

引用格式:陈祖刚,杨雅萍.耦合尺度的地理实体空间相关度算法的建立与应用[J].地球信息科学学报,2018,20(1):37-47. [Chen Z G, Yang Y P. A case of establishment and application of spatial correlation degree algorithm for geographic entities coupling scales[J]. Journal of Geo-information Science, 2018,20(1):37-47.] DOI:10.12082/dqxxkx.2018.170323

耦合尺度的地理实体空间相关度算法的建立与应用

陈祖刚,杨雅萍*

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101; 2. 中国科学院大学,北京 100049

A Case of Establishment and Application of Spatial Correlation Degree Algorithm for Geographic Entities Coupling Scales

CHEN Zugang, YANG Yaping*

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The traditional correlation degree algorithms for geographic entities have many disadvantages, such as non-applicable for some kinds of geographic entities and some types of topological relations, and not considering the dependency of spatial scale that results in poor discernibility of data. In this study, a new algorithm is proposed which computes the spatial correlation degree according to the specified spatial scale which is represented by a spatial extent. Based on the first law of geography and the theories on spatial correlation degree put forward by Egenhofer, the equations of spatial correlation degree was obtained by analyzing the topological and metric relations between different kinds of geographical entities such as points, lines and polygons. By comparison, the algorithm in this study can compute the correlation degree between geographic entities of different types and topological relations, alter the correlation degree with the change of the specified spatial scale, which is consistent with the generic intuition of human beings. At last, we introduced an application of the algorithm by taking geospatial data retrieval as an example. Compared with the traditional retrieval methods based on keyword matching, our algorithms can improve the *F1-measure* in geographic information retrieval (GIR) and give the accurate scores of correlation degree so that the retrieval results can be ranked. The algorithm is an elementary research that can be applied in the research fields of GIR, scientific data discovery, data recommendation, linked data, and so on.

Key words: spatial scale; geographic entity; correlation degree; geographic information retrieval; data discovery

*Corresponding author: YANG Yaping, E-mail: yangyp@igsnrr.ac.cn

摘要: 传统的地理实体空间相关度算法存在适应的实体和拓扑关系类型较少、没有考虑空间尺度依赖性而导致数据区分能力差的问题。本研究提出一种能依据指定的空间尺度(本文所指“空间尺度”是指定的地理空间范围),计算出相应的地理实体

收稿日期 2017-07-11;修回日期:2017-08-06.

基金项目: 中国工程科技知识中心地理资源与生态分中心建设项目(CKCEST-2017-1-8);国家地球系统科学数据共享服务平台(2005DKA32300);江苏省地理信息资源开发与利用协同创新中心资助项目。[**Foundation items:** Branch Center Project of Geography, Resources and Ecology of Knowledge Center for Chinese Engineering Sciences and Technology, No.CKCEST-2017-1-8; National Earth System Science Data Sharing Infrastructure, No.2005DKA32300; Project of Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application.]

作者简介: 陈祖刚(1989-),男,河南信阳人,博士生,主要从事地学数据挖掘研究, E-mail: czgbjy@yeah.net

***通讯作者:** 杨雅萍(1964-),女,北京市人,高级工程师,主要从事地球科学数据与共享研究。E-mail: yangyp@igsnrr.ac.cn

空间相关度的算法。该算法以地理学第一定律和Egenhofer关于空间相关度的论述为理论依据,分析点、线、面实体的拓扑关系和度量关系而建立不同的相关度计算公式。通过对比分析,本算法不仅能计算出不同类型和不同拓扑关系下的地理实体间相关度,而且计算结果随着空间尺度的变化而改变,与人类通常的认知相符合。最后,以地理空间数据检索为例,介绍了本算法的应用。与传统的关键词匹配检索方法相比,应用本算法能提高数据检索的 $F1-measure$ 值,并且能对文档按照与检索词的相关度进行排序。本算法可应用于地理信息检索、数据发现、数据推荐和关联数据等领域。

关键词 空间尺度;地理实体;相关度;地理信息检索;数据发现

1 引言

尺度一般是指时间的长短或空间范围的大小,即所谓时间尺度和空间尺度;地学研究中常采用狭义的尺度概念,即空间尺度^[1]。空间尺度的定义较为模糊,其内涵包括认知尺度、空间细节层次、地图比例尺、空间粒度、空间范围等^[2]。研究过程和研究结论依赖于特定空间尺度的特点,称为尺度依赖性。20世纪50年代,Robinson最早发现地学和社会学研究中的尺度依赖性问题^[3],发现在某一尺度下得出的结论不能无差别的适用于另一尺度;随后Openshaw等^[4]提出的可变元面积(Modifiable Areal Unit Problem, MAUP)问题,Goodchild^[5]、Dudley^[6]、Marceau等^[7]研究了MAUP对于统计模型和地学分析的影响,证实此类地学模型的尺度依赖性。

相似度或相关度是人类认知过程中的基础内容,它是归类的准则、也是联想记忆和演绎推理的依据^[8-9]。相关度在信息检索、信息集成和数据挖掘等领域有着重要应用^[10]。在地理信息检索(Geographic Information Retrieval, GIR)和时空数据挖掘等领域,地理实体的空间相关度也是研究的基础问题^[11]。地理实体的空间相关度指的是地理实体间的位置、距离、方位、形状和大小等几何属性的相关度程度。

已有的地理实体相关度(下文除特殊指定外,实体均指地理实体,相关度均指空间相关度)算法主要分为2类:①基于空间关系的方法。空间关系包括拓扑关系、度量关系和方位关系^[12]。如Hill^[1]和Walker等^[13]分别提出了基于面状空间要素重叠面积大小的地理实体相关度算法;Janeé等^[14]提出了基于Hausdorff距离的算法;Beard and Sharma^[15]提出了基于度量关系和空间拓扑关系的算法;Li等^[16]提出了一种综合利用拓扑关系、度量关系和方位关系计算地理实体相关度的模型;Frontiera^[17]提出一种利用空间拓扑关系和度量关系因子进行逻辑回归的实体相关度算法;国内学者如刘家骏对文本中以空间陈述形式存在的地理信息的模糊性因素进行分析,通过引入不确定场模型来描述参照对象和空

