

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 Hongwei, Zhu Yunqiang, Hou Zhiwei et al. Construction of Geospatial Metadata Association Network. Scientia Geographica Sinica, 2016, 36(8): 1180-1189.] doi: 10.13249/j.cnki.sgs.2016.08.008

# 地理空间元数据关联网络的构建

赵红伟<sup>1,2,3</sup>, 诸云强<sup>1,2</sup>, 侯志伟<sup>1,2,3</sup>, 杨宏伟<sup>4</sup>

(1. 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. 中国科学院地理科学与资源研究所, 北京 100101; 3. 中国科学院大学, 北京 100049; 4. 中国石油规划总院, 北京 100000)

**摘要:** 利用资源描述框架(RDF)设计地理空间元数据关联模型, 根据地理空间元数据之间的语义关系和语义相关度的计算, 以构建以元数据为节点、元数据之间的语义关系为边、语义相关度为权重的关联网络。在这一网络中, 一个节点是一个地理空间元数据的资源描述图, 包含属性特征(数据来源、空间特征、时间特征、内容)及其关系特征(元数据之间的语义关系、语义相关度)。实验及其分析表明, 地理空间元数据关联网络可以有效地支持地理空间数据语义关联检索、推荐等应用, 这与传统的基于关键词的元数据检索方式相比, 具有更高的准确度。

**关键词:** 地理空间元数据; 关联数据; 语义相似度; 关联网络

**中图分类号:** TP391      **文献标识码:** A      **文章编号:** 1000-0690(2016)08-1180-10

地理空间数据, 是指描述地球表面位置的信息<sup>[1]</sup>; 是指具有明确位置数据的信息, 通常情况下是在某一坐标系统下的几何体<sup>[2]</sup>。从传统的地理信息系统(GIS)到近年来分布式网络环境下的地理空间数据共享<sup>[3]</sup>和基于Web的地图服务等, 地理空间数据取得了广泛的应用和跨越式的发展。随着3S技术的发展, 地理空间数据来源日益广泛、内容更加丰富、存储格式趋于多样。通过元数据用户可迅速了解数据的名称、质量、组织方式等详细信息。目前, 国内外地理空间元数据标准日趋完善<sup>[4]</sup>, 但是, 元数据基于传统方式的组织形式和检索方式, 已不能够满足用户对语义检索的需求, 如: 用户通过关键词“江苏”只能检索到元数据中具有“江苏”字符的地理空间数据, 而在空间中包含于“江苏”的“南京”、包含“江苏”的“长江三角洲”、邻接“江苏”的“上海”等均不能够被检索到。除此之外, 地理空间数据还具有时间、内容、来源等多维语义关系, 如何充

分利用元数据中蕴含的丰富的语义信息, 准确、快速发现, 并推荐地理空间数据, 成为地理空间数据共享迫切需要解决的问题。关联数据的提出<sup>[2]</sup>为这一问题提供了有效的解决方式。

关联数据不仅仅是把数据发布到网络中构建被人类理解的文档网络, 还构建了数据与数据之间能被计算机理解的结构化、语义化的链接。通过已有数据找到与之相关的其他数据, 以开发构建更加智能化应用服务<sup>[5]</sup>。自“关联数据”提出以来, 越来越多的网络资源正在向着数据网络(Web of Data, 即Web中可被机器理解的语义数据)方向前进<sup>[6]</sup>, 地理空间元数据关联网络正是这其中一种。地理空间元数据关联网络本质上是地理空间数据与数据之间通过元数据的描述信息来建立关联, 是地理空间元数据关联的一种实现方式。

近年来, 国内外学者对地理空间关联数据的研究取得了丰硕的研究成果<sup>[6-9]</sup>: 英国的陆地测量部是第一个将多种地理空间数据以开放链接的形

**收稿日期:** 2015-11-23; **修订日期:** 2016-05-04

**基金项目:** 国家自然科学基金项目(41371381)、科技部科技基础性工作专项项目(2013FY110900)、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)资助。[Foundation: National Nature Science Foundation of China (41371381), Science and Technology Basic Work of Science and Technology (2013FY110900), the National Key Scientific Instrument and Equipment Development Project (2012YQ06002704), Science and Technology Project of Yunnan Province (2012CA021).]

**作者简介:** 赵红伟(1987-), 女, 山东聊城人, 博士研究生, 主要研究方向为地理空间数据语义关联、地理空间数据共享。E-mail: zhao-hw.10s@igsnrr.ac.cn

式发布的国家制图机构<sup>①</sup>; Linked GeoData<sup>②</sup>以 RDF<sup>③</sup>的形式应用 Open StreetMap 数据并应用声明的 SPARQL 语言对其进行检索; Longley P A 等人将西班牙的地理空间数据以 RDF 的形式进行发布<sup>④</sup>。以上研究,将地理空间元数据的属性与关联词汇集或本体进行连接,通过计算或推理词汇或本体概念的语义关系得到属性之间语义关系,进而得到地理空间元数据某一项属性的语义关系。而地理空间元数据往往具有多个属性信息,因此,将大大增加通过已知地理空间元数据找到与之综合语义相关较大的其他元数据计算的复杂度。

本文提出地理空间数据元数据关联网络,通过计算元数据间多维语义的综合相关关系和相关度,建立地理空间数据的直接关联。这可以降低元数据检索中算法的复杂性,提高检索效率,根据语义相关度进行检索结果的排序等。

