文章编号：1007 4619(2005)06 0733 09

从 GIS数据库中挖掘空间关联规则研究

马荣华[1]，马晓冬[2]，蒲英霞[2]
(1. 中国科学院南京地理与湖泊研究所，南京 210008; 2. 南京大学城市与资源学系，南京 210093)

摘 要： GIS数据库中空间关联规则发现是 SDMKD的重要内容，广泛涉及到知识的表示和推理，需要地理空间知识的深入参与。在地理空间认知的基础上，结合认知逻辑，通过 ILP对空间关联规则进行形式化描述，特别分析了其中涉及的空间谓词；通过例子说明了形式化空间关联规则的具体应用。从 GIS数据库中挖掘空间关联规则的主要问题是多层、多关系的规则挖掘问题，不同专题图层不同空间对象之间空间谓词的高效计算与存储表达是解决问题的关键；把空间关系非空间化，将连续数据离散化，从而把求解问题转换成布尔型关联规则问题进行讨论，基于此而探讨了一种通过 SJIP表组织空间谓词，然后根据目标对象的概念层级自顶向下、逐层细化的空间关联规则挖掘方法。

关键词： 空间关联规则；数据挖掘；GIS；认知；空间谓词
中图分类号： P208 文献标识码： A

1 引 言

空间数据挖掘和知识发现 (SDMKD)几乎与数据挖掘和知识发现 (DMKD)同时兴起于 20世纪 90年代[1, 2]，本质是从空间数据库或空间数据仓库中提取不明确的、隐含的知识、空间关系或其他模式[1, 3]，目的是发现、解释或预测空间现象或事件。作为最重要的一类空间数据库，GIS数据库中的隐含知识对基于知识的 GIS和智能空间分析是一个潜在的丰富资源，从 GIS数据库中发掘有用的隐含知识将提高 GIS的应用水平，并对建立智能化的 GIS起到极大的促进作用。GIS的地理空间认知特性决定了 GIS数据挖掘的特殊性，其关键是 GIS数据库中隐含知识的定义和获取[4]，广泛涉及到空间知识的表示和推理，需要综合知识的支持与结合[5]，需要用一定的形式化方法来建立现实知识和隐含知识的关系模型。从 GIS数据库可以发现的基本知识类型包括普遍的几何知识、空间分布规律、空间关联规则、空间聚类规则、空间特征规则、空间区分规则、空间演变规则、面向对象的知识等[6]；GIS数据库中空间关联规则发现是 SDMKD的重要内容，又被称作关联位置模式 (co location pattern)[7]，是 GIS数据库中反映空间对象结构以及空间关系的一些隐含规则，是 GIS空间关系或空间数据的不同子集之间的层次结构与相互关系的主要体现，涉及空间对象的拓扑关系、距离关系、方位关系以及它们的关系组合等；空间关联规则模式与关联规则模式一样，属于描述型模式[8]，具有较为严格的逻辑关系，因此可以在地理空间认知的基础上通过认知逻辑语言进行描述和表达，可以借鉴一般关系型事务数据库的关联规则挖掘方法，把空间数据转换为非空间数据，通过有空间数据标记的非空间数据来实现空间关联规则挖掘。

2 地理空间认知基础

空间关联规则挖掘的主要对象是 GIS数据库，GIS数据库的组织与加工实现了从现实世界到计算机地理空间世界的转变，是对现实世界的抽象与表达，抽象的过程是人们对现实世界进行认知的过程，表达的过程是人们对现实世界进行计算机再现的过程，因此 GIS数据库表达包含着地理空间认知的有关知识和内容。GIS数据挖掘是对存储于 GIS数据

收稿日期：2004 03 09；修订日期：2004 08 13
基金项目：国家自然科学基金 (40301038)。
作者简介：马荣华 (1972—)，男，山东临沂人，博士，目前从事 GIS数据挖掘与认知理论、水质遥感研究，发表论文 10余篇。E mail: rhma @ niglas. ac. cn。

库中的地理数据的概括、抽象与再加工,是从计算机世界到现实世界的反馈过程,是对地理空间世界的再认识,其中包涵地理思维的地理空间认知起着重要的作用。地理空间认知是对地理空间信息的表征,是一种逻辑思维加工过程,涉及三种空间框架[9]:地理空间(geographical space)、认知空间(cognitive space)和 Benediktine空间(cyberspace,赛博空间)。地理空间强调形态结构,包括几何的、拓扑的和尺度维的,认知空间强调空间知识的获取和学习,突出空间表达的用户理解,Benediktine空间即电脑(赛博)空间,以人机交互为基础,强调语义属性的存储以及它们之间的功能关系。认知空间为地理空间和赛博空间提供了表达与解释的桥梁和纽带,为 GIS空间信息的表达和应用提供了一个特别的视角。地理空间是一个连续的统一体,地理对象(现象)之间具有空间关联性和空间异质性;时空框架中地理对象的绝对和相对位置依其尺度和时间而变化,但可以对已有信息重新组织以获取变化后的信息属性,不同地理对象(现象)之间相互联系的空间扩展特征形成了空间关联的基础,隐含的空间关联规则也就成了 GIS空间数据挖掘的主要内容。