间关系对模糊性的贡献程度,提出了空间相关性模糊度量的计算方法^[18];赵宏伟等^[19]提出了利用地理空间拓扑关系、度量关系和专家打分给出的权重系数计算地理实体相关度的方法。②基于地名词典和本体的方法。如Rodríguez和Egenhofer^[11]提出了一种匹配距离的实体类相似度计算方法;李红梅等^[20]以本体做为揭示实体深层次语义信息的方法论,提出了符合认知特性的地理实体类型语义相关度计算模型;杨娜娜等^[21]基于地理空间概念本体,加入深度、密度权重因子,提出了基于本体的地理实体相关度计算方法;Janowicz等^[22-23]提出了一种基于地理空间描述逻辑的SIM-DL模型计算实体相关度。

现有的地理实体相关度算法主要存在以下问题:①构建地理本体需要完整的概念体系和概念之间的空间关系,难度大、耗时长;②地名词典、地理语义目录不能够表达地理实体间的拓扑关系;③基于空间关系的实体相关度算法,大多只适合面状实体,对点状和线状实体支持不够,文献[19]提出了一种较为全面的算法,但其拓扑关系基于4-交模型,能处理的拓扑关系种类有限;④所有的实体相关度算法都没有考虑空间尺度依赖性,在不同的空间尺度上使用同一种计算模型,计算出的实体相关度无差别,与人类的直觉不符。譬如,在全国尺度上,北京和广州二个点的相对距离较远,相关度较小;在全球尺度上,北京和广州的相对距离较近,其相关度理应比全国尺度上的相关度大,而目前所有的实体相关度算法均未考虑空间尺度的影响,计算出的数值是固定的。

本研究拟基于地理实体的空间关系,考虑空间尺度对实体相关度的影响,建立适合不同空间尺度和不同实体类型的相关度算法体系,构建地理实体空间相关度问题的新解决方案。

2 研究思路

地理实体的空间相关度只考虑地理实体的几何属性。以线状的公路和河流为例,地理实体空间相关度只考虑公路和河流的空间位置、长度和形状

等,而不考虑其自然属性的差别(公路和河流),地理实体的空间相关度取值范围可以设为 $[0,1]$,其中0代表两个地理实体的几何属性完全不相关,1代表两个地理实体的几何属性完全等同。一条河流和一条公路在空间上完全重叠并等同,它们的空间相关度为1。

人类靠直觉感知地理实体的空间相关度,主要受到3个方面因素的影响:空间拓扑关系、空间度量关系、方位关系^[24]。拓扑关系用来描述2个空间对象是否相交、相接、包含,相等或者重叠等^[25]。空间度量关系包括面积、周长、长度、形状、密度、分布模式等,能被用来描述单个的实体或实体间的关系,例如2个面之间的距离或者重叠的面积等。空间方位关系指的是指2个地理实体的方向关系,如东、西、南、北等。人类感知地理实体空间相关度的过程受以下2个定理支配^[26]:

定理一(地理学第一定理):任何事物都是相关的,但是距离较近的事物相关性更高^[27];

定理二:在地理空间中,拓扑关系被认为是决定地理实体相关度最重要的信息,而度量关系,如距离和形状,只是相关度的改进因素^[28]。

由于空间方位关系对实体的相关度影响较小,本研究只考虑空间拓扑关系和空间度量关系来计算地理实体的相关度。空间度量关系常用的关系为距离关系,地理实体间的距离关系分为绝对距离关系和相对距离关系。绝对距离表示两个空间对象在某一度量单位下的距离;相对距离通过与第3个对象的比较,间接表示2个对象间的距离^[29]。本研究中把“空间尺度”定义为指定的地理空间范围,以空间尺度为第3方比较对象,以相对距离作为地理实体间距离关系的度量方法。

综上,本研究的具体思路如下:以空间尺度为第三方比较对象,以相对距离为地理实体间的距离关系度量方法,以拓扑关系作为空间相关度的决定因子,以度量关系作为空间相关度的改进因子,构建地理实体的空间相关度算法。

3 地理实体相关度计算模型

与自然语言表达中的地名不同,地理实体是通过语义清晰并且表达明确的方法对地理区域和实体的位置进行编码,并用能投影到地理空间中的地理坐标(如经纬度)或者二维平面坐标进行表达

的。因而计算地理实体的相关度,需要建立具有统一空间参考的基础地理数据库,以要素类的形式存储各种不同的实体,选取任意2个实体,考量其拓扑关系、度量关系以及其所属的空间尺度,利用这3个因子建立相关度的计算公式。

3.1 空间拓扑关系

拓扑关系是指拓扑变换下的不变量,它表达了空间对象之间固定的、不随观察角以及放大缩小变化的性质,是空间实体间一种最稳定的关系^[30]。为了区分不同的拓扑关系,用数学和逻辑的方法对拓扑关系进行描述。目前,得到公认的拓扑关系描述模型主要有二大类:Randell等^[31]提出的区域连接演算 RCC (region connection calculus) 理论和 Egenhofer等^[32]提出的求交模型。区域连接演算理论(简称RCC),是以Clarke^[33]提出的基于连接的个体演算理论为基础,由Randell等^[31]提出的一种用于描述空间关系的一阶逻辑理论。它包含RCC-8和RCC-5 2种基本拓扑关系集合。RCC理论所做的拓扑区分比较符合人类对空间拓扑关系的认知,缺点是拓扑关系的表达能力有限。

求交模型分为4-交模型和9-交模型。4-交模型是将2个二维的简单空间实体 X, Y 分为内部 X^0, Y^0 和边界 $\partial X, \partial Y$ 2个点集,可由边界与内部之间的关系确定 X, Y 之间的拓扑关系。9-交模型是Egenhofer等^[28]以点集拓扑学为理论依据,基于4-交模型提出的一种拓扑关系表达框架。它通过考察空间对象的内部、边界以及外部的相交情况,来区分各种拓扑关系。9-交模型可描述一切可能的空间几何对象间的空间关系,但存在冗余度大的问题,有多种拓扑关系并没有实际意义。

由于基于RCC理论和4-交模型的拓扑模型能表达拓扑关系种类有限,如RCC理论和4-交模型无法表达或者区分线和线的相交关系与重叠关系,本研究选取9-交模型作为实体拓扑关系的表达模型,并把一些类似或者在进行相关度计算具有相同影响的拓扑关系进行归纳合并,总结梳理出表1中的多点-多点、点-线、点-面、线-线、线-面、面-面的拓扑关系类型。