## 1 地理空间元数据关联模型

### 1.1 地理空间元数据关联网络

地理空间元数据关联网络本质上是以元数据为节点,元数据之间的语义关系为边,语义相关度为边的权重的有向图。它以地理空间元数据为基元,主要为地理空间数据的语义发现、语义推荐等应用服务。其中,地理空间元数据包含空间、时间、内容、数据组织方式、数据质量等多方面的信息,元数据之间通过这些信息的综合语义关系建立元数据与元数据之间的关联。从语义关联上讲,如果将元数据的每一个特征均考虑到关联网络中,不仅增加网络的复杂性,而且会减弱主要语

义的应用。因此,在构建语义关联网络过程中,必须对元数据信息进行取舍,使得网络具有灵活性、可控性,网络应用目标更加明确、集中。

本文综合考虑地理空间数据的基本特征和用户关注的地理空间元数据主要信息选取用于关联的元数据特征,通过被选特征之间的语义关系建立地理空间元数据关联模型。地理空间元数据关联模型应包含:用于关联的元数据描述信息和元数据与元数据间的综合语义关系及语义相关度。

### 1.2 地理空间元数据描述模型

地理空间元数据描述模型的主体类为地理空间元数据(Geospatial Metadata)(图1)。空间特征(空间名称)、时间特征(时间词汇)、内容特征(内容关键词、内容分类)是地理空间数据本质特征,是用户语义检索主要关注的特征;数据来源(提供者)是地理空间元数据的必要特征,是衡量数据质量的重要指标,也是构建关联数据时描述模型的必要属性<sup>[10]</sup>。

1) 空间特征:是指地理空间元数据中表达空间特征的名词(一个或多个),在元数据关联网络构建过程中被映射到空间基础数据库中的空间实体,利用空间实体建立元数据间的空间语义关系。

2) 时间特征:是指地理空间数据集中的现象或事件在现实中发生或存在的时间(一个或多个)。可以是时间点,也可以是时间段,一个时间关键词对应一个时间类的实例。时间类包含起始时间、终止时间、时间间隔和时间单位4个属性信息。

3) 内容特征:内容关键词是描述地理空间数据集内容的关键词集合;内容分类是关键词所属的类别。

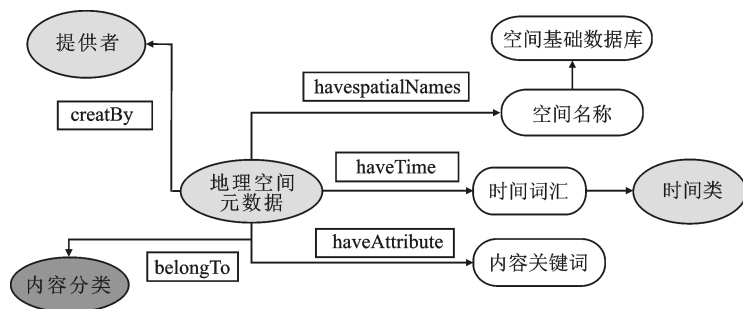


图1 地理空间元数据描述模型

Fig.1 The descriptive model of geospatial metadata

① <https://www.ordnancesurvey.co.uk/business-and-government/products/opedat>

② <http://linkedgeodata.org/About>

③ <http://www.phpstudy.net/e/rdf/>

④ <http://geo.linkeddata.es/>

4) 数据来源:是指数据的提供者,从一定程度上保证了数据质量,是地理空间元数据描述模型的必要特征。包含提供者的名字、邮箱、所属公司/单位3个属性。

### 1.3 地理空间元数据关联模型

地理空间元数据关联模型通过元数据描述信息综合语义关系构建(图2),包含语义关系(定性)和语义相关度(定量)两部分。定性的语义关系包含空间拓扑关系(spatialTopology)、时间拓扑关系(temporalTopology)、内容类别(contentCategory)3个语义维度。定量的语义相关度(semanticRelevancy)的计算不仅考虑以上3类语义关系,还考虑空间度量关系、时间度量关系、内容字面相关度等。

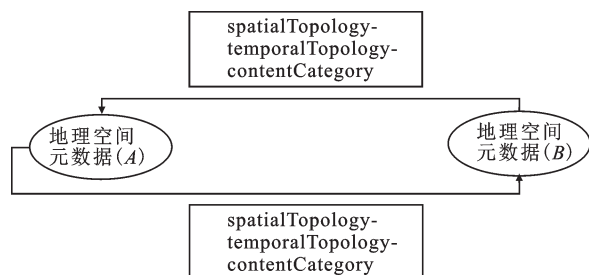


图2 地理空间元数据关系模型

Fig.2 The relationship model of geospatial metadata

1) 空间语义关系。Sloman 等人<sup>[11]</sup>对空间概念和实体做了解释:概念是一类实体或一组实体特征的描述,地理空间概念是用来描述地理特征类型。地理空间的语义相关主要由关系表征,关系表示概念之间的一类关联。地理空间数据的空间语义关系主要包含拓扑关系、方向关系、度量关系。根据常见的语义应用(数据检索、语义查询等),本文考虑空间拓扑关系和空间度量关系。

(1) 空间拓扑关系:影响较大的空间拓扑模型是Randell等人提出的区域连接演算RCC(region connection calculus)理论。RCC理论以空间中的区域为对象,而不是几何学中的无维度的点,用“连接”来表达空间对象间的基本关系,共有8种可能性(表1),可知Inside与Coverby两种拓扑关系可以合为一种,本研究取谓词Inside;同理,Contains与Covers两种拓扑关系共同采用谓词Contains。