空间知识是人类最早获取的一类知识[10],空间的认知表达是地理信息科学研究的重要领域之一[11],形成了一定的认知图式[9]:容器(container,有内部、外部和边界,区域可以通过具有特定属性的容器来表达)、平面(surface,描述和表达连续数据)、远近(near far,描述和表达特征的突出程度)、垂面(verticality)、路径(path,描述和表达"源地—目的地"的概念)、链接(link,描述地理特征的连接性和邻接性)、中心外围(center periphery);上述认知图式可用下述的空间元概念(spatial primitive concepts)[9,12]来描述:标识(identity)、位置(location)、方向(direction)、距离(distance)、数量(magnitude)、尺度(scale)、时间(time or change)。空间元概念是对 GIS空间关联规则挖掘进行形式化描述的空间关系谓词表达的基础。

3 形式化描述

3.1 规则定义

仅用 ILP(Inductive Logic Programming,归纳逻辑设计)对空间关联规则进行描述与表达[13,14]缺乏足够的谓词支持,特别是对规则产生过程中真假值的表达;认知逻辑[15]恰好能够弥补这一缺陷。以认知逻辑为基础,结合 ILP,可以对空间关联规则(包括规则的产生过程)进行更加全面有效的描述和表达。认知逻辑规则描述语言的模型 M的组成如下[15]:

(1)一个可能世界的非空集 W;

(2)W上的一个二元关系 R,作为可达关系;

(3)对每个可能世界 w赋以一个体域 D_w的域函数 D;

(4)一个解释函数 I,它对每个命题逻辑语言 L_a中的常量 c赋以实体 I(c),而对每个 L_a中的 n元谓词 P就每个世界 w∈W均赋以一个 $(D_w)^n$子集 I_w(P)。

模型描述常用的连词包括与(∧)、或(∨)、否(~)、蕴涵或条件( )、双条件( )以及等于(=)等,常用的量词包括全称量词( )与存在量词( )等。相关定义如下[13,14,16]:

定义 1 SDMKD的可能世界非空集为 GIS数据库,记为 SDB,用 G(S)来表示;SDMKD的目标对象集记为 S,S G(S),具体目标对象记为 s,s∈S,具体的目标对象集用 O[s]表示,如式(1)[16],O[s] G(S);SDMKD的相关对象集记为 R,第 j个相关对象集记为 R_j,R_j R,R G(S),具体的相关对象记为 r_i,$r_i \in R_j$,R_j R;记先验知识(Background Knowledge)为 BK。

$$O[s] = O[s|s] \vee \{O[r_i|s] \mid \text{tuple}\theta \in G(S): \theta(s, r_i)\}_{1 \leqslant i \leqslant n} \quad (1)$$

式中,O[s|s]包含 s与 r_i的空间关系,O[r_i|s]包含 r_i与 $s' \in S$的空间关系;实例说明如表 1。

定义 2 A={a_1, a_2,…, a_n}记作原子(atom),A中原子之间的连接称作原子集(atom sets);模式集 L是一组基于 A的原子集,关键原子决定着模式 P(P∈L),把模式 P转换成原子查询要形成定量连接公式 eqc(P)。如果 eqc(P)在 O[s]∨BK中为真,则称模式 P覆盖 O[s]。

定义 3 简记 O[s]为 O,O_p为 O的子集即 O_p O,则模式 P的支持度记为 σ(P)=|O_p|/|O|。

定义 4 G(S)中的空间关联规则是一种二元关系,形式上定义为:P Q(s%, c%),式中 P∈L,Q A,P∧Q= ,P∨Q中至少有一个粒子表示空间关系。s%和 c%分别表示规则支持度和可信度;s=σ(P∨Q),c=φ(Q|P)=σ(P∨Q)/σ(P),前者是对关联规则重要性的衡量,支持度越大,关联规则越重要,应用越广泛,后者是对关联规则准确度的衡量,有些关联规则可信度很高,但支持度很低,说明该关联规则实用的机会很小,因此也不重要。

表 1　目标对象集 O[Xinliuhe]组成列表

Table 1　Listing of object aggregate O[Xinliuhe]

O[Xinliuhe|Xinliuhe]	O[Laoliuhe|Xinliuhe]	O[E1|Xinliuhe]
is_a(Xinliuhe, River)	is_a(Laoliuhe, River)	is_a(E1, Road)
To_meet(Xinliuhe, Loujiang)	To_meet(Xinliuhe, Laoliuhe)	To_intersect(Yanglintang, E1)
To_meet(Xinliuhe, Laoliuhe)	To_meet(Xingyanghe, Laoliuhe)	To_inside(B1, E1)
To_meet(Xinliuhe, Shibagang)	To_meet(Yantiehe, Laoliuhe)	To_intersect(Qiputang, E1)
To_meet(Xinliuhe, Tietang)	Line_to_region(Taicang, Laoliuhe)	……
To_meet(Xinliuhe, Yangtze River)	……	O[G1|Xinliuhe]
Line_to_region(Xinliuhe, Taicang City)	O[Loujiang|Xinliuhe]	is_a(E1, Expressway)
Line_to_region(Xinliuhe, Liuhezhen)	is_a(Loujiang, River)	……
To_intersect(Xinliuhe, E2)	……	O[G3|Xinliuhe]
To_intersect(Xinliuhe, G1)	O[Shibagang|Xinliuhe]	is_a(G3, Nation_Road)
To_intersect(Xinliuhe, G3)	is_a(Shibagang, River)	……
To_intersect(Xinliuhe, P1)	……	O[P1|Xinliuhe]
To_intersect(Xinliuhe, C1)	O[Tietang|Xinliuhe]	is_a(P1, Province_Road)
To_intersect(Xinliuhe, C2)	is_a(Tietang, River)	……
To_intersect(Xinliuhe, C3)	……	O[C1|Xinliuhe]
To_intersect(Xinliuhe, C4)	O[Yangtze River|Xinliuhe]	is_a(C1, Major_road)
O[Loujiang|Xinliuhe]	is_a(Yangtze River, River)	……
is_a(Loujiang, River)	……	O[C2|Xinliuhe]
To_meet(Jinjihe, Loujiang)	O[Taicang City|Xinliuhe]	is_a(C2, Major_road)
Line_to_line(P1, Loujiang)	is_a(Taicang City, town)	……
To_meet(Taicanggang, Loujiang)	To_intersect(P1, Taicang City)	O[C3|Xinliuhe]
Line_to_region(Kunshan City, Loujiang)	Close_to(Kunshan City, Taicang City)	is_a(C3, Major_road)
……	……	……