3.2 空间度量关系与空间尺度

空间度量关系能被用来描述单个地理实体或者地理实体之间的的关系。如重叠比例、空间距离等。空间尺度的概念比较模糊,其一般包括空间范围和粒

表1 实体空间拓扑关系表
Tab. 1 The topology relations between spatial entities

拓扑	点-点	点-线	线-点	点-面	面-点	线-线	线-面	面-线	面-面
相等 Equals		***	***	***	***		***	***	
包含 Contains		***		***			***		
被包含 Within			***		***			***	
重叠 Overlaps		***	***	***	***		***	***	
相交 Crosses	***	***	***	***	***				***
相接 Touches	***								
相离 Disjoins									

注:点一:● 点二:○ 线一:— 线二:—○ 面一:□ 面二:□○ ***代表不存在拓扑关系

度2种含义^[34],本研究中的空间尺度指的是包含所有要研究的地理实体的空间范围。相关定义如下:

空间重叠比例:几何实体重叠部分的面积/长度/数量与实体总面积/长度/数量的比值。

空间距离:空间实体主要涉及到点、线、面3种几何形态,点-点、点-面、面-面的距离指其几何中心的欧氏距离;点-线、线-面的距离指点和面的几何中心到线的最短距离;线-线的距离指线的最短距离。

空间尺度长度:指定的包含所有研究对象(地理实体)的地理空间范围的边界上任意两点间的最大距离。

耦合尺度的空间距离比:两个实体的空间距离与空间尺度长度的比例。

相关度基本权重:两空间实体在某一拓扑关系时专家所给予的相关度的最小值。

相关度控制权重:两空间实体在某一拓扑关系下,空间度量关系能决定的相关度变化的最大值。

3.3 空间相关度计算公式

根据定理二,拓扑关系决定地理实体的相关度的基本大小,而度量关系是实体间的相关度的调节因素。地理实体的相关度用式(1)计算:

$$R(A,B) = W_{Ti} + W_{TC} \times M_{AB} \quad (1)$$

式中: $R(A,B)$ 是实体A和B之间的空间相关度, W_{Ti} 是两空间实体在某一拓扑关系下,相关度的最小值,即相关度基本权重,其对实体的相关度大小具有决定作用。 W_{TC} 是相关度控制权重; M_{AB} 是空间实体的度量关系大小,不同类型的地理实体之间的度量关系计算公式不同,下面分别予以讨论。

(1) 点和点实体相关度

本研究以多点(multi-point)A和多点B之间的空间关系代表点和点实体的空间关系,如表1所示,多点和多点之间有5种拓扑关系,即相等、包含、被包含、重叠和相离。当两实体相等时,其相关度为1。本研究不对包含和被包含2种拓扑关系予以区分,当两点实体的空间拓扑关系为包含\被包含和叠加时,其相关度受2个因素的影响:①重叠点的数量占实体的点的总数的比例;②两实体之间的空间距离。当点和点之间的拓扑关系为相离时,其相关度只受空间距离的影响。考虑不同的空间尺度对相关度的影响,相关度计算公式如下

当A与B包含\被包含:

$$R(A,B) = W_{Ti} + W_{TC} \times \frac{P(A \cap B)}{\text{Max}(P(A), P(B))} \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (2)$$

当A与B叠加:

$$R(A,B) = W_{t2} + W_{t2C} \times \frac{P(A \cap B)}{\text{Max}(P(A), P(B))} \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (3)$$

当A与B相离:

$$R(A,B) = W_{t4} + W_{t4C} \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (4)$$

式中: W_n 是指定拓扑关系下的相关度基本权重; W_{nC} 是相应的相关度控制权重; $P(A \cap B)$ 代表重叠点的数量; $P(A), P(B)$ 分别代表多点A、B中点的数量; $D(A,B)$ 代表A和B的空间距离; $S(A,B)$ 代表空间尺度长度。

(2) 点和线实体相关度

如表1所示, 点和线或者线和点实体(A和B)之间存在3种拓扑关系, 即包含/被包含、相接和相离。当被包含时, 点和线的空间距离是0, 由于本研究只考虑实体的几何属性, 且认为线内部是均质的, 此时点和线实体的相关度只与线本身的长度有关。当点和线的拓扑关系是相接(Touches)时, 点和线实体的相关度也只与线本身的长度有关。当点和线的拓扑关系是相离(Disjoints)时, 点和线实体的相关度只与其空间距离有关, 考虑空间尺度的作用, 相关度的计算公式如下:

当A与B包含/被包含时,

$$R(A,B) = W_{t2} + W_{t2C} \times \left(1 - \frac{L(A,B)}{S(A,B) + L(A,B)}\right) \quad (5)$$

当A与B相接时,

$$R(A,B) = W_{t3} + W_{t3C} \times \left(1 - \frac{L(A,B)}{S(A,B) + L(A,B)}\right) \quad (6)$$

当A与B相离时,

$$R(A,B) = W_{t4} + W_{t4C} \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (7)$$

式中: W_n 是指定拓扑关系下相关度基本权重; W_{nC} 是相应的相关度控制权重; $L(A, B)$ 是指线的长度; $D(A, B)$ 代表A和B的空间距离; $S(A,B)$ 代表空间尺度长度。

(3) 点和面实体相关度

点和面或者面和点实体(A和B)也存在3种拓扑关系, 即包含/被包含、相接、相离。当点和面实体的拓扑关系为包含/被包含时, 空间实体相关度主要受面的面积大小和点在面的内部位置影响; 当点和面的拓扑关系为相接时, 实体的相关度主要受面的

面积和点与面的空间距离的影响; 当点和面的拓扑关系为相离时, 相关度主要受空间距离的影响。考虑空间尺度的作用, 其计算公式如下:

当A与B包含/被包含时,

$$R(A,B) = W_{t2} + W_{t2C} \times \left(1 - \frac{A(A,B)}{S_A(A,B) + A(A,B)}\right) \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (8)$$

当A与B相接时,

$$R(A,B) = W_{t3} + W_{t3C} \times \left(1 - \frac{A(A,B)}{S_A(A,B) + A(A,B)}\right) \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (9)$$

