

# 地理空间数据本质特征语义相关度计算模型

赵红伟<sup>1,2</sup>, 诸云强<sup>1,3</sup>, 杨宏伟<sup>4</sup>, 罗 侃<sup>1,2</sup>

(1. 中国科学院地理科学与资源研究所, 中国科学院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; 3. 江苏省地理信息资源开发与利用协同创新中心, 南京 210023;

4. 中国石油规划总院, 北京 100083)

**摘要:** 关联数据是跨网域整合多源异构地理空间数据的有效方式, 语义丰富的关联是准确、快速发现目标数据的关键。根据地理空间数据在空间、时间、内容上的语义关系, 提出地理空间数据本质特征语义相关度计算模型。通过构建本质特征的关联指标体系, 分层次逐级计算地理空间数据的语义相关度。与传统的语义相关度计算方式不同, 以地理元数据为语料库, 充分考虑地理空间数据的特点及空间、时间、内容在检索中不同的重要程度, 分别采用几何运算、数值运算、词语语义相似度计算和类别层次相关度计算的方式, 构建地理空间数据的语义相关度计算模型。该模型具有构建简单、适用于多源异构数据、充分结合了数学运算和专家经验知识等特点。实验表明: 模型能够有效地计算地理空间数据本质特征的语义相关度, 并具备一定的扩展性。

**关键词:** 地理空间数据; 空间特征; 时间特征; 内容特征; 语义相关度

DOI: 10.11821/dljy201601006

## 1 引言

随着3S技术的发展, 地理空间数据的内容日益丰富、来源越来越广泛、存储格式多样化。传统基于关键词的数据检索方式, 很难满足用户需求。如“江苏省1:10万土地利用数据”(A)与“无锡市1:100万草地覆被数据”(B)两条数据, 如果用户需要江苏省土地利用数据, 通过关键字“江苏省”、“土地利用”等查询, 只能查询到数据集A而不能查询到数据集B, 但是, 数据集B在空间上(无锡市)属于江苏省, 在内容上(草地覆被)是土地利用的一种。因此, 科研人员虽处于“信息的海洋”, 却常面临“信息泛滥、知识匮乏”的困境<sup>[1]</sup>。在大数据环境下, 如何准确快速地发现数据, 成为地理空间数据共享应用面临的关键问题。关联数据的提出<sup>[2]</sup>为这一问题的解决提供了最佳实践。通过建立数据集A和数据集B之间的语义关联来实现数据的语义搜索。然而, 仅仅依靠语义关联还不能够解决检索中的排序问题, 因此还需要计算数据集之间的语义相关度。

语义相关度不仅包含词汇间的相似性, 而且包括词汇之间根据各种语义关系具有的关联性<sup>[3]</sup>, 例如: 对于“江苏省”和“无锡市”这两个词而言, 虽然两者词汇相似性非常

收稿日期: 2015-06-21; 修订日期: 2015-11-18

基金项目: 国家自然科学基金项目(41371381); 科技基础性工作专项项目(2013FY110900); 中国科学院科研信息化“科技领域云”项目; 国家重大科学仪器设备开发专项(2012YQ06002704); 云南省科技计划项目(2012CA021)

作者简介: 赵红伟(1988-), 女, 山东聊城人, 博士研究生, 主要研究方向为空间数据挖掘, 地理空间关联数据。  
E-mail: zhaohw.10s@igsrr.ac.cn

通讯作者: 诸云强(1977-), 男, 江西广丰人, 博士, 研究员, 主要研究方向为地学数据共享关键技术、地学科研信息化环境、资源环境信息系统。E-mail: zhuyq@lreis.ac.cn

低，但其空间相关性却很高（无锡市属于江苏省）。除了空间关系，地理空间数据集之间还具有多种语义关系如属性类别关系、时间关系等。目前，国内外学者主要通过地理本体<sup>[4-7]</sup>、地名词典<sup>[8,9]</sup>、地理语义目录<sup>[10]</sup>等方式构建地理语义关系来辅助计算地理空间数据的语义相关性。然而，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；地名词典、地理语义目录不能够表达地理空间特征的拓扑关系、度量关系等。因此，以地理空间元数据为语料库，选取用户检索中主要关注的空间、时间、内容三个特征，构建地理空间数据本质特征语义相关度计算模型。该模型通过建立空间、时间、内容三个维度的关联指标体系，并根据不同维度的语义特点，利用地理空间元数据提供的语义信息分别计算语义相关度，进而实现地理空间数据之间的语义关联，支持地理空间数据的精准搜索和排序。

2 地理空间数据本质特征语义关联指标体系

内容、空间、时间是多源地理空间数据的本质特征，每个特征的语义关联都是由多种语义关系构成的，这些语义关系在不同程度上影响地理空间数据的语义相关度。通过对本质特征的分析建立地理空间数据本质特征三级关联指标体系（表1）。每个指标的权重由专家打分确定。空间度量关系和时间度量关系如重叠比例、空间距离等，一方面可以辅助量化空间拓扑关系，另一方面可提高空间语义相关度计算的准确性。

（1）内容语义相关度，用Fsem表示。指地理空间数据集所表达的内容信息的相关程度。一部分取决于数据内容描述词汇的相似性，如土地覆被、土地利用的语义相似性很大；另一部分取决于内容所属的类别相关性，如果园与农用地词汇相似性非常低，但果园属于农用地，在类别上有一定的相关性。两部分分别用内容词汇语义相似度（F1）和类别相关度（F2）两个二级指标表示。

类别相关度包含类别层次相关度和类别相关比例两个三级指标。类别层次相关度是指在同一分类体系中，两个数据所属类别的相关程度。在某些情况下，同一地理空间数据集会同时属于多个类别，如“杭嘉湖地区1:10万土地利用、水资源与水利工程（2000年）”数据集既属于土地资源类，又属于水资源类。因此，应用类别相关比例这一指标来度量多类别数据集之间的相关度。