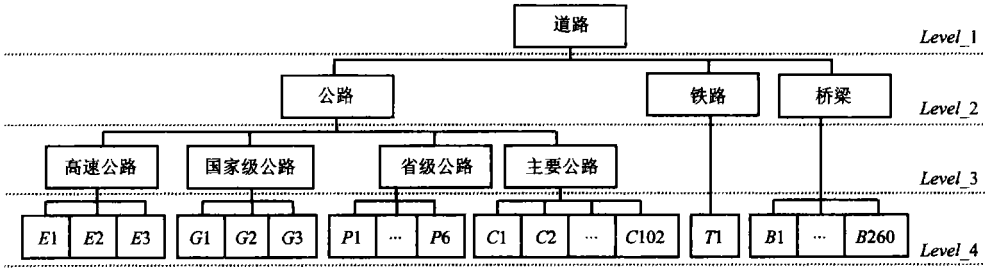
定义 5　设 mins[l]和 minc[l]为第 l层概念等级的最小支持度和最小可信度,如果 σ(P)≥mins[l],则模式 P在 l上是频繁的,且 P的所有祖先在各自的等级层次上都是频繁的;如果 φ(Q|P)≥minc[l],则空间关联规则 P　Q的可信度在 l上是高的;如果 P∨Q在 l上频繁且可信度高,则称 P　Q为强规则。

示例描述　任务:挖掘河流(S)与道路(R1)、桥梁(R2)以及城镇(R3)之间的空间关联规则;数据集 G(S):苏州地区河流、道路(包括桥梁)、城镇等 GIS数据库;先验知识(BK):道路的概念层级关系(图 1)。

先验知识的概念层级关系通过 is_a来断定,is_a在具体语言环境中具有重载功能,其中涉及的空间关系通过空间谓词来实现。这里仅通过新刘河 Xinliuhe(s)与公路(R1)、城镇(R2)的关系来说明空间关联规则形式化描述中 O[s]的表达(表 1)。

表 1表明,O[Xinliuhe]　G(S)不仅包含目标对象 Xinliuhe∈S与相关对象公路如 324省道以及相关对象城镇如太仓市之间的空间关系,如 to_intersect(Xinliuhe, P1)以及 line_to_region(Xinliuhe, Taicang City),还包含每一个相关对象与目标对象集 S中的其他对象 s′∈S之间的空间关系,如 line_to_region(Taicang City, Laoliuhe)。可以看到,O[Xinliuhe]∨BK覆盖了 eqc(P)　=　is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, road),因此可以说 P　=　is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, road)覆盖了表 1中的 O[Xinliuhe],该模式隐含的空间关联规则表示为:is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, road)　is_a(Y, bridge)(80%, 100%)。

该空间关联规则存在于 level_2中(图 1),另外在 level_3以及 level_4上也存在类似的规则,但支持度以及可信度有所不同,这就是多层空间关联规则挖掘。不同等级的道路在经过不同等级的河流时,桥梁的等级程度也不同,因此可以通过一定的算法(如决策树支持算法)先获取数据库中如图 1所示的道路层级一样的河流层级结构以及桥梁层级结构,进而获取目标对象与相关对象之间不同层级的空间关联规则,这就是跨层空间关联规则挖掘。



E1: 沪宁高速 E2: 沿江高速 E3: 苏嘉杭高速 T1: 沪宁铁路

G1: 204国道 G2: 312国道 G3: 204国道复线

P1: 324省道 P2: 209省道 P3: 苏常线 P4: 沪宜线 P5: 张扬线 P6: 锡沙线

C1: 沪浮璜线 C2: 沪浮线 C3: 渔港线 C4: 牌新长线 C5: 越湖线 C6: 苏沪机场路

C7: 西南环线 C8: 木东线 C9: 望湖线 C10: 渡假区专线 C11: 湖滨线 C12: 太西线

C13: 海新线 C14: 常福线 C15: 常昆线 C16: 海丰线 …… C101: 周湖线 C102: 周师线

B_i: GIS数据库中第 i座桥梁, $1 \leqslant i \leqslant 260$

图 1 苏州市路网层级划分

Fig 1 A spatial concept hierarchy for the roads in Suzhou region

3.2 规则谓词

谓词逻辑是空间关联规则形式化描述与求解的重要内容,通过谓词逻辑演算的各种等价关系可以获得归结规则,并通过归结推理的方法进行问题求解。从 GIS数据库中挖掘空间关联规则可能使用的谓词主要包括 6类:

(1)认知模态谓词 包括必然(□)和可能(◇);据此, P Q(s%, c%)可分解为: s=100, □(P Q)即 P Q一定会成立; 0<s<100, ◇(P Q)即 P Q可能会成立;

(2)一般时态谓词 包括总会(G,将会总是)、总有(H,过去总是)、将会(F,将来会有)和曾有(P,曾经有过);例如, □G(P Q),即 P Q将会总是成立是必然的; ◇G(P Q),即 P Q将会总是成立是可能的;

(3)GIS时态谓词 时间是重要的空间元概念之一, GIS空间关联规则的形成具有一定的时态环境,涉及 GIS时态拓扑关系,主要包括分支、线性以及循环等三种时态结构[17],要表达的时态谓词主要有: Before/after, equal, start/finish, meet, overlap, end, during,可以分别用 ti before(ti after)、ti equal, ti start(ti finish)、timeet, ti overlap, ti end 以及 ti during等算子来表达;

(4)GIS空间谓词 位置、方向和距离是空间元概念的重要内容,是描述、表达与获取 GIS空间关联规则的最重要的组成部分,根据它们的表述内容可以把 GIS空间谓词分有 3类:表示拓扑结构的谓词(用 to disjoint, to intersect, to contain, to inside, to meet, to equal, to cover以及 to cover by等算子来表达)、表示空间方向的谓词(用 direct, east of, west of, south of, north of, southeast of, northeast of, southwest of以及 southeast of等算子来表达)和表示距离的谓词(用 dist, close_to/near_to以及 far_away等算子来表达);

(5)点线面相互关系谓词 即点对象、线对象和面对象之间的空间关系谓词,可以分别用 point to line, point to region, line to line, line to region以及 region to region等算子来表达;

(6)分类谓词 如 is_a等。

4 挖掘方法

从 GIS数据库中挖掘空间关联规则不同于一般的关系数据挖掘,涉及到空间关系计算,这与 GIS数据库的存储模式以及数据结构息息相关。本文在诸多学者研究[13, 14, 16, 18, 19]的基础上,借鉴 Apriori算法[20],探讨一种基于 ILP的以参考变量为中心的多层多关系空间关联规则挖掘方法,基本框架如图 2,涉及两个关键问题: (1)不同专题图层之间不同对象空间关系谓词的计算、存储与表达, (2)不同概念层级之间空间关联规则的产生。

GIS中的空间关系是隐含的,显式地表明这些空间关系并不容易,特别是不同专题图层之间不同对象的空间关系的计算与表达较为复杂[21, 22],空间连接索引(Spatial Join Index, SJI)[21, 23]可以有效地

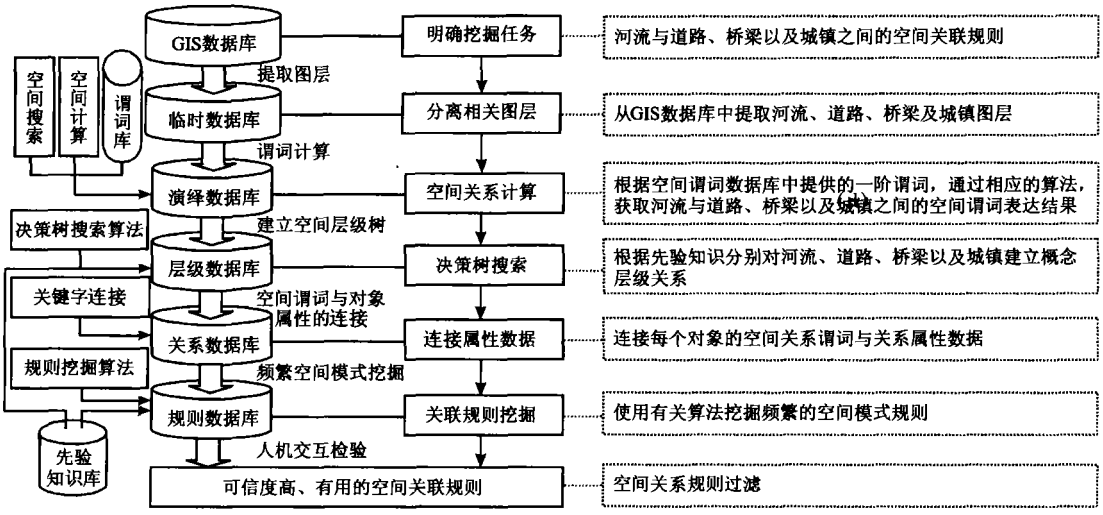


图 2　GIS空间关联规则挖掘的基本框架

Fig. 2　Framework of spatial association rules mining from GIS databases

简化这一过程。SJI是一张两个不同专题图层之间的匹配表,为了获取并记录两个不同专题图层之间不同对象的空间关系谓词,增加一个 PredicateID项,存储空间谓词取值编码,称此表为空间连接谓词索引表 (Spatial Join Index for Predicates, SJIP, 图 3(a))。