当A与B相离时,

$$R(A,B) = W_{t4} + W_{t4C} \times \left(1 - \frac{D(A,B)}{S(A,B)}\right) \quad (10)$$

式中: $A(A,B)$ 是指面的面积; $S_A(A,B)$ 是指空间尺度的面积; $D(A,B)$ 是实体A和B的空间距离; W_n 是相关度基本权重; W_{nC} 是控制权重。

(4) 线和线实体相关度

线和线(A和B)实体之间存在7种拓扑关系, 即相等、包含、被包含、叠加、相接、相交和相离。当线和线实体相等时, 其相关度为1; 当线和线实体是包含/被包含或者叠加关系时, 其相关度与空间重叠比例有关, 也与线自身的长度有关; 当线和线实体是相接或者相交关系时, 其几何相关度与相交的点数有关, 也与线自身的长度有关; 当线和线实体是相离关系时, 其相关度与空间距离有关。考虑空间尺度的作用, 其计算公式如下:

当A与B包含/被包含时,

$$R(A,B) = W_{t1} + W_{t1C} \times \left(\frac{L(A \cap B)}{\text{Max}(L(A), L(B))}\right) \times \left(1 - \frac{\text{Max}(L(A), L(B))}{S(A,B) + \text{Max}(L(A), L(B))}\right) \quad (11)$$

当A与B叠加时,

$$R(A,B) = W_{t2} + W_{t2C} \times \left(\frac{L(A \cap B)}{\text{Max}(L(A), L(B))}\right) \times \left(1 - \frac{\text{Max}(L(A), L(B))}{S(A,B) + \text{Max}(L(A), L(B))}\right) \quad (12)$$

当A与B相交或相接时,

$$R(A,B) = W_{t3} + W_{t3c} \times \frac{P(A \cap B)}{1 + P(A \cap B)} \times \left(1 - \frac{\text{Max}(L(A), L(B))}{S(A,B) + \text{Max}(L(A), L(B))} \right) \quad (13)$$

当A与B相离时,

$$R(A,B) = W_{t4} + W_{t4c} \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (14)$$

式中: $L(A \cap B)$ 是指A和B空间重叠长度; $L(A), L(B)$ 分别指A和B的长度; $S(A, B)$ 是指空间尺度长度; $P(A \cap B)$ 是指A和B相交或者相接的点的数量; $D(A, B)$ 指A和B的空间距离; W_n 是相关度基本权重; W_{nc} 是控制权重。

(5) 线和面实体相关度

线和面或者面和线实体(A和B)之间存在4种拓扑关系,即包含\被包含、相交、相接和相离。当线和面实体是包含\被包含和相交关系时,空间相关度与线的长度和面的周长比例有关,也与线和面的空间距离有关;当线和面实体是相接关系时,空间相关度与线的长度和空间距离有关;当线和面实体是相离关系时,空间相关度与空间距离有关,考虑空间尺度的作用,计算公式如下:

当A与B包含\被包含时,

$$R(A,B) = W_{t2} + W_{t2c} \times \left(\frac{L(A \cap B)}{\text{Max}(L(A), L(B))} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (15)$$

当A与B相交时,

$$R(A,B) = W_{t3} + W_{t3c} \times \left(\frac{L(A \cap B)}{\text{Max}(L(A), L(B))} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (16)$$

当A与B相接时,

$$R(A,B) = W_{t5} + W_{t5c} \times \left(1 - \frac{L(A,B)}{S(A,B) + L(A,B)} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (17)$$

当A与B相离时,

$$R(A,B) = W_{t6} + W_{t6c} \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (18)$$

式中: $L(A \cap B)$ 是指线和面叠加的线的长度; $L(A), L(B)$ 分别指线的长度和面的周长; $D(A, B)$ 是指A和B

的空间距离; $S(A, B)$ 是指空间尺度长度; $L(A, B)$ 是指线的长度; W_n 是基本权重; W_{nc} 是控制权重。

(6) 面和面实体相关度

面和面(A和B)实体存在6种拓扑关系,即相等、包含、被包含、叠加、相接和相离。当面和面实体相等时,空间相关度为1,当面和面是包含或被包含和叠加关系时,其空间相关度主要与空间重叠比例相关,也受到空间距离的影响;当面和面是相接和相离关系时,空间相关度主要受空间距离影响,考虑空间尺度的作用,其计算公式如下:

当A与B包含\被包含时,

$$R(A,B) = W_{t1} + W_{t1c} \times \left(\frac{A(A \cap B)}{\text{Max}(A(A), A(B))} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (19)$$

当A与B叠加时,

$$R(A,B) = W_{t2} + W_{t2c} \times \left(\frac{A(A \cap B)}{\text{Max}(A(A), A(B))} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (20)$$

当A与B相接时,

$$R(A,B) = W_{t3} + W_{t3c} \times \left(\frac{L(A \cap B)}{\text{Max}(L(A), L(B))} \right) \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (21)$$

当A与B相离时,

$$R(A,B) = W_{t4} + W_{t4c} \times \left(1 - \frac{D(A,B)}{S(A,B)} \right) \quad (22)$$

式中: $A(A \cap B)$ 是A与B重叠的面积; $A(A), A(B)$ 是指A和B的面积; $L(A \cap B)$ 是A和B相接时重叠的边长; $L(A), L(B)$ 是A和B的周长; $D(A, B)$ 是A和B的空间距离; $S(A, B)$ 是指空间尺度长度; W_n 是相关度基本权重; W_{nc} 是控制权重。

权重系数可以由多位专家打分求取平均值给出。专家打分需遵循以下基本原则:①对于点和点、点和线、点和面、面和面类型的实体,包含\被包含关系下的实体相关度不得大于相等关系(如果存在,下同)下的实体相关度;叠加关系下的实体相关度不得大于包含\被包含关系下的实体相关度;相接关系下的实体相关度不得大于叠加关系下实体相关度;相离关系下实体相关度不得大于相接关系下的实体相关度。②基本权重和控制权重之间的关

系。具体如下:

(1)当拓扑关系为“包含/被包含”时,“包含/被包含”的极限为“外部、边界、内部的点完全相同”,即相等,也就是当度量关系的值为1($M_{AB}=1$)时, $W_{T1}+W_{T1C}=1$;