(2) 空间度量关系:主要指距离关系,也包括其他与尺度有关的关系,如面积、体积等。距离关系的表示分为绝对距离和相对距离关系。绝对距

表1 RCC-8 拓扑关系

Table 1 Region connection calculus

序号	关系代码	关系名称	语义解释
1	DC	Disjoints	A、B 相离(不相交)
2	EC	Meets	A、B 相接
3	EQ	Equals	A、B 相等
4	NTPP	Inside	A 包含于 B,且两者边界不交
5	NTPPi	Contains	A 包含 B,且两者边界不交
6	TPP	Coverby	A 包含于 B,且两者边界相交
7	TPPi	Covers	A 包含 B,且两者边界相交
8	PO	Overlaps	A、B 部分重叠

离关系直接表示两个空间对象之间的距离,相对距离关系通过与第3个对象的比较,间接表示两个对象间的距离。绝对距离可以是定性关系,也可以是定量关系,而相对距离一般是定性的。由于定性距离在不同的空间尺度下,具有不同的距离语义认知,比如,在全球尺度上“北京”与“广州”距离较近,而在全中国尺度上“北京”与“广州”非常远。因此,本文采用绝对距离的定量关系,如定义1。

定义1,空间距离:空间实体主要涉及到点、线、面3种几何形态,本文中点-点、点-面、面-面的距离指几何中心的欧式距离;点-线、线-面的距离指点、面的几何中心点到线的垂直距离;线-线的距离指线的最短距离。

2) 时间语义关系。地理空间数据集的时间主要是指地学现象或过程发生、演化、完结的时间,本文采用公历时间和时钟时间对其描述,相关的时间语义关系主要有时间拓扑关系和时间度量关系。在时间关系研究方面,时间区间代数(Interval Algebra)理论<sup>[12]</sup>以时间段为基元,总结了Before、After等13种基本的时间关系及其推理规则和算法,成为时间关系研究的基础。

(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,时间拓扑关系分为时间点-时间点、时间点-时间段、时间段-时间段3种。

定义2,时间点:时间轴上的每个点表示一个时刻,它没有长短,只有先后,是一个序数,用 $t_i$  ( $i=1,2,3,\dots,n$ )表示。

定义3,时间段:时间轴上的一段时间表示一个时间段,时间段可以表示为 $T=[ts, te]$ ,ts和te分别表示时间段的开始时间点和结束时间点,且 $ts < te$ 。



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(相离);② 时间点 $B$ -时间段 $A$ 之间存在4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;③ 时间段-时间段的拓扑关系。

本文采用Allen归纳出的13种时态关系,分别为 before、overlap、meet、equal、start、finish、during 及其逆关系(equal没有逆关系)。

(2) 时间度量关系:主要指时间距离关系,与空间距离相似,时间距离也分为绝对距离和相对距离,绝对距离包含定性距离与定量距离,同空间度量关系相似,本文采用绝对距离中的定量距离,如定义4。

定义4,时间距离:时间段的中心值或时间点的时间轴上的距离。

(3) 内容语义关系。内容语义是对地理事物和现象普通专题特征含义的表达,主要包含复合关系、从属关系、分类与概括关系等<sup>[13]</sup>。根据研究目标,本文只考虑内容的分类与概括关系。

地理空间数据内容分类是指数据按专题要素进行分类,分类体系都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(图3)。主要包括父子关系( $T_1$ 是 $X$ 的父类; $X$ 是 $T_1$ 的子类)、兄弟关系( $X$ 与 $Y$ 有共同的父类),共3种类别关系。

综上所述,地理空间元数据具有6种空间拓扑关系(表1),13种时间拓扑关系和3种内容分类关系(图3)。地理空间元数据之间在空间、时间、内容三维语义空间中,理论上共有234种的语义关系。本文分别定义空间关系谓词、时间关系谓词和内容类别关系谓词,对三者组合来定义234种地理空间元数据间的语义关系谓词。

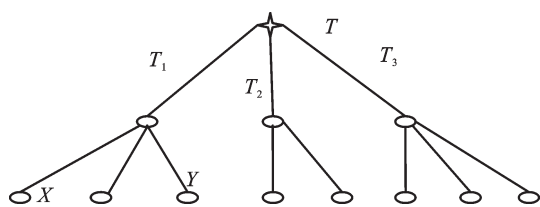


图3 内容分类树

Fig.3 Classification tree of the contents

空间拓扑关系谓词包含: Disjoints, Meets,

Equals, Inside, Contains, Overlaps;时间拓扑关系谓词包含: Equals, Contains, During, Finishes, Finished-By, Starts, StartedBy, Overlaps, OverlapedBy, Meets, Metby, Before, After;内容类别谓词包含: FatherOf, SonOf, BrotherOf。地理空间元数据的关系谓词为:“spatial”+空间关系谓词+“-temporal”+时间关系谓词+“-class”+内容类别关系谓词。

## 2 地理空间元数据语义相关度的计算

地理空间元数据的语义相关度是指语义关系的定量描述,直接反映语义关系的强弱。在一些不需要了解具体语义关系的应用中有重要价值,如提高元数据的检索效率并对结果进行排序。本文通过分别计算空间语义关联强度、时间语义关联强度和-content语义关联强度,构建基于三者的地理空间元数据语义相关度计算的线性模型。

$$Geos = W_s S_s + W_t T_s + W_f F_s \quad (1)$$

其中, $Geos$ 为地理空间数据语义相似度, $S_s, T_s, F_s$ ,分别为空间、时间、内容语义相关度,计算公式见(2~8)。 $W_s, W_t, W_f$ ,为相应的关联权重值<sup>①</sup>,且满足  $W_f + W_s + W_t = 1$ 。