（2）空间语义相关度，用Ssem表示。指地理空间数据所表达的空间实体间的空间关联程度，包括拓扑关系、度量关系和方位关系。方位关系在检索排序中的影响较小，采用拓扑关系和度量关系计算空间语义相关度。

空间拓扑关系主要包括相交、包含、相接等。同一拓扑关系，如包含关系，多个空间对象的距离、面积不同，其语义相关度也不同。如图1所示的面一面之间的拓扑关系，如采用空间实体几何中心的欧式距离，B同时包含A、C、D，距离（CB）=距离（AB）、面积（A）>面积（C），一般认为空间相

表1 地理空间数据本质特征语义关联指标体系		
Tab. 1 The semantic relevance indices system of essential features of geospatial data		
一级指标	二级指标	三级指标
内容语义相关度 (Fsem)	内容词汇语义相似度(F <sub>1</sub> )	
	类别相关度(F <sub>2</sub> )	类别层次相关度(F <sub>21</sub> ) 类别相关比例(F <sub>22</sub> )
	空间语义相关度 (Ssem)	空间拓扑关系相关度 (S <sub>1</sub> ) 空间度量关系相关度(S <sub>2</sub> ) 空间重叠比例(S <sub>21</sub> ) 空间距离(S <sub>22</sub> )
时间语义相关度 (Tsem)	时间拓扑关系相关度 (T <sub>1</sub> )	
	时间度量关系相关度 (T <sub>2</sub> )	时间重叠比例 (T <sub>21</sub> ) 时间距离 (T <sub>22</sub> )

关度 (AB) > 空间相关度 (CB); 距离 (DB) > 距离 (CB)、面积 (D) = 面积 (C), 根据地理学第一定律<sup>[1]</sup>距离越近的两个事物相关性越紧密, 则空间相关度 (CB) > 空间相关度 (DB)。因此, 空间语义关系在考虑空间拓扑关系的基础上, 应进一步考虑空间重叠比例和空间距离等度量关系。

(3) 时间语义相关度, 用 Tsem 表示。指地理空间数据所表达时间 (对于监测类的数据, 可用采集时间代替) 的关联程度。与空间语义相关度相似, 时间语义相关度包括时间拓扑关系 (T<sub>1</sub>) 和时间度量关系 (T<sub>2</sub>) 两个二级指标。时间度量关系由时间重叠比例和时间距离构成。

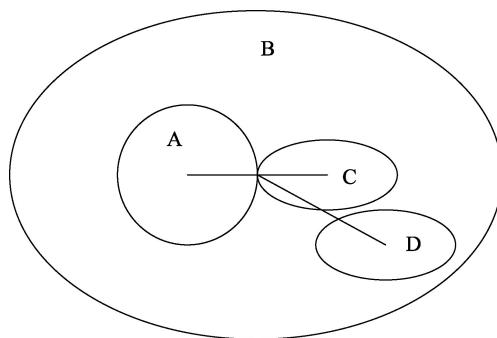


图1 空间拓扑关系和空间度量关系示意图  
Fig. 1 The diagram of spatial topological relations and measure relationship

### 3 地理空间数据语义关联模型

#### 3.1 语义相关度计算模型

地理空间数据语义关联度由三个一级指标直接计算得出, 如式 (1) 所示。每个一级指标由相应的二级、三级指标逐层计算得到。

$$GeoSem = W_F Fsem + W_S Ssem + W_T Tsem \quad (1)$$

式中:  $GeoSem$  为地理空间数据语义相似度;  $W_F$ 、 $W_S$ 、 $W_T$  分别为内容、空间、时间关联权重值, 且满足  $W_F + W_S + W_T = 1$ 。

地理空间数据语义相关度计算流程 (图2): 首先从地理空间元数据语料库中提取各个三级指标值, 并进行标准化处理; 然后分别计算内容相关度、空间相关度、时间相关度; 最终通过加权求和, 即通过式 (1), 得到综合地理空间数据语义相关度。

#### 3.2 内容相关度

内容语义相关度是指地理空间数据表示的内容、要素属性之间的相关程度, 由内容的词汇语义相似度和内容的类别相关度确定, 相应的计算模型如下:

$$Fsem = W_{F1} F_1 + W_{F2} F_2 \quad (2)$$

式中:  $Fsem$  是内容语义相关度;  $W_{F1}$  和  $W_{F2}$  分别为内容语义相似度、类别层次相关性的权重值, 两者满足  $W_{F1} + W_{F2} = 1$ ;  $F_1$  和  $F_2$  分别指内容词汇语义相似度和类别相关度。

**3.2.1 内容词汇语义相似度** 目前, 语义相似度算法主要是基于本体词典或知识库的规则方法以及基于大规模语料库的统计方法。采用基于《知网》的语义相似度的度量方式, 首先从元数据中提取内容关键词集合, 然后应用刘群等开发的词汇语义相似度软件 WordSimilarity<sup>[12]</sup> 计算地理空间元数据内容关键词的语义相似度。设数据集 A 和数据集 B 的关键词集合分别为  $(a_1, a_2, \dots, a_n)$  和  $(b_1, b_2, \dots, b_m)$ , 其中,  $n$  和  $m$  为关键词的个数。数据集 A 和数据集 B 的语义相似度计算如下:

$$F_1 = \frac{\sum_{i=1}^n \sum_{j=1}^m WS(a_i, b_j)}{n \times m} \quad (3)$$

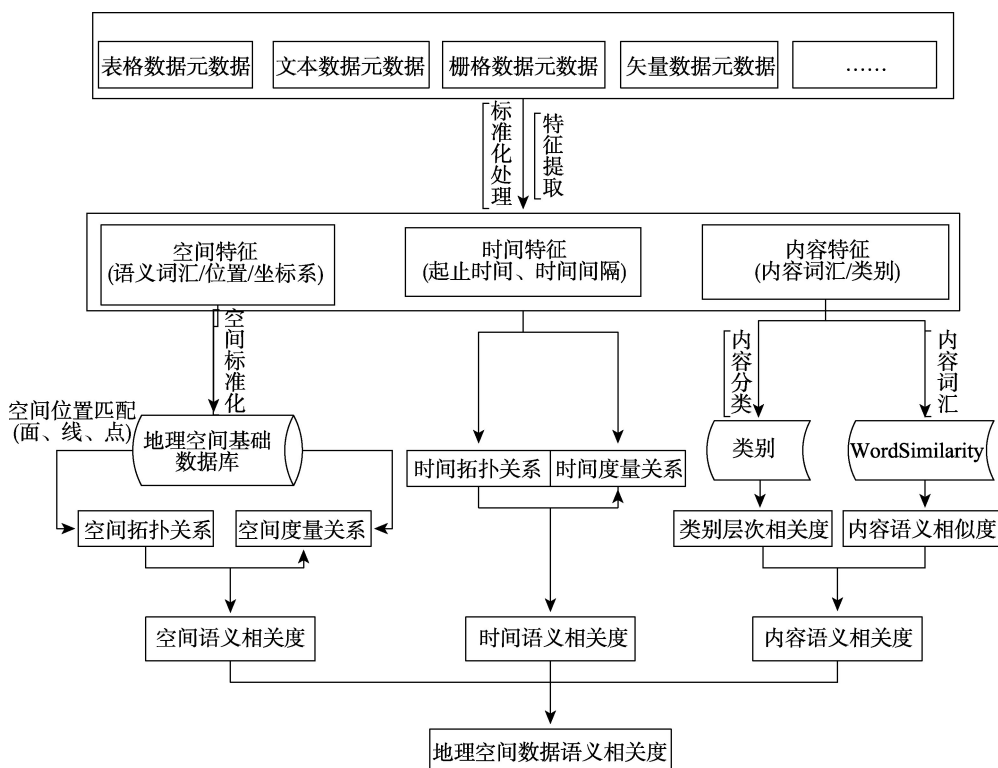


图2 地理空间数据本质特征语义相关度计算流程图

Fig. 2 The calculation flow chart of semantic relevance

式中:  $WS(a_i, b_j)$  为关键词  $a_i$  和  $b_j$  的词汇语义相似度值; 对于关键词  $a_i$  和  $b_j$ , 如果  $a_i$  有  $n$  个义项 (概念):  $S_{a1}, S_{a2}, \dots, S_{ak}$ ,  $b_j$  有  $m$  个义项:  $S_{b1}, S_{b2}, \dots, S_{bl}$ , 则  $a_i$  和  $b_j$  的相似度为最大的义项相似度值<sup>[12]</sup>:

$$WS(a_i, b_j) = \max(WS(S_{ai}, S_{bj})) \quad (i=1, \dots, k, j=1, \dots, l) \quad (4)$$

由于所有的义项根据上下位关系构成了一个树状的义项层次体系, 假设两个义项在这个层次体系中的路径为  $d$ , 两个义项之间的语义相似度:

$$WS(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (5)$$

式中:  $p_1$  和  $p_2$  表示两个义项;  $d$  是  $p_1$  和  $p_2$  在义项层次体系中的路径长度;  $\alpha$  的含义是当相似度为 0.5 时的词语距离值<sup>[12]</sup>。

### 3.2.2 内容类别相关性 (1) 类别层次相关性

地理空间数据内容分类是指数据按专题要素进行分类, 分类体系可以使用层次化的树状结构来描述类与类之间的逻辑关系, 因此, 计算类与类的相关性需要处理分类树中父子节点、兄弟节点等不同类型的关系。地理空间数据类别语义相关度对于数据挖掘、知识发现、类型数据库综合有重要理论意义, 国内外学者对其多有研究<sup>[13-15]</sup>。通过对比分析, 采用 Yao 等的算法<sup>[15]</sup>计算内容类别层次相关性。

设分类树的根节点为  $T$  (图 3),  $T_1$ 、 $T_2$ 、 $T_3$  分支为  $T$  的子树, 计算任意两个非根节点  $X$  和  $Y$  的相关性分两种情况:

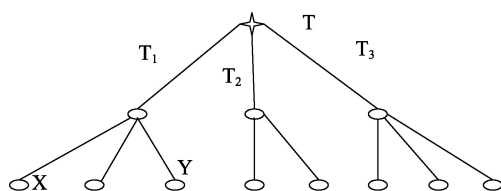


图3 X和Y在同一子树上

Fig. 3 X and Y in the same subtree

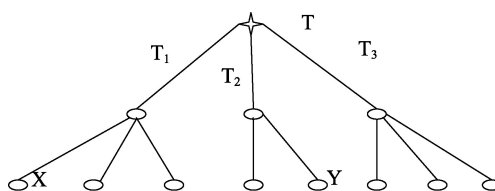


图4 X和Y在不同子树上

Fig. 4 X and Y in the different subtrees

当X和Y在同一子树上时(图3),X与Y的语义相关度  $sim(X,Y)$  的计算公式如下:

$$sim(X,Y) = \frac{l}{l + \alpha(x,y) \times d_x + (1 - \alpha(x,y)) \times d_y} \quad (6)$$