(2)当拓扑关系为“重叠”时,“重叠”的极限是“包含/被包含”,也就是当 $M_{AB}=1$ 时, $W_{T1}=(W_{T2}+W_{T2C})$;

(3)当拓扑关系为“相接”时,“相接”的极限为“具有共同的内部点”,即叠加,也就是当 $M_{AB}=1$ 时, $W_{T2}=(W_{T3}+W_{T3C})$;

(4)当拓扑关系为“相离”时,“相离”的极限为“具有共同的边界点,内部点不相交”,即相接,也就是当 $M_{AB}=1$ 时, $W_{T3}=(W_{T4}+W_{T4C})>W_{T4}$ (权重系数均为非负实数);

对于线和线实体,把线和线的相交关系与相接关系视为等同,其权重系数关系也符合以上原则。

同理,对于线和面实体类型:相交关系下的实体相关度不得大于包含/被包含关系下的相关度,相接关系下的实体相关度不得大于相交关系下的实体相关度,相离关系下的实体相关度不得大于相接关系下的实体相关度。权重系数存在关系: $1=(W_{T1}+W_{T1C})>W_{T1}=(W_{T2}+W_{T2C})>W_{T2}=(W_{T3}+W_{T3C})>W_{T3}=(W_{T5}+W_{T5C})>W_{T5}=(W_{T6}+W_{T6C})>W_{T6}$;

因此,地学专家只需根据以上原理给出各种拓扑关系下地理实体空间相关度的基本权重值,相关度的控制权重值可由权重值间的关系求取。

空间尺度可以根据需要指定,其最小范围必须包含要研究的所有空间实体。例如研究河南省范围内的空间实体相关度,空间尺度最小范围需指定为河南省,但也可以指定为中国或者亚洲等。

4 结果与分析

表2 本研究权重系数取值表

Tab. 2 The weights of the algorithm in the case

基本权重	值	控制权重	值
W_{T1}	0.667	W_{T1C}	0.333
W_{T2}	0.5	W_{T2C}	0.167
W_{T3}	0.333	W_{T3C}	0.167
W_{T4}	0	W_{T4C}	0.333
W_{T5}	0.167	W_{T5C}	0.166
W_{T6}	0	W_{T6C}	0.167

基于以上不同类型的实体间相关度的计算方法,本研究邀请地学专家对不同拓扑关系下的基本

权重进行评分,获取基本权重和控制权重的取值(表2),然后选取区域和全国两个尺度,分别计算地理实体在不同空间尺度上的相关度。

此外,文献[19]中提出了一种基于4-交模型的点、线、面实体之间的拓扑关系和度量关系的地理实体空间相关度的算法,这种算法可以作为传统的基于绝对距离的地理实体空间相关度算法的代表。本研究用本文提出的算法和文献[19]中的算法做对比分析,比较2种方法优劣。

如图1所示,选取河南省范围内的村庄,道路,行政区以及经济概念区实体,构成不同类型和不同拓扑关系的点线面组,分别以河南省和全国行政区划范围为空间尺度,应用3部分的公式,同时使用文献[19]中提出的方法计算地理实体的空间相关度。

经计算得到的实体相关度如表3所示。由表3可知,本算法能以拓扑关系为主要区分依据,计算点、线、面实体之间的空间相关度。相同拓扑关系下,空间距离越近的实体,其相关度越高。另外,在不同的空间尺度下,2个实体的相关度是不同的,譬如焦作市陈家沟村和信阳市郝堂村,在区域尺度上,其相关度为0.134,在全国尺度上其相关度为0.311。随着空间尺度的增大,两实体的相关度也是增大的(相同的空间实体除外),这与人类通常的直觉是相符的。

传统的基于绝对距离的地理实体空间相关度算法在不同的空间尺度上的计算结果不变,而且其能处理的地理实体的类型有限(譬如,不能计算多点和多点类型实体的相关度),其能处理的拓扑关系种类也有限(譬如,不能计算线和线相交关系下的实体相关度)。此外,传统地理实体空间相关度算法计算出的地理实体的相关度的数值的连续性差,在拓扑关系渐变的情况下,传统算法计算的实体相关度数值出现跳跃性的变化,与人类通常的直觉不符(如拓扑关系为相接的洛阳市和郑州市的空间相关度为0.357,而拓扑关系为相离的洛阳市和漯河市的空间相关度变为 3.47×10^{-6} ,相关度跳跃了 10^5 个数量级)。