### 2.1 空间语义相关度的计算

空间语义关系主要包含定性的拓扑关系和定量的度量关系。通常在同一空间尺度、同一拓扑关系中,距离越近、重叠长度/面积越大,空间语义相关度越高。因此,本文通过计算同一拓扑关系下不同度量关系相关度来计算空间语义相关度。

$$S_s = W_{s_{1min}} + (W_{s_{1max}} - W_{s_{1min}}) S_2 \quad (2)$$

其中, $W_{s_{1min}}$ 为地理空间元数据之间的空间拓扑关系为 $S_1$ 时的基本权重; $W_{s_{1max}}$ 为相应最大权重, $S_2$ 为在拓扑关系 $S_1$ 中度量关系的关联值。

$$S_2 = W_{s_{21}} S_{21} + W_{s_{22}} S_{22} \quad (3)$$

$$S_{22} = 1 - D_s / (R_A + R_B) \quad (4)$$

其中, $S_{21}$ 为重叠长度/面积占数据集 $A, B$ 的空间实体长度/面积比例的均值, $S_{22}$ 为距离相关度。 $W_{s_{21}}, W_{s_{22}}$ 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1), $R_A$ 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。由于 $S_{21}, S_{22}$ 是归一化值,公式(3、4)可应用于不同的空间尺度。

① 本文中的关联权重值由专家打分判断。

## 2.2 时间语义相关度的计算

时间语义关系包含定性的拓扑关系和定量的度量关系。与空间关系相似,同一时间尺度、时间拓扑关系中,距离越近、重叠的时间越长,时间语义相关度越大。时间语义相关度计算方法:

$$T_s = W_{T_{1\min}} + (W_{T_{1\max}} - W_{T_{1\min}})T_2 \quad (5)$$

其中,  $W_{T_{1\min}}$ 、 $W_{T_{1\max}}$  为时间拓扑关系时  $T_1$  最小关联权重和最大关联权重,  $T_2$  为时间度量关系相关度。

当时间实体  $A$ 、 $B$  相离时(after/before):

$$T_2 = 1/D_T \quad (6)$$

当时间实体  $A$ 、 $B$  相离时,  $D_T > 1$ , 因此,  $T_2 < 1$ 。

当时间实体  $A$ 、 $B$  不相离时,

$$T_2 = W_{T_{21}}T_{21} = W_{T_{22}}T_{22} \quad (7)$$

$$T_{22} = 1 - D_T/R_{T_A} = R_{T_B} \quad (8)$$

其中,  $T_{21}$  为时间重叠部分占时间  $A$ 、 $B$  比例的均值,  $W_{T_{21}}$ 、 $W_{T_{22}}$  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,  $R_{T_A}$  和  $R_{T_B}$  分别为时间  $A$  和时间  $B$  长度的一半。由于  $T_{21}$ 、 $T_{22}$  是归一化值,因此,公式(7、8)可应用于不同的时间尺度。

## 2.3 内容语义相关度的计算

除了类别语义关系,本文还考虑了内容关键词之间的相同比例来计算内容语义相关度。

$$F_s = W_{F_1}F_1 + W_{F_2}F_2 \quad (9)$$

其中,  $F_s$  是内容语义相关度,  $W_{F_1}$ 、 $W_{F_2}$  分别为关键词相同比例、类别层次相关性的权重值,两者满足

$W_{F_1} + W_{F_2} = 1$ 。 $F_1$ 、 $F_2$  分别指内容关键词相同比例和类别相关度,具体计算方法如下。

设数据集  $A$ 、 $B$  的关键词集合分别为  $(a_1, a_2, \dots, a_n)(b_1, b_2, \dots, b_m)$ , 其中,  $n$ 、 $m$  为关键词的个数。数据集  $A$ 、 $B$  的语义相似度计算:

$$F_1 = \sum_{i=1}^n \sum_{j=1}^m WS(a_i, b_j) / n \times m \quad (10)$$

其中,  $WS$  为词汇相似度。当关键词  $a_i, b_j$  相同时,  $WS(a_i, b_j)$  取值为 1; 当关键词  $a_i, b_j$  不同时,  $WS(a_i, b_j)$  取值为 0。

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。国内外学者对其多有研究<sup>[14-16]</sup>, 通过对比分析, 本文采用 Yaolin L 等<sup>[16]</sup>算法计算内容类别层次相关性。

## 3 地理空间元数据关联网络实例

首先从地理空间元数据中提取空间、时间、内容、来源特征; 其次, 根据提取的元数据特征和地理空间元数据关联模型构建描述对象和关系的词汇集; 最后, 根据元数据特征计算元数据之间的语义关系和语义相关度(图 4)。

### 3.1 实验数据及数据预处理

1) 地理空间元数据。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<sup>①</sup>, 该平台的元数据以 ISO 19100 地理信息系统

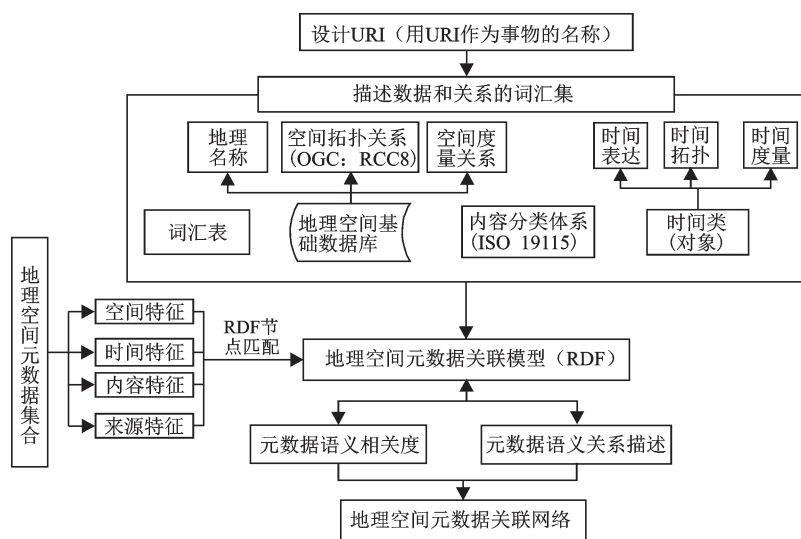


图 4 地理空间元数据关联网络构建的技术路线

Fig. 4 Flow chart of geographic spatial metadata association network construction