构建 SJIP之前,需要先构建空间谓词取值表(图 3(b)),即按照 2.2节设计的空间谓词计算不同专题图层之间各对象的空间关系;空间谓词使得不同空间对象的空间关系实现了非空间显式化表达,并通过关系数据库来储存,从而为布尔型关联规则问题的提出与解决提供了一个良好的基础,使得从GIS数据库中挖掘空间关联规则的问题转化为了从关系数据库中挖掘关联规则的问题,属于多关系表的数据挖掘范畴,达到了简化研究问题的目的。然后通过先验知识或者通过相关决策树归纳算法[22, 24]建立道路、河流以及城镇的概念层级决策树,设置层级编码、以及每一层的最小支持度 (minsup[l])和最小可信度 (minconf[l]);计算不同专题图层不同对象之间的空间谓词,同决策树层级编码一起

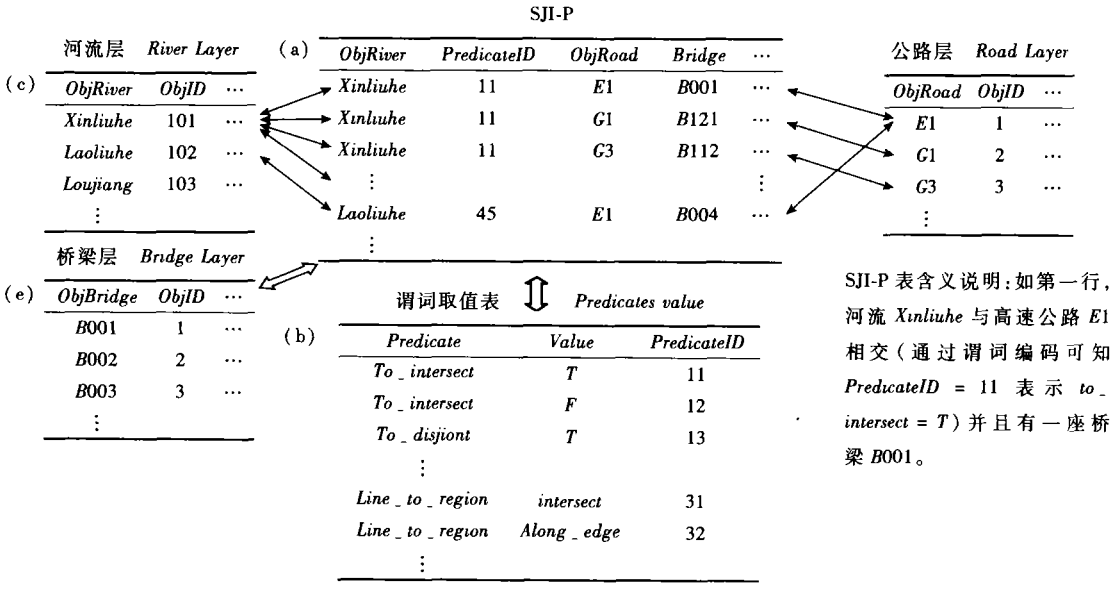


图 3　空间谓词连接索引

Fig. 3　Spatial join index for predicates (SJIP)

加入到 SJIP表中 (图 4)。便于解释说明,不妨把图 3中真实的 SJIP表进行简化 (图 5(a),(b)),并通过 Bridge项把图 5中的 SJIP表 1和 SJIP表 2连接到一起 (图 5(c)),挖掘任务的参照目标项 (如图 5中 SJIP表 1中的 ObjRoad)包含了先验知识决策树的层级信息,如 E表示高速公路、P表示省级公路、C表示主要公路以及 T表示铁路等。这样就把多关系的空间关联规则挖掘问题转换为顾客交易数据库中项集[19,20]间的关联规则挖掘问题,然后以挖掘任务的目标对象为中心,自顶向下,逐步求精,发现较低概念层次的空间关联规则;概念层次越高,规则的最小支持度就越高。

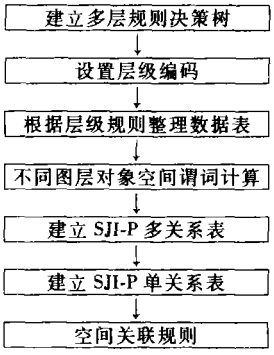


图 4 多层多关系表挖掘

Fig 4 Multi level and multi relational mining

假设同一条河流与同等级公路之间的空间关系为一个交易,交易号 (TID)为 7—9位编码,如 {11, B, 31, G, E1} (前两位 11表示空间谓词取值即河流与道路相交,第 3位 B表示相交处有桥、N表示无桥,第 4—5位表示河流穿越城市,第 6位 G表示公路、T表示铁路、B表示桥梁,后面几位表示具体的道路),简化后的图 5(c)位于图 1的 level_2层级上,过滤掉与挖掘任务无关的数据项 (如 ObjRiver, ObjTown等),转换为图 6(a),称为空间关联规则编码交易表,记为 T[2],对其进行遍历,即搜索河流、道路相交的数据项,设最小支持度为 5 (为方便,用绝对值表示),导出 level_2的频繁 1项集 L(2,1)、频繁 2项集 L(2,2)、频繁 3项集 L(2,3)以及频繁 4项集 L(2,4) (图 6(b),(c),(d),(e));然后使用能够表达河流与道路、桥梁之间基本关联规则的频繁集即 L(2,3),再导出位于 level_3上的空间关联规则编码交易表,记为 T[3] (图 7(a)),设最小支持度为 2,再导出 level_3的频繁集 L(3,4)以及 L(3,5) (图 7(b),(c));同样导出最底层的 T[4] (图 8(a))以及最小支持度为 1的频繁集 (图 8(b))。上述过程可以描述为:首先计算不同专题图层的目标对象和相关对象之间的空间谓词,通过 SJIP表将其组织成一个关系型的事务数据库,进而转化为空间交易数据库,然后根据目标对象的概念层级对其进行空间关联规则挖掘;挖掘过程根据概念层级建立的空间谓词粒度而不断交替循环,首先挖掘高层粗粒度空间谓词的频繁模式和关联规则,然后对频繁模式中的空间对象进一步计算低层细粒度的空间