5 应用案例

本研究在多个研究领域得到应用,如地理空间数据关联、地理空间数据发现等,本文以地理空间数据检索为例,应用本研究提出的算法。

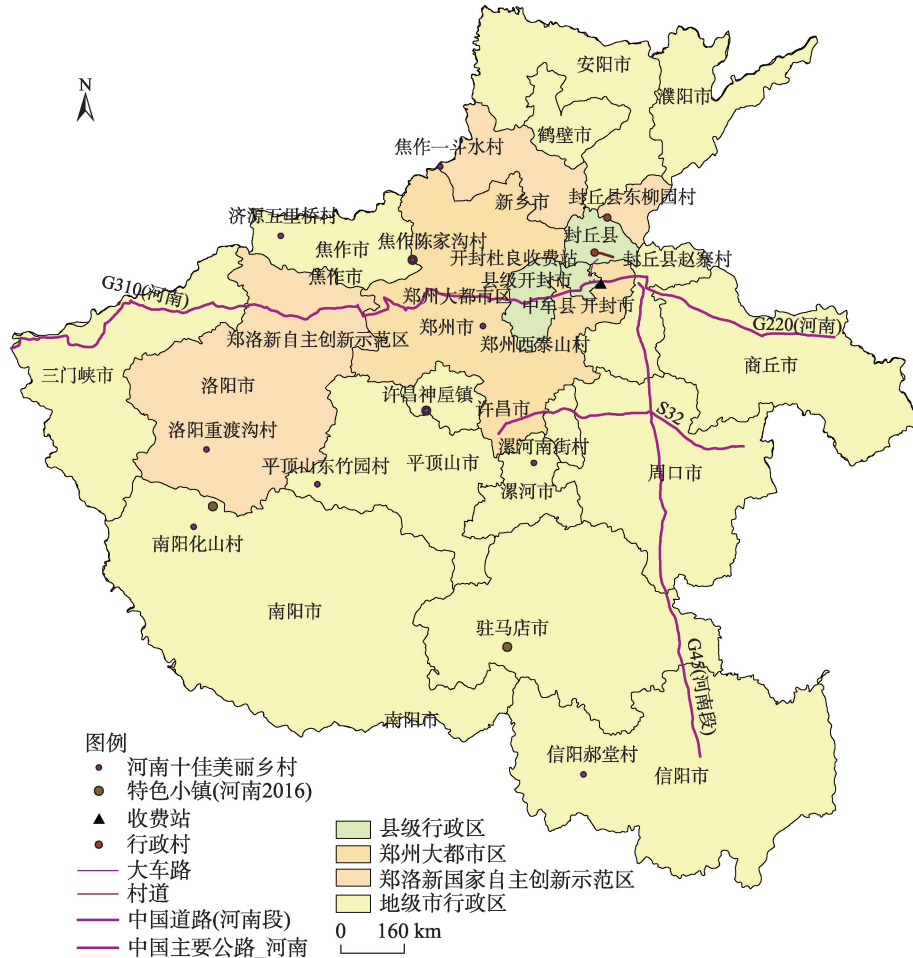


图1 实体分布示意图

Fig. 1 Distribution of spatial entities

国家科技基础条件平台——地球系统科学数据共享平台(<http://www.geodata.cn/>)是中国地球科学领域唯一的国家级科学数据共享平台,该平台拥有丰富的地学数据,并采用元数据的方式对其进行管理。用户通过检索元数据而查找所需的数据。该平台的元数据以ISO19100地理信息系类标准为基础^[19],每条地理空间元数据包含标题、主题词、空间位置(范围)、时间、学科类别等30余种信息。本研究从地球系统科学数据共享平台上提取85条元数据作为实验数据,采用不同的数据检索方法测试检索的准确率和召回率。

一般情况下,利用地理信息检索(Geographic Information Retrieval, GIR)检索网页文档中的地理信息时,需要同时从主题和空间位置两个方面评估文档和用户所需文档间的相关性^[26],本设计如下算法评估用户输入的主题词和空间位置与科学数据元数据的主题词和空间位置的相关度,从而实现数据检索。

$$t = w_1 \cdot x + w_2 \cdot y \quad (23)$$

式中: t 为用户输入的检索词和科学数据元数据的匹配度; x 为用户输入的主题词和科学数据元数据的主题词项的相关度; y 为用户输入的空间位置和科学数据元数据的空间位置项的相关度; w_1 和 w_2 为权重系数,其和为1,可由权重评估-层次分析法^[35]计算获取。

权重评估-层次分析法详细的步骤如下:首先需要建立对上一层目标有影响的所有因素两两之间的相对重要性比较矩阵。然后请领域专家用1-9给出影响因素两两之间的相对重要性分数。最后求比较矩阵的特征向量,经归一化处理即是各影响因素的权重值。当影响因素超过2个时,比较矩阵需要做一致性检验。一致性检验通过的标准是CR值小于0.1。

准确率(Precision)是检索到的相关数据与所有检索结果数量的比率,召回率(Recall)是检索到的

表3 不同尺度上空间相关度对比表

Tab. 3 The contrast of spatial relevance in different scales and methods

实体一	实体二	类型	拓扑关系	相关度 (全国尺度)	相关度 (区域尺度)	相关度 (传统算法)
漯河南街村	漯河南街村	点-点	相等	1.000	1.000	1.000
河南特色小镇(2016)	许昌神垕镇	多点-点	包含/被包含	0.749	0.744	*
河南十佳美丽乡村(2014)	河南特色小镇(2016)	多点-多点	重叠	0.533	0.531	*
焦作陈家沟村	信阳郝堂村	点-点	相离	0.311	0.134	9.760×10^{-7}
开封杜良收费站	G220(河南)	点-线	包含/被包含	0.659	0.617	1.000
封丘县赵寨村村道	封丘县赵寨村	点-线	相接	0.499	0.496	*
郑州西泰山村	封丘县赵寨村村道	点-线	相离	0.327	0.284	3.965×10^{-6}
南阳化山村	河南省南阳市	点-面	包含/被包含	0.664	0.624	0.667
新乡市封丘县	封丘县东柳园村	点-面	相接	0.499	0.494	*
漯河南街村	河南省许昌市	点-面	相离	0.330	0.308	2.650×10^{-5}
东郑线	东郑线	线-线	相等	1.000	1.000	1.000
G220(河南)	东郑线	线-线	包含/被包含	0.721	0.707	0.528
023乡道	齐边线	线-线	重叠	0.50209	0.50208	0.502
G220(河南)	G310(河南)	线-线	相接	0.411	0.383	0.333
S32	G45(河南段)	线-线	相交	0.411	0.387	*
G220(河南)	S32	线-线	相离	0.328	0.292	4.840×10^{-6}
齐边线	县级开封市	线-面	包含/被包含	0.50175	0.50174	0.670
G220(河南)	河南省开封市	线-面	相交	0.357	0.356	0.385
开柳公路	县级开封市	线-面	相接	0.333	0.328	0.333
河南省漯河市	G220(河南)	线-面	相离	0.163	0.133	4.099×10^{-6}
郑州大都市区	郑州大都市区	面-面	相等	1.000	1.000	1.000
河南省郑州市	郑州市中牟县	面-面	包含/被包含	0.728	0.723	0.531
郑州大都市区	郑洛新自主创新示范区	面-面	重叠	0.553	0.548	0.600
河南省洛阳市	河南省郑州市	面-面	相接	0.495	0.461	0.357
河南省洛阳市	河南省漯河市	面-面	相离	0.320	0.221	3.470×10^{-6}
河南省洛阳市	河南省信阳市	面-面	相离	0.310	0.122	1.810×10^{-6}