① <http://www2.geodata.cn/index.html>

类标准为基础。实验选取100条元数据,从中提取地理空间数据集的空间、时间、内容、来源(所有者)等信息并进行预处理。其中时间信息包含起始时间、终止时间、时间间隔、间隔单位(年、月、日等);内容分类采用ISO 19115元数据内容分类标准,共19类<sup>[17]</sup>;来源信息包含作者(数据拥有者)名字、联系方式、单位等。

2) 地理空间基础数据库。本文将地理空间元数据中提取的空间关键词映射到地理空间基础数据库的方式建立地理空间元数据之间的空间拓扑关系和空间度量。地理空间基础数据库中图层数据的质量直接影响空间语义关系和空间语义相似度的精度。根据实验采用的元数据,地理空间基础数据库主要包含中国边界图层、中国省界图层、中国县界图层、中国乡镇图层、中国主要河流图层、中国陆地地形图层、中国主要湖泊图层等。数据库中所有的矢量/栅格图层采用统一坐标系——北京1954地理坐标系。

### 3.2 元数据关联网络构建

地理空间元数据之间通过语义关系建立关联、通过语义相关度来衡量关联强度,如果把地理空间元数据看做网络中的一个节点,语义关系看做元数据之间的有向边,那么语义相关度即为边的权重。因此,元数据关联网络的构建主要包含语义关系和相应的语义相关度值。

1) 语义相关度关联权重。由前文语义相关度计算方法可知,地理空间元数据之间的语义相关度的计算需要对多个权重值进行赋值。由于6种空间拓扑关系——Disjoints, Meets, Overlaps, Inside, Contains, Equals, 权重值呈递增,且空间度量关系起到权重调控的作用,空间拓扑基本权重( $W_{s_{i_{min}}}$ )和拓扑最大权重( $W_{s_{i_{max}}}$ )可通过训练数据集得到较理想的值,最终确保每组元数据空间相关度的相对大小。时间拓扑基本权重( $W_{t_{i_{min}}}$ )和拓扑最大权重( $W_{t_{i_{max}}}$ )也可得到。

除了以上两类权重,其余权重的赋值采用专家打分的方法。实验征求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公式(1、3、7、8)中的权重进行打分(表2)。

2) 关联网络结果及分析。将地理空间元数据的空间名词映射到地理空间基础数据库中的空间实体,计算空间实体之间的空间拓扑关系和空间相关度来计算元数据的空间关系;用公历时间

表2 权重打分结果

Table 2 Weights scores for each relationship by experts

权重项目	权重值	权重项目	权重值
内容关系( $W_f$ )	41	关键词相同比例权重( $W_{f1}$ )	58
		内容分类( $W_{f2}$ )	42
		小计	100
空间关系( $W_s$ )	35	空间重叠比例( $W_{s21}$ )	60
		距离相关度( $W_{s22}$ )	40
		小计	100
时间关系( $W_t$ )	24	时间重叠比例( $W_{t21}$ )	60
		距离相关度( $W_{t22}$ )	40
		小计	100
合计	100		

值计算时间拓扑关系和时间相关度;利用ISO19115内容分类体系计算类别相关度,该标准只包含19个一级类。类别相关性只有BrotherOf和No(没有关系)两种情况。本文随机选取藏高原东缘气象气候研究数据(1956~1993年)(简称青藏高原数据集)和江苏省1:10万土地利用数据(1980s)(简称江苏省数据集)两条元数据与其他元数据的语义关系进行分析(表3、4)。设青藏高原东缘气象气候研究数据(1956~1993年)为A,表3中1~25条元数据为B;江苏省1:10万土地利用数据(1980s)为C,表4中1~25条元数据为B。每个元数据是一个节点,语义关系是由A/C指向B/D的边,语义相关度是边的权重值。

表3中,空间概念“青藏高原”是指青藏高原在中国境内的空间范围。元数据A是主语,元数据1~25是宾语,语义关系是相应的关系谓词。元数据A在空间上与元数据2相等(spatialEquals)、时间上包含元数据2(temporalContains)、与元数据2是有共同的父类(BrotherOf),两者的语义相关度是0.952 2——紧密相关。元数据A与其他元数据之间的关系可知,语义相关性由元数据3到元数据25依次降低。当元数据A与其他元数据具有相同的语义关系时,语义相关度不一定相等,这是由于语义相关度考量了定量的空间度量关系、时间度量关系和内容关键词相同比例。在实际应用中,当用户输入“青藏高原”、“气象气候”等关键词时,完全匹配的“青藏高原东缘气象气候研究数据(1956~1993年)”首先被检索到,同时还会检索到与之相关的其他数据,检索结果按照与目标数据的语义相关度由大到小排序。这将大大增加数



表3 与“青藏高原东缘气象气候研究数据(1956~1993年)”关联较紧密的元数据

Table 3 The closely related metadata to “the climate research data of eastern margin of Tibetan Plateau (1956-1993)”