(a) SJI-P 表1

ObjRiver	PredicateID	ObjRoad	Bridge
Xinliuhe	11	GE1	B001
Xinliuhe	11	GP1	B041
Xinliuhe	11	GG1	B121
Xinliuhe	11	GG3	B112
Xinliuhe	11	GC45	B205
Xinliuhe	11	GC68	B213
Xinliuhe	11	GC89	B145
Xinliuhe	11	GC113	no

(b) SJI-P 表2

ObjRiver	PredicateID	ObjTown	Bridge
Xinliuhe	31	Liuhezhen	B041
Xinliuhe	31	Liuhezhen	B205
Xinliuhe	32	Luduzhen	B213
Xinliuhe	31	Taicangshi	B001
Xinliuhe	31	Taicangshi	B121
Xinliuhe	31	Taicangshi	B112

(c)

ObjRiver	PredicateIDRoad	ObjRoad	Bridge	PredicateIDTown	ObjTown
Xinliuhe	11	GE1	B001	31	Taicangshi
Xinliuhe	11	GP1	B041	31	Liuhezhen
Xinliuhe	11	GG1	B121	31	Taicangshi
Xinliuhe	11	GG3	B112	31	Taicangshi
Xinliuhe	11	GC45	B205	31	Liuhezhen
Xinliuhe	11	GC68	B213	32	Luduzhen
Xinliuhe	11	GC89	B145	00	—
Xinliuhe	11	GC113	—	00	—

图 5 SJIP多关系表到单关系表的转换

Fig 5 From the multi relational SJIP to the single relational one

(a)

TID	项
T1	{11, B, 31, G, E1}
T2	{11, B, 31, G, P1}
T3	{11, B, 31, G, G1}
T4	{11, B, 31, G, G3}
T5	{11, B, 31, G, C45}
T6	{11, B, 32, G, C68}
T7	{11, B, 00, G, C89}
T8	{11, N, 00, G, C113}

(b)

项集	支持度
{11, , , , }	8
{ , , , G, }	8
{ , B, , , }	7
{ , , 31, , }	5

(c)

项集	支持度
{11, B, , , }	7

(d)

项集	支持度
{11, B, , G, }	7

(e)

项集	支持度
{11, B, G, 31, }	5

(a) 空间关联规则编码交易表 T[2]

(b) Level_2的频繁 1项集 L(2, 1)

(c) Level_2的频繁 2项集 L(2, 2)

(d) Level_2的频繁 3项集 L(2, 3)

(e) Level_2的频繁 4项集 L(2, 4)

mins[2] =5

图 6　Level_2上 T[2]的遍历

Fig 6　Searching T[2] at level_2

(a) *mins*[3]=2

TID	项
*T*1	{11, *B*, 31, *G*, *E*1}
*T*2	{11, *B*, 31, *G*, *P*1}
*T*3	{11, *B*, 31, *G*, *G*1}
*T*4	{11, *B*, 31, *G*, *G*3}
*T*5	{11, *B*, 31, *G*, *C*45}
*T*6	{11, *B*, 32, *G*, *C*68}
*T*7	{11, *B*, 00, *G*, *C*89}

(b)

项集	支持度
{11,*B*,* * *,*G*,*G** * *}	2
{11,*B*,* * *,*G*,*C** * *}	2

(c)

项集	支持度
{11,*B*,31,*G*,*G** * *}	2

(a) 空间关联规则编码交易表 *T*[3]

(b) *Level*_3 的频繁 4-项集 *L*(3,4)

(c) *Level*_3 的频繁 5-项集 *L*(3,5)

图 7　Level_3上 T[3]的遍历

Fig 7　Searching T[3] at level_3

(a) *mins*[4]=1

TID	项
*T*1	{11, *B*, 31, *G*, *E*1}
*T*2	{11, *B*, 31, *G*, *P*1}
*T*3	{11, *B*, 31, *G*, *G*1}
*T*4	{11, *B*, 31, *G*, *G*3}
*T*5	{11, *B*, 31, *G*, *C*45}

(b)

项集	支持度
{11, *B*, 31, *G*, *E*1}	1
{11, *B*, 31, *G*, *P*1}	1
{11, *B*, 31, *G*, *G*1}	1
{11, *B*, 31, *G*, *G*3}	1
{11, *B*, 31, *G*, *C*45}	1

(a) 空间关联规则编码交易表 *T*[4]

(b) *level*_4 的频繁 5-项集 *L*(4,5)

图 8　Level_4上 T[4]的遍历

Fig 8　Searching T[4] at level_4

谓词,再挖掘相应的频繁模式和关联规则,直至不能发现新的频繁模式为止。上述过程产生了如下的主要空间关联规则:

(1) level_2上,规则 1: is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, road) is_a(Y, bridge) (87 5%, 100%),该规则意味着如果河流与公路相交,则相交处有桥的支持度为 87 5%,可信度为 100%;规则 2: is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, road)∧[line_to_region(X, w) = intersect]∧is_a(w, town) is_a(Y, bridge) (62 5%, 100%),该规则意味着如果河流与公路相交,并且穿越城镇,则相交处有桥的支持度为 62 5%,可信度为 100%; (2) level_3上,规则 1: is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, nation_road) is_a(Y, bridge) (28 6%, 100%),该规则意味着如果河流与国家级公路相交,则相交处有桥的支持度为 28 6%,可信度为 100%;规则 2: is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, major_road) is_a(Y, bridge) (28 6%, 66 7%),该规则意味着如果河流与主要公路相交,则相交处有桥的支持度为 28 6%,可信度为 66 7%;规则 3: is_a(X, river)∧to_intersect(X, R)∧to_inside(Y, R)∧Y\=X∧is_a(R, nation_road)∧[line_to_region(X, W) = intersect]∧is_a(w, town) is_a(Y, bridge) (28 6%, 100%),该规则意味着如果河流与国家级公路相交,并且相交处穿越城镇,则相交处有桥的支持度为 28 6%,可信度为 100%; (3) level_4上,有 5条类似上述的规则。

5 结 论

(1)从 GIS数据库中挖掘空间关联规则是数据挖掘也是空间数据挖掘研究的一个特例,存在于计算机世界到现实世界的反馈过程中,需要充分理解GIS的有关知识,并涉及到地理空间认知的相关理论和内容。因此从 GIS数据库中挖掘知识是一个多学科交叉的综合问题。先验知识往往起到深层次发现知识的基石作用[5],特别是目标对象的概念层级关系,对频繁模式的关联规则发现具有一定的引导作用,同时增加了关联规则的实用性和可信度,但引起了多层、跨层的规则挖掘问题,基于传统关联规则的挖掘方法是解决该问题的思路之一,其中又以参考变量为中心的模型为主。把空间关系非空间化,将连续数据离散化,从而把求解问题转换成布尔型关联规则问题,基于此而探讨了一种通过 SJIP表组织空间谓词,然后根据目标对象的概念层级自顶向下、逐层细化的空间关联规则挖掘方法。

(2)谓词逻辑是认知逻辑在 GIS空间关联规则挖掘过程中首先碰到的问题,适用于表达空间关系和背景知识;但不同专题图层不同空间对象之间的空间谓词计算与表达是 GIS空间关联规则挖掘面临的难点之一。以认知逻辑提供的规则描述语言模型为基础,给出了空间关联规则的形式化表达式,特别通过例子说明了目标对象集的组成,可以更加充分地理解规则的支持度与可信度。认知逻辑主要研究知识和信念的形式化问题,处理的是有关知识和信念等认知概念的逻辑性质和关系问题[15];因此可以通过认知逻辑语言对其表达式进行细化和约简,使该表达式能够清楚地表明空间关联规则的挖掘过程与所需条件。

(3)空间关系非空间化就是把 GIS中隐含的空间关系通过空间谓词显示地表达出来,是对 GIS数据库进行重新整理的过程,势必会损失信息,影响挖掘结果,特别影响到通过定量分析模型而挖掘的隐含定量知识,空间关联规则属于定性描述规则,除非在定性获取规则的过程中使用定量化分析(如空间回归等)手段,否则不会对结果产生直接影响;损失信息对挖掘结果的影响评价要以具体的挖掘目标为前提,结合应用实例,通过定量模型来进行。

(4)从 GIS数据库中发现的强空间关联规则有很多,其中多余规则占 30%— 60%[19],另外用户并不是对每一条都感兴趣,因此必须剔除多余的、不必要的规则;可以在认知逻辑的基础上,通过规则约简建立多余规则与原规则的逻辑关系,进而剔除多余规则。

参 考 文 献 (References)

[1] Li D R, Wang S L, Shi W Z, et al. On Spatial Data Mining and Knowledge Discovery (SDMKD) [J]. Geomatics and Information Science of Wuhan University, 2001, 26 (6): 491— 499. [李德仁,王树良,史文中等. 论空间数据挖掘和知识发现 [J]. 武汉大学学报 (信息科学版), 2001, 26 (6): 491— 499.]

[2] Fayyad U M, Piatetsky Shapiro G, Smyth P, et al. Advances in Knowledge Discovery and Data Mining [M]. AAAI Press, Menlo Park, CA. 1996.

[3] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases [A]. Egenhofer M J, Herring J R. Advances in Spatial Databases [C], LNCS951, Springer Verlag, Berlin, 1995: 47— 66.

[4] Sester M. Knowledge Acquisition for the Automatic Interpretation of Spatial Data [J]. International Journal of Geographic Information Science, 2000, 14 (1): 1— 24.

[5] Ma R H, Huang X Y, Zhu C G. Knowledge Discovery with ESDA from GIS Database [J]. Journal of Remote Sensing, 2002, 6 (3): 102— 107. [马荣华,黄杏元,朱传耿. 用 ESDA技术从 GIS数据库中发现知识 [J]. 遥感学报, 2002, 6 (3): 102— 107.]