备注: *代表无法处理此种类型的拓扑关系;传统算法指的是文献[19]中提出的算法

相关数据与数据库中所有相关数据的数量比率。 $F1$ -measure是综合准确率和召回率的数据检索能力评价指标。其定义如下:

$$F1 - measure = \frac{2 \cdot P \cdot R}{P + R} \quad (24)$$

式中: P 是指检索的准确率; R 是指检索的召回率; $F1$ -measure值越高说明检索方法越有效。

本研究拟用本文提出的算法和传统的关键词匹配算法分别实现对样本数据的检索,以 $F1$ -measure值度量算法的性能。

传统的关键词匹配算法在进行地理空间数据检索时,比较用户输入的主题词、空间位置关键词与地理科学数据元数据的主题词、空间位置关键词,若二者匹配,则相应的 x 值和 y 值为1,若不匹配,则 x 值和 y 值为0,也就 t 的取值为0或者1。使用本研究的算法进行地理空间数据检索时,当用户输入的主题词与元数据中主题词匹配时 x 为1,否则为0;计算 y 值时,把用户输入的空间位置关键词和元数据

的空间位置描述项与基础地理数据库中的地理实体一一映射,计算地理实体间的相关度作为 y 值。

基于以上实验方案,分别开发基于以上2种方法的地理空间数据检索系统,并利用权重评估-层次分析法计算 w_1 和 w_2 的值,分别为0.667和0.333,向系统中输入关键词,例如“土地利用 上海市”,点击检索按钮,数据检索系统分根据2种算法计算地理空间数据的主题词以及空间范围和关键词中的主题词和空间范围的相关度,从而获取匹配度。规定传统关键词匹配算法计算的匹配度为1的地理空间数据为检索结果,使用本研究提出的算法计算的匹配度大于或者等于0.889(即: $0.667 \times 1 + 0.333 \times 0.667 = 0.889$)的地理空间数据为检索结果(即认为元数据主题词包含用户输入的主题词,且空间范围包含或者等于用户指定的空间范围的地理空间数据为检索结果)。2种方法检索结果如表4、5所示。

根据本案例中用户输入检索词的意图,以检索到上海市或者包含上海市的土地利用数据为查找

表4 关键词匹配法检索结果
Tab. 4 The retrieved results of the keywords
matching method

数据名称	匹配度
上海市 1:10 万土地利用数据(2008 年)	1.000
上海市 1:10 万土地利用数据(1980s)	1.000
上海市 1:10 万土地利用数据(1995 年)	1.000

表5 本研究提出算法检索结果
Tab. 5 The retrieved result of our method

数据名称	匹配度
上海市 1:10 万土地利用数据(2008 年)	1.000
上海市 1:10 万土地利用数据(1980s)	1.000
上海市 1:10 万土地利用数据(1995 年)	1.000
长三角 1:10 万土地利用数据(2005 年)	0.892
长三角地区 1980s、1995 年、2000 年 1:25 万土地利用数据集	0.892
中国分省土地利用面积数据(1980s、1995 年、2005 年)	0.889
中国 1:100 万土地利用区划(1996 年)	0.889
中国地区土地利用/土地覆盖数据集	0.889
中国 1 km 网格土地利用数据(1980s、1995 年、2000 年)	0.889

表6 2种检索方法的准确率、召回率和 *F1-measure* 值
Tab. 6 The contrast of precision, recall and *F1-measure*
of the two methods

方法名称	准确率 /%	召回率 /%	<i>F1-measure</i> /%
方法一(使用本研究提出的算法)	77.8	100	87.5
方法二(关键词匹配算法)	100	42.9	60.0

正确,2种方法的准确率、召回率以及 *F1-measure* 值分别如表6所示。

分析表6可以发现,本研究提出的算法能在保证较高准确率的前提下,大大提高科学数据检索的召回率,能发现包含位置检索关键词所指定的地理空间范围的科学数据,具有较强的数据发现能力,较大幅度的提高了地理空间数据检索的 *F1-measure* 值。此外,本算法能对检索结果按匹配度进行排序,有利于用户快速获取所需的数据。

6 结论与讨论

本研究针对现有地理空间实体相关度算法存在的问题,建立了一套适合不同空间尺度、不同实体类型和不同拓扑关系的实体相关度计算方法。从区域和全国2个尺度上,验证了本方法的适用性。并以地理空间元数据检索为例,应用了本研究的算法,证明了本研究的实用价值。

本研究的创新点为:①发现了地理实体空间相

关度的尺度依赖性,进而考虑空间尺度的作用,建立耦合尺度的实体相关度计算模型。②针对不同类型的空间实体和实体间不同的拓扑关系,建立全面系统的相关度计算方法。

本算法计算出的地理实体相关度数值分布均匀,区分度良好,这对于基于地理实体相关度的一些数据挖掘算法,如人工神经网络算法模拟精度和计算效能的提升具有重要意义,因为人工神经网络对连续型数值模拟的结果比较好,对跳跃型的数值模拟精度比较差;本算法实现简单,只需建立地理空间基础数据库就可以实现。此外,本算法具可拓展性。如在某些情况下,地理空间实体的相关度要求具有方向性,本算法作轻微改动,如把 *Max* 函数改成取某一值即可满足需求。随着大数据和人工智能时代的来临,本算法的作用会逐步凸现。

本算法也存在以下问题:①空间尺度范围的确定问题:本算法要求空间尺度范围必须包含所有的要参与计算相关度的地理实体,但是在实际应用过程中,地理实体的数量是不确定的或者是变化的,这增加了空间尺度范围选择的难度,下一步将根据研究的实际情况对空间尺度范围进行预估,坚持从大原则选择空间尺度。②本算法只能对基础地理数据库中已有的空间实体评估相关度。③实际应用中,通常要求实现文字和基础地理空间数据库中的地理实体进行映射,由于文字表达的复杂性,通常难以实现自动化映射,这在很大程度上制约了本算法的应用,这是下一步要努力的方向。

参考文献(References):

- [1] Hill L L. Access to Geographic concepts in online bibliographic files: Effectiveness of current practices and the potential of a graphic interface[D]. Pittsburgh: University of Pittsburgh, 1990.
- [2] 李霖,应申.空间尺度基础性问题研究[J].武汉大学学报·信息科学版,2005,30(3):199-203. [Li L, Ying S. Fundamental problems on spatial scales[J]. Geomatics and Information Science of Wuhan University, 2005,30(3):199-203.]
- [3] Robinson W S. Ecological correlations and the behavior of individuals[J]. International Journal of Epidemiology, 2011,40(4):351-357.
- [4] Openshaw S, Taylor P J. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem[M]. London: Pion, 1979:127-144.
- [5] Goodchild M F. The Aggregation Problem in Location-Allocation[J]. Geographical Analysis, 1979,11(3):240-255.
- [6] Dudley G. Modifiable areal units and human geographical inquiry: An empirical investigation, in Department of Geography[D]. Ontario: University of Waterloo, 1991.