式中:  $l$  表示X和Y的最近共同父类到T的距离(边的数量);  $d_x$  和  $d_y$  分别表示X和Y的最近共同父类到X和Y的距离(边的数量);  $\alpha(x,y)$  表示最近共同父类到X和Y的距离,计算方法如下:

$$\alpha(x,y) = \begin{cases} \frac{d_x}{d_x + d_y} & d_x \leq d_y \\ 1 - \frac{d_x}{d_x + d_y} & d_x > d_y \end{cases} \quad (7)$$

当X和Y不在同一子树上时(图4),分别属于子树  $T_1$  和子树  $T_2$ , X和Y的最近共同父类是T, X与Y的语义相关度  $sim(X,Y)$  计算公式如下:

$$sim(X,Y) = \frac{\beta}{\beta + \alpha(x,y) \times d_x + (1 - \alpha(x,y)) \times d_y} \quad (8)$$

式中:  $\beta$  表示X和Y所在子树的相关度,取值在[0,1],根据实际应用由领域专家给出。本文中,  $\beta$  表示不同领域地理空间数据之间的相关度,如农业和林业、农业和气候等。

另外,计算任意节点X与根节点T的相关度公式如下:

$$sim(X,T) = \frac{1}{d_x} \quad (9)$$

式中:  $d_x$  表示X到T的距离(边的数量)。

(2) 类别相关部分比例

$$F_2 = W_{F_{21}} F_{21} + W_{F_{22}} F_{22} \quad (10)$$

$$F_{21} = \frac{n_f}{Nf_A} \quad (11)$$

$$F_{22} = \frac{n_f}{Nf_B} \quad (12)$$

式中:  $n_f$  为类别相关的个数;  $Nf_A$  表示数据集A总类别数据;  $Nf_B$  表示数据集B总类别数;  $W_{F_{21}}$  和  $W_{F_{22}}$  分别为类别相关部分占数据集A和数据集B的比例的权重值,在此认为两者同等重要,即  $W_{F_{21}} = W_{F_{22}} = 0.5$ 。

### 3.3 空间语义相关度

从地理空间元数据中提取的空间信息一般为文本格式,如行政区划、特征区域(如黄土高原、长江三角洲、京津冀)、道路名称(如国道311)、特征位置(如气象站点、山峰)等地理名称。如何根据地理名称来计算数据集之间的空间拓扑关系和空间度量关



系，是空间语义相关度计算的根**本**。本文首先建立具有统一空间参考的基础地理数据库，并按照面、线、点的顺序将文本格式的地理名称与基础地理数据库中空间数据图层的属性值进行匹配，从而将地理名称映射到空间几何实体，进而依据匹配到的空间几何实体来计算空间拓扑关系和空间度量关系。

**3.3.1 空间拓扑关系** 目前，普遍应用的拓扑关系模型是由 Egenhofer 等建立的 4 交模型和 9 交模型<sup>[16,17]</sup>。根据空间特征关联的特点，采用 4 交模型表示空间实体的拓扑关系。在实际应用中，基础地理数据库中的矢量数据共有点、线、面三种类型，任意两种类型的拓扑关系如表 2 所示。

**3.3.2 空间度量关系** 空间度量关系如重叠比例、空间距离等，一方面可以辅助量化空间拓扑关系，另一方面可提高空间语义相关度计算的准确性，包含两个指标：空间重叠比例（ $S_{21}$ ）和空间距离（ $S_{22}$ ），相关定义如下：

- 定义 1，空间重叠比例：几何实体重叠部分的面积/长度与实体总面积/长度的比值。
- 定义 2，空间距离：空间实体主要涉及到点、线、面三种几何形态，点一点、点一面、面一面的距离指几何中心的欧式距离；点一线、线一面的距离指点**和**面的几何中心到线的最短距离；线一线的距离指线的最短距离。
- 定义 3，空间距离比：两个空间实体的空间距离与实体外包圆半径和之比。
- 定义 4，基本权重：两个空间实体满足一种拓扑关系时专家所给予的最小权重。
- 定义 5，控制权重：考虑空间度量关系情况下，一种拓扑关系所能达到的最大权重。如重叠的极限为两个实体完全相互重叠，即相等，这时取最大权重为 1。

空间度量关系不能一概而论，如点一线、点一面相交的图形是点，因此， $S_{21}$ 是没有实际意义的，空间距离（ $S_{22}$ ）控制度量关系。如果面一面的关系是 Touches，那么  $S_{21}$  指面一面相接线的长度占面周长的比例；如果面一面的关系是 Contains/Overlaps，那么  $S_{21}$  指相交面积占面的面积的比例。因此，空间度量关系的计算还要考虑具体的拓扑关系，即度量关系是用来区分具有相同拓扑关系的几何实体之间的相关度。同一拓扑关系有基本权重（ $W_{S1min}$ ），和控制权重（ $W_{S1max}$ ）。不同实体类型间相同拓扑关系、相同实体类型间不同拓扑关系的空间重叠比例（ $S_{21}$ ）与空间距离（ $S_{22}$ ）的重要程度不同。

- （1）点一线拓扑关系：由于点一线相交图形为点，因此  $S_{21}$  不具有实际意义。点在线上时，距离线中心越近的点，点一线之间的关联强度越强
- （2）点一面拓扑关系：由于点一面相交图形为点， $S_{21}$  不具有实际意义。根据地理学第一定律，距离面中心越近的点与面的相关度越大。

表 2 空间实体拓扑关系  
Tab. 2 Spatial topology relationships

空间拓扑	英文	点一点	点一线	点一面	线一线	线一面	面一面
相等	Equals	★	***	***		***	
相接	Touches	***					
穿过	Crosses	***	***	***			***
重叠	Overlaps	***	***	***			
包含/被包含	Contains/Within	***					
相离	Disjoints	★ ★					