[6] Di K C, Li D R, Li D Y. A Framework of Spatial Data Mining and Knowledge Discovery [J]. Journal of Wuhan Technical University of Surveying and Mapping, 1997, 22 (4): 328— 332. [邸凯昌,李德仁,李德毅. 空间数据发掘和知识发现的框架 [J]. 武汉测绘科技大学学报, 1997, 22 (4): 328— 332.]

[7] Shekhar Shashi, Huang Y. Discovering Spatial Co location Patterns: A Summary of Results [A]. In: Proceedings of 7th International Symposium on Spatial and Temporal Databases (SSTD01) [C], 2001.

[8] Shi Z Z. Knowledge Discovery [M]. Beijing: Tsinghua University Press, 2002. [史志植. 知识发现 [M]. 北京:清华大学出版社, 2002.]

[9] Fabrikant S I, Buttenfield B P. Formalizing Semantic Spaces for Information Access [J]. Annals of the Association of American Geographers, 2001, 91 (2): 263— 280.

[10] Taylor H A, Tversky B. Perspective in Spatial Descriptions [J]. Journal of Memory and Language, 1996, 35: 371— 391.

[11] Montello D R. Cognition of Geographic Information [DB/OL]. In: UCGIS. Research Priorities for Geographic Information Science. http://www.ncgia.ucsb.edu/other/ucgis/research_priorities/paper4.html. 2000.3.

[12] Geolledge R G. Primitives of Spatial Knowledge [A]. Nyerges T L, Mark D M, Laurini R, et al. Dordrecht Cognitive Aspects of Human computer Interaction for Geographic Information Systems, vol 83, NATO ASI Series D: Behavioral and Social Sciences [C], The Netherlands: Kluwer Academic Publishers, 1995.

[13] Popelinsky L. Knowledge Discovery in Spatial Data by Means of ILP [A]. Zytkow J M, Quafalou M. Principles of Data Mining and Knowledge Discovery [C], LNAI 1510, Springer Verlag

Berlin, 1998: 185—193.

[14] Malerba D, Lisi F A. An ILP Method for Spatial Association Rule Mining [A]. In: Working Notes of the First Workshop on Multi relational Data Mining. Freiburg [C], Germany, 2001: 291—314.

[15] Zhou C L. Introduction to Epistemic Logic [M]. Beijing: Tsinghua University Press, 2001. [周昌乐. 认知逻辑导论 [M]. 北京:清华大学出版社, 2001.]

[16] Malerba D, Lisi F A, Appice A, et al. Mining Spatial Association Rules in Census Data: A Relational Approach [A]. In: Proceeding of the ECML/PKDD' 02 Workshop on Mining Official Data. University Printing House, Helsinki, 2002: 80—93.

[17] Egenhofer M J, Golledge R G. Spatial and Temporal Reasoning in Geographic Information Systems [M]. Oxford: Oxford University Press, 1998.

[18] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases [A]. In: Proceedings of the ACM SIGMOD Conference on Management of data [C], 1993: 207—216.

[19] Han J, Fu Y. Mining Multiple level Association Rules in Large Databases [J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11 (5): 798—805.

[20] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules [A]. In: Proceedings of 1994 International Conference on Very Large Data Bases [C], Santiago, Chile, 1994: 487—499.

[21] Zeitouni K, Yeh L, Aufaure M. Join Indices as a Tool for Spatial Data Mining [A]. In: International Workshop on Temporal, Spatial and Spatio Temporal Data Mining. Lecture Notes in Artificial Intelligence [C], Springer, Lyon, France, 2000: 102—114.

[22] Chelghoum N, Zeitouni K, Boulmakoul A. A Decision Tree for Multi layered Spatial Data [A]. In: 10th International Symposium on Spatial Data Handling. SDH [C], Ottawa, Canada, 2002.

[23] Valduriez P. Join Indices [J]. ACM Transactions on Database Systems, 1987, 12 (2): 218—246.

[24] Knobbe J, Siebes A, Van der Wallen D M G. Multi relational Decision Tree Induction [A]. In: Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD' 99 [C], 1999.

Spatial Association Rule Mining from GIS Database

MA Rong hua[1], MA Xiao dong[2], PU Ying xia[2]

(1. Nanjing Institute of Geography and Limnology, CAS, Nanjing 210008, China;
2. Dept. of Urban & Resources Science, Nanjing University, Nanjing 210093, China)

Abstract: To discover spatial association rule is one of the important contents for spatial data mining and knowledge discovery (SDMKD), which extensively concerns with spatial, especially geographic spatial knowledge expression and reasoning. Combining with epistemic logic, formal expression is described by Inductive Logic Programming (ILP), on the basis of geographic spatial cognition. And then the spatial predicates, possibly used in SDMKD from GIS, are analyzed and listed. Subsequently, the formal expression of spatial association rules is explained through taking examples for the relation rules between roads, rivers and towns in Suzhou region. The major problem to mine spatial association rule from GIS is to mine multi level and multi relational rules. And the key to solve the problem is how to effectively compute, store and express the spatial predicates among different spatial objects of different thematic layers. Firstly, the spatial data are transformed to the non spatial data, which can be described in the related tables. And then the original problems are transformed into the problems in Boolean logic rules. Finally, on the basis of the mentioned above, according to the concept hierarchy of spatial objects, the spatial association mining approach, from top to bottom and deepening step by step, is introduced. And the spatial predicates are organized by spatial join index for predicates.

Key words: spatial association rule; data mining; GIS; cognition; spatial predicate