编号	元数据条目	语义关系 ( $A \rightarrow B$ )	语义关系释义	语义相 关度
1	青藏高原东缘气象气候研究数据(1956~1993年)	spatialEquals-temporalEquals-classBrotherOf	空间:A与B相等;时间:A与B相等 内容:A与B同类	1
2	青藏高原分区域气候数据(1984年)	spatialEquals-temporalContains-classBrotherOf	空间:A与B相等;时间:A包含B 内容:A与B同类	0.9522
3	1951~2000年中国公里网格多年平均风速	spatialInside-temporalDuring-classBrotherOf	空间:A包含于B;时间:A包含于B 内容:A与B同类	0.9222
4	纳木错和慕士塔格大气化学成分数据集(2005~2009年)	spatialContains-temporalBefore-classBrotherOf	空间:A包含B;时间:A在B之前 内容:A与B同类	0.6568
5	四川省紫色土区气象研究数据(1997~2003年)	spatialOverlaps-temporalBefore-classBrotherOf	空间:A与B相交;时间:A在B之前 内容:A与B同类	0.6524
6	1971~2000年浙江省1:25万累年平均降水量空间分布图	spatialDisjoints-temporalOverlappedBy-classBrother	空间:A与B相离;时间:A被B相交 内容:A与B同类	0.6037
7	1971~2000年浙江25万累年月平均极端最低和最高气温、月平均气温空间分布图	spatialDisjoints-temporalOverlappedBy-classBrother	空间:A与B相离;时间:A被B相交 内容:A与B同类	0.6037
8	青藏高原农田系统生态因子数据集(1960~2000年)	spatialEquals-temporalOverlappedBy-classNo	空间:A与B相等;时间:A被B相交 内容:A与B不同类	0.5698
9	青藏高原草地分布面积、类型、经济特性数据集(1974~1976年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5625
10	青藏高原草地资源物种营养成分数据集(1974~1976年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5625
11	1981年青藏高原水分、水文研究数据	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5507
12	1:100万青藏高原水系流域图(1980~1982年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5507
13	青藏自然区划背景数据(1984年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5422
14	青藏高原湖泊水化学成分图(1990年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5298
15	青藏高原野生动物数据(1991年)	spatialEquals-temporalContains-classNo	空间:A与B相等;时间:A包含B 内容:A与B不同类	0.5277
16	青藏高原森林资源系统生态因子数据集(1980~2000年)	spatialEquals-temporalOverlappedBy-classNo	空间:A与B相交;时间:A被B相交 内容:A与B不同类	0.5170
17	1951~2000年中国公里网格风能数据	spatialInside-temporalDuring-classNo	空间:A包含于B;时间:A包含于B 内容:A与B不同类	0.5122
18	中国森林资源数据库(分省,分县,1950~1993年)	spatialInside-temporalDuring-classNo	空间:A包含于B;时间:A包含于B 内容:A与B不同类	0.5098
19	中国地质灾害事件数据集(1949~2008)	spatialInside-temporalDuring-classNo	空间:A包含于B;时间:A被B相交 内容:A与B不同类	0.5062
20	1961~2000年中国1:100万生态环境背景数据(水、热要素)	spatialInside-temporalOverlappedBy-classNo	空间:A包含于B;时间:A被B相交 内容:A与B不同类	0.4977
21	中国1:400万耕地质量数据(1980s)	spatialInside-temporalContains-classNo	空间:A包含于B;时间:A包含B 内容:A与B不同类	0.4821
22	中国海岸带和海涂资源20世纪80年代综合调查数据	spatialInside-temporalContains-classNo	空间:A包含于B;时间:A包含B 内容:A与B不同类	0.4821
23	西藏部分河流的水文数据(1980~1989年)	spatialOverlaps-temporalContains-classNo	空间:A与B相交;时间:A包含B 内容:A与B不同类	0.4821
24	中国农业物候(1980~1981年)	spatialInside-temporalContains-classNo	空间:A包含于B;时间:A包含B 内容:A与B不同类	0.4806
25	中国草地资源数据库(分省,分县,1980s)	spatialInside-temporalContains-classNo	空间:A包含于B;时间:A包含B 内容:A与B不同类	0.4805

表4 与“江苏省1:10万土地利用数据(1980s)”关联较为紧密的元数据

Table 4 The closely related metadata to “the 1:10 million land use data of Jiangsu province (1980s)”