注：\*\*\*表示在现实应用中，该类型的拓扑关系不具有实际意义。

(3) 线—线拓扑关系：相交的图形有点、线两种情况，相交的图形是点， $S_{21}$ 不具有实际意义，空间度量关系由距离控制；相交图形为线， $S_{21}$ 表示相交线段长度占两个线实体长度总和的比例。

(4) 线—面拓扑关系。相交的图形有点、线两种情况，当相交为点， $S_{21}$ 不具有实际意义，空间度量关系由距离控制；当相交为线， $S_{21}$ 表示相交线段长度占线实体长度的比例/占两个面实体周长总和的比例。

(5) 面—面拓扑关系。相交的图形有点、线、面三种情况，当相交为点， $S_{21}$ 不具有实际意义，空间度量关系由距离控制；当相交为线， $S_{21}$ 表示相交线的占两个面实体周长总和的比例；当相交为面， $S_{21}$ 表示相交部分的面积占两个面实体面积之和的比例。

通过以上分析，任意两个数据集的空间度量关系相关度的计算方法如下：

$$S_2 = W_{S_{21}} S_{21} + W_{S_{22}} S_{22} \quad (13)$$

$$S_{22} = 1 - \frac{D_s}{R_A + R_B} \quad (14)$$

式中： $S_2$ 为空间度量关系相关度； $S_{21}$ 为重叠长度/面积占数据集A、数据集B的空间实体长度/面积比例的均值； $S_{22}$ 为距离相关度  $W_{S_{21}}$  和  $W_{S_{22}}$  为相应指标的权重且满足。 $W_{S_{21}} + W_{S_{22}} = 1$ ； $D_s$ 为空间距离（定义2）， $R_A$ 和 $R_B$ 分别为数据集A的空间实体和数据集B的空间实体的外包圆半径。

**3.3.3 空间语义相关度计算** 根据层次计算方法，空间语义相关度计算模型可表示为：

$$Ssem = W_{S1\min} S_1 + (W_{S1\max} - W_{S1\min}) S_2 \quad (15)$$

式中： $Ssem$ 为空间语义相关度； $S_1$ 为数据集之间的空间拓扑关系，取值为1； $W_{S1\max}$ 和 $W_{S1\min}$ 为相应空间拓扑关系的最小关联权重和最大关联权重。

### 3.4 时间语义相关度

地理空间元数据包含了丰富的时间信息，主要包括地学现象或过程发生、演化、完结的时间，以及相应的地理空间数据采集、存储、处理和分析、再生产与应用过程中的时间。从实际检索应用上考虑，采用地学现象或过程发生和（或）完结的时间，记录方式采用公历时间。

**3.4.1 时间拓扑关系** 地理空间数据集记录的时间有时间点、时间段、复合时间等，复合时间由时间点、时间段符合而成。因此，时间拓扑关系可分为时间点—时间点、时间点—时间段、时间段—时间段三种。

#### (1) 时间点—时间点的拓扑关系

时间点之间存在两种拓扑关系：相等、不相等。相等时，相关度为1；不相等时，相关度为0。

#### (2) 时间点—时间段的拓扑关系

时间点B—时间段A之间存在四种拓扑关系：A包含B、B在A期间、B是A的开始时间、B是A的结束时间。从数据相关性的角度来看，四种时间拓扑关系起到的作用大致相同。本文认为以上四种时间拓扑关系权重相同。

#### (3) 时间段—时间段的拓扑关系

Allen 对时态拓扑关系描述和推理进行了研究，归纳出13种时态关系，分别为 before、overlap、meet、equal、start、finish、during 及其对应的逆关系，equal 没有逆关系，如表3所示<sup>[18]</sup>。其中，2~4的六种时间关系具有相同的拓扑相关性，因此，本研究认为六种时间关系的拓扑权重相同。

**3.4.2 时间度量关系** 与空间度量关系相似，时间度量关系用来调控时间拓扑关系，每种时间拓扑关系都具有相应的基本权重 ( $W_{T1\min}$ ) 和控制权重 ( $W_{T1\max}$ )。包含时间重叠比例 ( $T_{21}$ ) 和时间距离 ( $T_{22}$ ) 两个指标，相关定义如下：

定义6，时间重叠比例：时间重叠长度与时间范围A或时间范围B长度的比值。

定义7，时间距离比：两个时间范围中间时间点的距离与两个时间半径和的比值。

时间度量关系相关度计算如下：

$$T_2 = W_{T_{21}} T_{21} + W_{T_{22}} T_{22} \tag{16}$$

$$T_{22} = 1 - \frac{D_T}{R_{tA} + R_{tB}} \tag{17}$$

式中： $T_2$  为时间度量关系相关度； $T_{21}$  为时间重叠部分占时间A和时间B比例的均值； $W_{T_{21}}$  和  $W_{T_{22}}$  为相应指标的权重，且满足  $W_{T_{21}} + W_{T_{22}} = 1$ ； $D_T$  为时间距离； $R_{tA}$  和  $R_{tB}$  分别为时间A和时间B长度的一半。

**3.4.3 时间语义相关度计算模型** 同理，时间语义相关度计算模型可表示为：

$$Tsem = W_{T1\min} T_1 + (W_{T1\max} - W_{T1\min}) T_2 \tag{18}$$

式中： $T_1$  为时间拓扑关系，取值为1； $W_{T1\min}$  和  $W_{T1\max}$  为相应时间拓扑关系的最小关联权重和最大关联权重。

4 实验分析

4.1 实验数据与实验方法

(1) 实验数据集

实验数据来源于国家科技基础条件平台——地球系统科学数据共享平台 (<http://www2.geodata.cn/>)。该平台的元数据以ISO19100地理信息系类标准为基础，每条地理空间元数据包含了丰富的空间、时间、内容特征。

