovery of General Knowledge in Large Spatial Databases. Far East Workshop on Geographic Information Systems, Singapore, 1993

6　陈　军, 赵仁亮. GIS 空间关系的基本问题与研究进展. 测绘学报, 1999, 28(2): 95～102

7　Gold C M. The Meaning of "Neighbor". Lecture Notes in Computer Science. Pisa: Springer-Verlag, 1992. 220～235

8　Srikant R, Agrawal R. Mining Generalized Association Rule. The 21th VLDB Conference, Zurich, 1995

9　Han J, Fu Y. Discovery of Multi-level Association Rules from Large Databases. The 21th VLDB Conference, Zurich, 1995

10　Srikant R, Agrawal R. Mining Quantitative Association Rules in Large Relational Tables. The 1996 ACM SIG-

MOD Conference, Montreal, Quebec, 1996

11 Rahayana S, Siberschatz A. On the Discovery of Interesting Patterns in Association Rules. The 24the VLDB Conference, New York, 1998

12 边馥苓, 沙宗尧, 陈江平. 基于粗规则对象空间信息表的最小规则集生成. 武汉大学学报·信息科学版, 2001, 26(5): 399~403

13 Fernando B, Juan-Carlos C, Nicolas M. TBAR: An Efficient Method for Association Rule Mining in Relational Databases. Data & Knowledge Eng., 2001, 37(1): 47~64

作者简介: 沙宗尧, 博士生。现从事GIS 应用、GIS 空间数据模型和数据挖掘研究。

E-mail: zongyaosha@163.com

Comprehensive Knowledge Discovery: Theory, Concept and Application

SHA Zongyao[1] *BIAN Fuling*[1] *CHEN Jiangping*[1]

(1 School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)

Abstract: The principle of comprehensive knowledge discovery is proposed in this paper. We first propose some theories and concepts as the base of our research, which include spatial knowledge representation system, spatial object classification, spatial relations, comprehensive knowledge discovery, comprehensive knowledge and spatial union information table (SUIT), etc. In theory, SUIT records all information contained in the studied object. But in reality because of the complex and varieties of spatial relations, we only select some possible factors that we are interested in. The selected factors constitute the data to be processed. In this study, we select spatial association as the research emphasis, which was defined as sharing voronoi edge between spatial entities by C. M. Gold in 1992. The index value of spatial association is also introduced in our study. In order to find out the spatial association for spatial entities in a given study area, a highly efficient comprehensive knowledge discovery algorithm called recycled algorithm (RA) is also suggested.

As an example, a case study about comprehensive knowledge discovery for spatial association and attributes is studied. We compare the yields of different agriculture crops in two areas with similar climate characters and find that the difference of the same crop is notable. Through comprehensive knowledge discovery, we find that the reason lies in the spatial association relations. The study shows that the principle and methods proposed in this paper have an effective impact on comprehensive knowledge discovery and the discovered knowledge is both valuable and understandable.

Key words: comphensive knowledge discovery; spatial association; spatial union information table; knowledge discovery algorithm

About the author: SHA Zongyao, Ph. D candidate. His main research areas are GIS application, spatial data model and data mining

E-mail: zongyaosha@163.com