编号	元数据条目	语义关系 ( $C \rightarrow D$ )	语义关系释义	语义 相关度
1	江苏省1:10万土地利用数据(1980 s)	spatialEquals-temporalEquals-classBrotherOf	空间: $C$ 与相等;时间: $C$ 与 $D$ 相等 内容: $C$ 与 $D$ 同类	1
2	中国土地资源数据库(1980~2001年,分省、 分县)	spatialInside-temporalStarts-classBrotherOf	空间: $C$ 包含于 $D$ ;时间: $C$ 开始于 $D$ 内容: $C$ 与 $D$ 同类	0.8404
3	江苏省1:10万土地利用数据(1995年)	spatialEquals-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相等;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.7646
4	江苏省1:10万土地利用数据(2000年)	spatialEquals-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相等;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.7631
5	江苏省1:10万土地利用数据(2005年)	spatialEquals-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相等;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.7623
6	江苏省1:10万土地利用数据(2008年)	spatialEquals-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相等;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.7620
7	江苏省1:10万土地利用数据(2010年)	spatialEquals-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相等;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.7619
8	上海市1:10万土地利用数据(1980 s)	spatialMeets-temporalEquals-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 与 $D$ 相等 内容: $C$ 与 $D$ 同类	0.7457
9	中国地区土地利用/土地覆盖数据集	spatialOverlaps-temporalBefore-classBrotherOf	空间: $C$ 包含于 $D$ ;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.6268
10	安徽省1:10万土地利用数据(2005年)	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 包含于 $D$ ;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 不同类	0.5502
11	中国农业物候(1980~1981)	spatialInside-temporalStartedBy-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 以 $D$ 开始 内容: $C$ 与 $D$ 不同类	0.5382
12	上海市1:10万土地利用数据(1995年)	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.5103
13	上海市1:10万土地利用数据(2000年)	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.5088
14	上海市1:10万土地利用数据(2005年)	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.5080
15	上海市1:10万土地利用数据(2008年)	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.5077
16	1996年浙江省1:25万数字化土地利用现状 图	spatialMeets-temporalBefore-classBrotherOf	空间: $C$ 与 $D$ 相接;时间: $C$ 在 $D$ 之前 内容: $C$ 与 $D$ 同类	0.4963
17	中国1:400万耕地质量数据(1980 s)	spatialInside-temporalEquals-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 与 $D$ 相等 内容: $C$ 与 $D$ 不同类	0.4537
18	中国海岸带和海涂资源20世纪80年代综合 调查数据	spatialInside-temporalEquals-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 与 $D$ 相等 内容: $C$ 与 $D$ 不同类	0.4537
19	中国历年县级的行政区划数据集(1980~ 2005)	spatialInside-temporalStarts-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 开始于 $D$ 内容: $C$ 与 $D$ 不同类	0.4269
20	1961~2000年中国1:100万生态环境背景数 据(水、热要素)	spatialInside-temporalDuring-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 包含于 $D$ 内容: $C$ 与 $D$ 不同类	0.4257
21	中国地质灾害事件数据集(1949~2008)	spatialInside-temporalDuring-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 包含于 $D$ 内容: $C$ 与 $D$ 不同类	0.4230
22	中国森林资源数据库(分省,分县,1950~1993 年)	spatialInside-temporalDuring-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 包含于 $D$ 内容: $C$ 与 $D$ 不同类	0.4204
23	1951~000年中国公里网格风能数据	spatialInside-temporalDuring-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 包含于 $D$ 内容: $C$ 与 $D$ 不同类	0.4188
24	1951~2000年中国公里网格多年平均风速	spatialInside-temporalDuring-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 包含于 $D$ 内容: $C$ 与 $D$ 不同类	0.4188
25	中国草地资源数据库(分省,分县,1980 s)	spatialInside-temporalStartedBy-classNo	空间: $C$ 包含于 $D$ ;时间: $C$ 以 $D$ 开始; 内容: $C$ 与 $D$ 不同类	0.3894



据的查全率和应用率。

与表4同理,元数据C是主语,元数据1~25是宾语,语义关系是相应的关系谓语。元数据C空间上与元数据2相等(spatialInside)、时间上开始于元数据2(temporalStarts)、与元数据2有共同父类(BrotherOf),两者语义相关度是0.840 4。元数据3-元数据7与元数据C虽然有相同语义关系,但语义相关度不同,这是由时间度量关系的不同导致的结果——时间上,距离元数据1越近,语义相关度越大。实际应用中,当用户对“江苏省1:10万土地利用数据(1980s)”感兴趣时,可以推荐与之语义相关度较大的数据集,给出语义相关类型。

## 4 结束语

地理空间元数据关联网络是基于关联数据技术及其在地理空间数据中的应用,为解决互联网大数据背景下海量、多源、异构的地理空间数据发现、共享等问题提出的元数据与元数据之间直接进行语义关联的数据网络。本文综合考虑地理空间数据特征和用户主要关注的特征,选取元数据中用于语义关联的信息来构建地理空间元数据描述模型,通过元数据描述信息建立地理空间元数据的语义关联,意在打破地理空间元数据间的语义壁垒并消除元数据孤岛现象。构建的地理空间元数据关联网络以元数据为节点,元数据之间的语义关系为有向边,语义相关度值为边的权重。

通过地理空间元数据关联网络实例可知:

① 地理空间元数据描述模型符合一般用户对地理空间数据的空间、时间、内容等检索条件的需求;② 地理空间元数据关联网络中语义关系的计算结果经验证(见表3、4)符合人们对空间拓扑关系、时间拓扑关系和内容分类关系的认知;③ 语义相关度的计算结果能够反映元数据之间语义相关度的相对大小,在语义检索排序中具有较高的应用价值。

该网络可支持多项语义应用:① 语义查询:通过关系谓词可对地理空间元数据进行空间拓扑关系、时间拓扑关系、内容类别关系查询;② 语义关系度量排序:通过对语义相关度可以对元数据之间的语义关联程度进行度量,进而对查询结果进行排序;③ 语义推荐:将与目标数据关联程度高(语义相关度值大)的数据推荐给用户。

地理空间元数据关联网络基于关联数据技

术,本文讨论并实验了地理空间元数据关联网络的构建方法,但还没针对关联网络应用开发出一个完整的原型系统。接下来的研究目标是将元数据关联网络以资源描述框架(RDF)的形式在网上发布,并开发出相应的语义检索原型系统。