实验选取地球科学数据共享平台100条数据，提取空间、时间、内容特征，并对其进行预处理以便与基础地理数据库中的属性进行匹配和进一步计算，部分数据处理结果如表4所示。

(2) 基础地理空间数据库

所选取的地理空间数据空间位置都在中华人民共和国内，因此实验建立的基础地理空间数据库包含选取的100条数据所在的全部空间范围。实验基础地理空间数据库包含中华人民共和国国界、中国省界、中国地区界、中国县界、中国单线河流等图层。

(3) 权重设置方法。实验中征求了8位地理科学、地球科学数据共享、地理本体、地理语义等相关领域的专家对一级、二级关联指标进行权重打分，平均结果如表5所示。

表3 时间段—时间段拓扑关系

Tab. 3 Time topology relationships

编号	中文	英文	图示
1	相等	Equals	
2	包含 在...期	Contains During	
3	结束于 以...结束	Finishes FinishedBy	
4	开始 以...开始	Starts StartedBy	
5	相交 被相交	Overlaps OverlappedBy	
6	相接 被相接	Meets MetBy	
7	早于 晚于	Before After	



表 4 地理空间元数据特征提取结果

Tab. 4 The extraction result of geospatial metadata features

数据条目	关键词	内容分类	起始年	终止年	空间特征
2000 年新疆土地覆被数据	土地, 土地覆被	规划地籍, 测绘	2000	2000	新疆维吾尔自治区
2010 年新疆土地覆被数据	土地, 土地覆被	规划地籍, 测绘	2010	2010	新疆维吾尔自治区
江苏沿海 1:10 万土地利用数据集 (1980 s-2010 年)	土地, 土地利用	规划地籍, 测绘	1980	2010	江苏省
长三角地区时间序列遥感影像数据集 (1990-2012 年)	遥感, 影像	测绘	1990	2012	长江三角洲
青藏高原 NPP 时空数据集 (1982-2006 年)	初级, 生产力, 土壤, 碳含量, 生物量	生物, 地学信息	1982	2006	青藏高原
中国 30 m 分辨率的降雨侵蚀力图 (1981-2010 年)	降雨, 侵蚀力,	气候/气象/大气, 地学信息	1981	2010	中华人民共和国
中国 1:25 万三级流域分级数据集	流域, 水资源,	地学信息, 内陆水	2002	2002	中华人民共和国
藏东南帕隆藏布流域冰川水文站点观测数据集	径流, 水资源	地学信息, 内陆水	2007	2008	帕隆藏布江
2009-2012 年南海海洋断面科学考察海面气象过程图集系列	气象, 海洋	气候/气象/大气, 海洋	2009	2012	南海
西藏纳木错流域冰川水文站点观测数据集 (2006-2008 年)	径流, 水资源	地学信息, 内陆水	2007	2008	纳木错
.....	.....	.....	.....	.....	.....

(4) 实验环境: Windows7 操作系统, Intel(R) Core(TM) i5-2400 CPU @3.10GHz, 4GB 内存, 程序实现为 Python 2.7。

4.2 结果分析

本实验根据层次计算法, 逐步计算地理空间数据两两之间的内容特征、空间特征、属性特征。由于篇幅限制, 随机选取“鄱阳湖湖口 2005 年日流量数据集”(简称为“鄱阳湖数据集”)和“上海市 1:10 万土地利用数据集”(简称为“上海市数据集”)两条数据集与其他数据集的语义相关度做分析。

表 6 给出 100 条实验数据中与“鄱阳湖数据集”相关度大于 0.1 的数据集排序。“鄱阳湖数据集”的空间特征、时间特征、内容分类、内容特征分别为:“鄱阳湖”、“2005 年”、“内陆水”、“水资源、水流量”。与之相关度较高的“中国 30m 分辨率的降雨侵蚀力图 (1981-2010 年)”, “中国区域地面气象要素数据集 (1981-2008 年)”, “中国 1:25 万三级流域分级数据集 (2002 年)”在空间和内容上均有一定的关联, 前两条数据在时间上包含“鄱阳湖”数据集。随着语义相关度数值的降低, 可以看出, 相应的数据集与“鄱阳湖”数据的语义相关性随之减弱。

表 7 给出 100 条实验数据中与“上海市数据集”相关度大于 0.2 的数据集排序。“上海市数据集”的空间

表 5 关联指标打分结果

Tab. 5 Scores of relevancy indices

一级指标	一级指标分值	二级指标	二级指标分值
内容关系	41	内容词汇语义相似度	58
		内容分类	42
		小计	100
空间关系	35	空间拓扑关系	60
		空间度量	40
		小计	100
时间关系	24	时间拓扑关系	60
		时间度量	40
		小计	100
合计	100		

表6 与“鄱阳湖湖口2005年日流量数据集”相关度 $\geq 0.1$ 的数据集排序  
 Tab. 6 The sorting of data sets which have the semantic relevancy with "The daily traffic data set of Poyang Lake (2005)" greater than 0.1