- [7] Marceau D J, Howarth P J, Gratton D J. Remote sensing and the measurement of geographical entities in a forested environment: The scale and spatial aggregation problem[J]. *Remote Sensing of Environment*, 1994,49(2):93-104.
- [8] Tversky A. Features of similarity[J]. *Psychological Review*, 1977,84(4):290-302.
- [9] Medin D L, Goldstone R L, Gentner D. Respects for Similarity[J]. *Psychological Review*, 1993,100(2):254-278.
- [10] Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2003,15(2):442-456.
- [11] Rodriguez M A, Egenhofer M J. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure[J]. *International Journal of Geographical Information Science*, 2004,18(3):229-256.
- [12] Egenhofer M J, Franzosa R D. Point-set topological spatial relations[J]. *International Journal of Geographical Information Systems*, 1991,5(2):161-174.
- [13] Walker D R F, Newman I A, Medyckyj-Scott D J, et al. A system for identifying datasets for GIS users[J]. *International Journal of Geographical Information Systems*, 1992,6(6):511-527.
- [14] Janée G, Frew J. Spatial Search, Ranking, and Interoperability[C]. *Workshop on Geographic Information Retrieval*, 2004.
- [15] Beard K, Sharma V. Multidimensional ranking for data in digital spatial libraries[J]. *International Journal on Digital Libraries*, 1997,1(2):153-160.
- [16] Li B, Fonseca F. TDD-A comprehensive model for qualitative spatial similarity assessment[J]. *Spatial Cognition and Computation*, 2006,6(1):31-62.
- [17] Frontiera P L. A probabilistic approach to spatial ranking for geographic information retrieval[D]. Berkeley: University of California, 2004.
- [18] 刘家骏,李浩然,钟翔,等.地理信息检索中空间相似性度量的一种模糊方法[J].*地理与地理信息科学*,2015,31(4):38-42. [Liu J J, Li H R, Zhong X, et al. A fuzzy method to measure spatial similarity in geographic information retrieval[J]. *Geography and Geo-Information Science*, 2015,31(4):38-42.]
- [19] 赵红伟,诸云强,杨宏伟,等.地理空间数据本质特征语义相关度计算模型[J].*地理研究*,2016,35(1):58-70. [Zhao H W, Zhu Y Q, Yang H W, et al. The computation model of semantic relevancy on essential features of geospatial data[J]. *Geographical Research*, 2016,35(1):58-70.]
- [20] 李红梅,翟亮,朱焜.基于本体的地理空间实体类型语义相似度计算模型的研究[J].*测绘科学*,2009,34(2):12-14. [Li H M, Zhai L, Zhu H. A study on calculative modeling of semantic similarities for geospatial entity classes based on ontology[J]. *Science of Surveying and Mapping*, 2009,34(2):12-14.]
- [21] 杨娜娜,张青年,牛继强.基于本体结构的地理空间实体语义相似度计算模型[J].*测绘科学*,2015,40(3):107-111. [Yang N N, Zhang Q N, Niu J Q. Computational model of geospatial semantic similarity based on ontology structure[J]. *Science of Surveying and Mapping*, 2015,40(3):107-111.]
- [22] Janowicz K. Towards a Similarity-Based Identity Assumption Service for Historical Places[C]. *The 4th International Conference on Geographic Information Science (GIScience)*, 2006.
- [23] Janowicz K. Sim-DL: Towards a semantic similarity measurement theory for the description logic ALCN^R in geographic information retrieval[C]. *The 2nd international workshop on semantic-based geographical information systems (SeBGIS06)*, 2006.
- [24] Bruns H T, Egenhofer M J. Similarity of spatial scenes [C]. *Seventh International Symposium on Spatial Data Handling*, 1996.
- [25] Clementini E, Felice P D. A Model for Representing Topological Relationships between Complex Geometric Features in Spatial Databases[J]. *Information Sciences*, 1996, 90:121-136.
- [26] Frontiera P, Larson R, Radke J. A comparison of geometric approaches to assessing spatial similarity for GIR[J]. *International Journal of Geographical Information Science*, 2008,22(3):337-360.
- [27] TOBLER W R. A computer movie simulating urban growth in the detroit region[J]. *Economic Geography*, 1970,46:234-240.
- [28] Egenhofer M J, Herring J R. Categorizing binary topological relations between regions, lines, and points in geographic databases[J]. *Statistics and Information Forum*, 1990.
- [29] 赵红伟,诸云强,侯志伟,等.地理空间元数据关联网络的构建[J].*地理科学*,2016,36(8):1180-1189. [Zhao H W, Zhu Y Q, Hou Z W, et al. Construction of geospatial meta-data association network[J]. *Scientia Geographica Sinica*, 2016,36(8):1180-1189.]
- [30] 黄茂军.地理本体的形式化表达机制及其在地图服务中的应用研究[D].武汉:武汉大学,2005. [Huang M J. Study on formal representation of geographic ontology and its application in map services[D]. WuHan: Wuhan University, 2005.]
- [31] Randell D A, Cui Z, Cohn A G. A spatial logic based on regions and connection[C]. *The 3rd International Conference on Knowledge Representation and Reasoning*, 1992.
- [32] Egenhofer M J. A Formal Definition of Binary Topological Relationships[C]. *Foundations of Data Organization and Algorithms 3rd International Conference*, 1989.
- [33] Clark B L. Individuals and Points[J]. *Notre Dame Journal of Formal Logic*, 1985,26(1):61-75.
- [34] 马蔚纯,赵海君,李莉,等.区域规划环境评价的空间尺度效应——对上海高桥镇和浦东新区的案例研究[J].*地理科学进展*,2015,34(6):739-748. [Ma W C, Zhao H J, Li L, et al. Spatial scale effects of environmental impact assessment of regional planning: The Gaoqiao Town and Pudong New District cases in Shanghai, China[J]. *Progress in Geography*, 2015,34(6):739-748.]
- [35] Saaty L T. How to make a decision: The analytic hierarchy process[J]. *European Journal of Operational Research*, 1990,48(1):9-26.