## 参考文献(References):

- [1] Béjar R, Latre M Á, Noguera-Iso J et al. An RM-ODP enterprise view for spatial data infrastructures[J]. Computer Standards & Interfaces, 2012, 34(2):263-272.
- [2] Hart G, Dolbear C. Linked data : A Geographic Perspective[M]. Boca Raton:Crc Press, 2013.
- [3] 郑文峰. 面向服务的空间数据共享[D]. 成都:成都理工大学, 2008.[Zheng Wenfeng. Geospatial Data Sharing Based on SOA. Chengdu: Chengdu University of Technology,2008.]
- [4] Yingjie H, Janowicz K, Prasad S et al. Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online[J]. Transactions in Gis, 2015, 19(3):398-416.
- [5] Bizer C, Heath T, Berners-Lee T. Linked Data—The Story So Far. Int[J]. J.semantic Web Inf.syst, 2009, 5(3): 1-22.
- [6] Longle P, Goodchild M, Maguire D et al. Geographic Information Systems and Science[M]. New York: Wiley , 2001.
- [7] James R, William W, Ben B. An Infrastructure for Publishing Geospatial Metadata as Open Linked Metadata[C/OL].[https://agile-online.org/Conference\\_Paper/CDs/agile\\_2012/proceedings/papers/Paper\\_Reid\\_An\\_Infrastructure\\_for\\_Publishing\\_Geospatial\\_Metadata\\_as\\_Open\\_Linked\\_Metadata\\_2012.pdf](https://agile-online.org/Conference_Paper/CDs/agile_2012/proceedings/papers/Paper_Reid_An_Infrastructure_for_Publishing_Geospatial_Metadata_as_Open_Linked_Metadata_2012.pdf)
- [8] Diederik T, Ann C, Thérèse S. Publishing metadata of geospatial indicators as Linked Open Data: A policy-oriented approach[C/OL].[https://agile-online.org/Conference\\_Paper/cds/agile\\_2014/agile2014\\_135.pdf](https://agile-online.org/Conference_Paper/cds/agile_2014/agile2014_135.pdf)
- [9] Yingjie H, Janowicz K, McKenzie G et al. A linked-Data-driven and semantically-enabled journal portal for scientometrics[C/OL].<http://geog.ucsb.edu/~hu/papers/SEJP.pdf>
- [10] Bizer C. Linked Data: Evolving the Web into a Global Data Space[J]. Synthesis Lectures on the Semantic Web Theory & Technology, 2011, (1):1.
- [11] Sloman SA, Love BC, Woo-Kyoung A. Feature centrality and conceptual coherence[J]. Cognitive Science, 1998, 22(2): 189-228
- [12] Allen J F. Maintaining knowledge about temporal intervals[J]. Communications of the ACM, 1983, 26(11): 832-843.
- [13] 李小娟. 基于特征的时空数据模型及其在土地利用动态监测信息系统中的应用[D].北京:中国科学院遥感应用研究所, 1999. [Li Xiaojuan. Research on the Feature-based Spatio-Temporal Data Model and Its Application in Landuse Dynamic Monitoring Information System. Beijing:Institute of Remote Sensing Applications Chinese Academy of Sciences, 1999.]
- [14] Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation[J]. Red, 2008, 30(2):

- 243-254.
- [15] Yang R, Kalnis P, Tung A K H. Similarity evaluation on tree-structured data[C]// Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005: 754-765.
- [16] Liu Y, Molenaar M, Kraak M J. Semantic similarity evaluation model in categorical database generalization[J]. Symposium on Geospatial Theory, 2002, 34(4): 279-285.
- [17] Open GIS Consortium. OpenGIS® Catalogue Services Specification 2.0-ISO19115/ISO19119 Application Profile for CSW 2.0[S]. 2004b.

## Construction of Geospatial Metadata Association Network

Zhao Hongwei<sup>1,2,3</sup>, Zhu Yunqiang<sup>1,2</sup>, Hou Zhiwei<sup>1,2,3</sup>, Yang Hongwei<sup>4</sup>

(1. State Key Laboratory of Resources and Environmental Information System, Chinese Academy of Sciences, Beijing 100101, China; 2. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China; 4. China Petroleum Planning & Engineering Institute, Beijing 100000, China)

**Abstract:** The rapid acquisition of geospatial data mainly depends on geospatial metadata. But the traditional organization of geospatial metadata and the keywords-based retrieval methods create barriers among metadata considering semantic relations between geospatial data such as spatial topology relationship, category relationship, resulting in a bottleneck in geospatial data sharing. In the context of big geospatial data, the development of linked data provides an effective practice for the semantic sharing and application of massive geospatial data. The linked geodata is intended to break the semantic barriers between geospatial data and form a data network with semantic realtions. Due to the complexity, diversity and uncertainty of geospatial data, linked geodata is often achieved through the association between metadata. Geospatial metadata contains a number of descriptive information. How to effectively organize vast amounts of geospatial metadata and map the metadata into the semantic space by simple way have become the hotspots in the field of geospatial data sharing. Construction of semantic associations among geospatial metadata is an effective means of performing semantic retrieval using related data technologies. Effective application of linked data depends on effective association models. Considering this, a method of constructing geospatial metadata association networks is proposed in this paper: firstly, a geospatial metadata association model is designed on basis of the resource description framework (RDF); secondly, a semantic relation between metadata is determined and the relationship is constructed; and finally, the degree of semantic relevance of the semantic relationship is calculated. In the association network, the metadata are nodes, the semantic relationships between the metadata are edges, and the degrees of semantic relevance are the weights of the edges. Every node is an RDF that has attribute properties, such as sources, spatial characteristics, temporal characteristics, and content, and has properties of semantic relationships. Experimental results showed that the constructed network could effectively support operations such as semantic association search and recommendation, and the retrieval results were more precise and accurate compared with traditional metadata retrieval methods based on keywords.

**Key words:** geospatial metadata; linked data; semantic relevance; association network