数据编号	数据条目	语义相关度
1	鄱阳湖湖口2005年日流量数据集	1
2	中国30 m分辨率的降雨侵蚀力图(1981-2010年)	0.414
3	中国1:25万三级流域分级数据集(2002年)	0.370
4	中国区域地面气象要素数据集(1981-2008年)	0.367
5	青藏高原地区水资源数据(1988年, 分县)	0.354
6	中国环境污染数据库(分省: 1981-2000年; 分城市: 1981-2001年)	0.352
7	东北平原与山地湖区10 km <sup>2</sup> 以上湖泊2008-2010水量观测数据集	0.320
8	东部平原湖区10 km <sup>2</sup> 以上湖泊2007-2009水量观测数据集	0.320
9	青藏高原湖区10 km <sup>2</sup> 以上湖泊2008-2010水量观测数据集	0.320
10	淮河流域2005-2006年面积10 km <sup>2</sup> 以上主要湖泊信息数据集	0.314
11	松花江流域2005-2006年面积10 km <sup>2</sup> 以上主要湖泊信息数据集	0.314
12	西南诸河流域2005-2006年面积10 km <sup>2</sup> 以上主要湖泊信息数据集	0.314
13	辽河流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
14	黄河流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
15	东南诸河流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
16	淮河流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
17	海河流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
18	松花江流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
19	珠江流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
20	海河流域2005-2006年面积10 km <sup>2</sup> 以上主要湖泊信息数据集	0.314
21	长江流域2005-2006年面积1 km <sup>2</sup> 以上湖泊基本信息数据集	0.314
22	长江中下游1980-1989年主要水文站日均流量数据集	0.281
23	杭嘉湖地区1:10万土地利用、水资源与水利工程(2000年)	0.276
24	辽河流域2000年湖泊分布数据集	0.274
25	青藏高原湖区10 km <sup>2</sup> 以上湖泊2008-2010年水质观测数据集	0.259
26	江苏省1:10万土地利用数据集(2005年)	0.253
27	安徽省1:10万土地利用数据集(2005年)	0.253
28	安徽省1:10万土地利用数据集(2000年)	0.253
29	上海市1:10万土地利用数据集(2005年)	0.253
30	重庆市1 km分辨率的NDVI数据集(2001-2010年)	0.144
31	青藏高原NPP时空数据集(1982-2006年)	0.107
32	江苏沿海1:10万土地利用数据集(1980 s-2010年)	0.107

特征、时间特征、内容分类、内容特征分别为：“上海市”、“2008年”、“测绘、规划地籍”、“土地利用”。与之相关度较高的有“上海市1:10万土地利用数据集”系列、“江苏省1:10万土地利用数据集”系列、“长三角1:10万土地利用数据集”系列。这是因为空间上“长三角”包含“江苏省”和“上海市”，且同为土地利用数据集。第10条、第11条数据集虽不是土地利用数据，但空间上属于“上海市”。第21条、第22条数据集空间上都包含“上海市”，但“上海市”与“长三角”的面积比例大于“上海市”与“中国”

表7 与“上海市1:10万土地利用数据集(2008年)”相关度 $\geq 0.2$ 的数据集排序  
Tab. 7 The sorting of data sets which have semantic relevancy with "Land use data set (1:100000) of Shanghai (2008)" greater than 0.2

数据编号	数据条目	语义相关度
1	上海市1:10万土地利用数据集(2008年)	1
2	上海市1:10万土地利用数据集(2005年)	0.720
3	上海市1:10万土地利用数据集(2000年)	0.720
4	上海市1:10万土地利用数据集(1995年)	0.720
5	上海市1:10万土地利用数据集(1980s)	0.720
6	江苏省1:10万土地利用数据集(2008年)	0.592
7	长三角1:10万土地利用数据集(1980s)	0.585
8	长三角1:10万土地利用数据集(2000)	0.585
9	长三角1:10万土地利用数据集(1995)	0.585
10	上海2006年统计年鉴数据集	0.420
11	1:25万上海市乡镇界线数据	0.400
12	江苏省1:10万土地利用数据集(2010年)	0.392
13	江苏省1:10万土地利用数据集(2005年)	0.392
14	江苏省1:10万土地利用数据集(1995年)	0.392
15	江苏省1:10万土地利用数据集(1980s)	0.392
16	江苏省1:10万土地利用数据集(2000年)	0.392
17	安徽省1:10万土地利用数据集(2005年)	0.320
18	安徽省1:10万土地利用数据集(2000年)	0.320
19	安徽省1:10万土地利用数据集(1980s)	0.320
20	浙江省1:10万土地利用数据集(1995年)	0.320
21	长三角地区时间序列遥感影像数据集(1990-2012年)	0.320
22	中国30 m分辨率的降雨侵蚀力图(1981-2010年)	0.311
23	长三角1990年统计年鉴数据集	0.285
24	长三角1995年统计年鉴数据集	0.285
25	中国区域地面气象要素数据集(1981-2008年)	0.285
26	中国环境污染数据库(分省:1981-2000年;分城市:1981-2001年)	0.279
27	华东部分省区(江苏、安徽、上海)1980-1997农作物数据集	0.265
28	杭嘉湖地区1:10万土地利用、水资源与水利工程(2000年)	0.260
29	中国1:25万三级流域分级数据集	0.257
30	长江下游河道、河口地形数据集(2008年)	0.200

的面积比例;时间上都包含“2008”,且时间中点均为1996,与2008的距离相等;内容上与土地利用无关;因此,第21条数据集与“上海市数据集”的关联度大于第22条数据集与“上海市数据集”的关联度。

5 结论与讨论

以提高地理空间数据检索的查全率和查准率为目标,根据地理空间数据特点及数据检索中用户关注的焦点,选取地理空间数据内容、空间、时间三大本质特征建立语义关

联指标体系。在此基础上,采用分层逐级计算的方式构建地理空间数据本质特征语义相关度计算模型。实验结果表明,该模型具有四点优势:① 构建简单、构建周期短。在提高地理空间数据的查全率、查准率的同时,避免了在语义检索中空间、时间、内容本体构建的复杂性、主观性。② 语义相关性的量化计算与领域专家较精确的语义判断相结合。在模型中,几何关系的计算依赖于基础地理空间数据库,空间拓扑关系、空间度量关系均可精确表达计算;时间语义相关度的计算有赖于时间的数值描述;内容语义相关度同时包含了内容特征语义相似度和内容类别的相关性。③ 具有一定的可扩展性。基础地理空间数据库、属性分类、时间描述方式均可根据实际应用进行扩展。④ 可应用于多源异构数据。该模型基于元数据,因此,不受数据格式的限制,不同的数据源均可应用。

通过实验分析,本模型虽然具有多种优势,但是还存在一定的不足。比如空间、时间、内容的特征提取由人工参与,具有一定的主观性;模型中权重的赋值依赖于专家知识。因此,在后续的工作中还要对模型进行优化、改进。如从多标准的元数据中自动或半自动提取空间、时间、内容特征并进行统一化表达;尝试利用训练数据集确定权重。

## 参考文献(References)

- [1] 金海,袁平鹏.语义网数据管理技术及应用.北京:科学出版社,2010.[Jin Hai, Yuan Pingpeng. Technology and Application of Semantic Web Data Management. Beijing: Science Press, 2010.]
- [2] Berners Lee T. Linked Data-Design Issues. <http://www.w3.org/DesignIssues/LinkedData.html>, 2006-07-27.
- [3] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995: 448-453.
- [4] Li W, Goodchild M F, Raskin R. Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. International Journal of Digital Earth, 2014, 7(1): 17-37.
- [5] Gao Y, Gao S, Li R, et al. A semantic geographical knowledge wiki system mashed up with Google Maps. Science China Technological Sciences, 2010, 53(S1): 52-60.
- [6] Fu G, Jones C B, Abdelmoty A I. Building a geographical ontology for intelligent spatial search on the web. In: Proceedings of Iasted International Conference on Databases & Applications, 2005: 167-172.
- [7] 宋佳,诸云强,王卷乐,等.基于GML的时空地理本体模型构建及应用研究.地球信息科学,2009,11(4): 442-451. [Song Jia, Zhu Yunqiang, Wang Juanle. A study on the model of spatio-temporal geo-ontology based on GML. Geo-Information Science, 2009, 11(4): 442-451.]
- [8] Rieker W F. Automated retrieval of information in the internet by using thesauri and gazetteers as knowledge sources. Journal of Universal Computer Science, 2002, 8(6): 581-590.
- [9] Schlieder C, Voegelé T, Visser U. Qualitative spatial representation for information retrieval by gazetteers. Proceeding of Cosit'01 Lncs, 2001: 336-351.
- [10] Farazi F, Maltese V, Dutta B, et al. A semantic geo-catalogue for a local administration. Artificial Intelligence Review, 2013, 40(2): 193-212.
- [11] Tobler W R. A computer movie simulating urban growth in the detroit region. Geospatial Semantics First International Conference Geos, 1970, 46: 234-240.
- [12] 刘群,李素建.基于《知网》的词汇语义相似度计算.见:第三届汉语词汇语义学术研讨会,2002: 59-76. [Liu Qun, Li Sujian. Word's semantic similarity computation method based on HowNet. In: The Third Session of The Chinese Vocabulary Semantics, 2002: 59-76.]
- [13] Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation. Red, 2008, 30(2): 243-254.
- [14] Yang R, Kalnis P, Tung A K H. Similarity evaluation on tree-structured data. Sigmod, 2005, (3): 754-765.
- [15] Liu Y, Molenaar M, Kraak M J. Semantic similarity evaluation model in categorical database generalization. Symposium on Geospatial Theory, 2002, 34(4): 279-285.
- [16] Egenhofer M J, Herring J. Categorizing binary topological relationships between regions, lines and points in geographic databases. The, 1992, (2): 1-28.
- [17] Egenhofer M J, Sharma J, Mark D M, et al. A critical comparison of the 4-intersection and 9-intersection models for spatial relations: Formal analysis. Autocarto, 2011, (5): 1-11.

[18] Allen J F. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 1983, 26(11): 832-843.

## The semantic relevancy computation model on essential features of geospatial data

ZHAO Hongwei<sup>1,2</sup>, ZHU Yunqiang<sup>1,3</sup>, YANG Hongwei<sup>4</sup>, LUO Kan<sup>1,2</sup>

(1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China; 4. China Petroleum Planning & Engineering Institute, Beijing 100083, China)

**Abstract:** Linked data is an effective way to integrate multi-source heterogeneous geospatial data cross domain. Semantic high association is the key point to find out target data accurately and quickly. Semantic relevancy directly reflects the value of semantic association between geospatial data, and has great value in retrieving and ranking the targets. According to the semantic relations of geospatial data in space, time and content, a semantic relevancy computation model focusing on essential features of geospatial data is proposed in this research. We compute the semantic relevancy hierarchically through building up a relevancy indices system for essential characteristics. Spatial semantic relevancy is calculated by taking spatial topology relationships and spatial measurement relationships into account. The spatial semantic relevancy is bigger when the distance is smaller and the relative area (or length) is bigger of two spatial objects in the same spatial topology relationship. Accordingly, the time semantic relevancy is calculated by taking into account time topology relationships and time measurement relationships. The time semantic relevancy is bigger when the distance is smaller and the relative time is bigger between two times. The content relevancy is calculated by taking into account the semantic similarity of content keywords and the category correlation degree. Taking geographic metadata as the corpus, this model, which is different from traditional ones, was built up by considering the characteristics of geospatial data and their different important degrees in retrieval and using the methods of geometry processing, numerical computation, semantic similarity calculation and analysis of category relevancy. This model has the advantages of simply building process, suitable for multi-source heterogeneous data, and fully combining mathematics computation and semantics judgment of experts. The result showed that the model can be used to calculate the semantic relevancy on essential characteristics of geospatial data effectively, which can improve the speed of searching targets and the accuracy of the retrieval, and has good expansibility.

**Keywords:** geospatial data; spatial characteristic; temporal characteristic; content characteristic; semantic relevancy