引文格式：李雯静,王艺,谢邹玥,等. 基于本体的地理数据语义一致性处理方法研究[J]. 地理空间信息, 2026, 24(4):1-8. doi:10. 3969/j.issn. 1672-4623.2026.04.001

基于本体的地理数据语义一致性处理方法研究

李雯静，王　艺，谢邹玥，杨媛琦，邓子溪
（武汉科技大学 资源与环境工程学院，湖北 武汉 430081）

摘　要：由于应用需求的差异以及地理数据尺度变化过程中信息表达方式的不同，不同尺度的地图数据往往存在语义不一致的问题。这种不一致主要表现在实体的分类层次、属性描述和功能表达上的差异，可能导致数据融合与共享的困难，以及数据集逻辑的不连贯，从而影响决策依据的完整性或准确性。以居民地为研究对象，提出一种基于本体的语义相似度计算与一致性处理方法。通过构建居民本体模型生成改进的MD模型，计算实体间的语义相似度，并结合优先级覆盖原则对实体的语义信息进行处理，提高语义信息的一致性。以武汉市青山区部分居民地实体为例，验证方法的有效性，实验结果表明，本研究所述方法有助于提高实体的语义相似度，增强不同尺度地理数据间语义的一致性。

关键词：本体模型；语义相似性；多尺度地理数据；一致性处理
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2026）04-0001-08

Research on Semantic Consistency Processing Method for Geographical Data Based on Ontology

LI Wenjing, WANG Yi, XIE Zouyue, YANG Yuanqi, DENG Zixi
(School of Resources and Environmental Engineering, Wuhan University of Science and Technology, Wuhan 430081, China)

Abstract: Due to differences in application requirements and variations in information representation during geographical data scale changes, semantic inconsistencies often arise in different scales geographical data. These inconsistencies mainly manifest in differences in the entity classification hierarchy, attribute descriptions, and functional expressions, which may lead to difficulties in data integration and sharing, as well as logical discontinuities in datasets, ultimately affecting the completeness or accuracy of decision-making references. Taking residential areas as the research objects, we proposed a semantic similarity calculation and consistency processing method based on ontology. By constructing a residential ontology model to generate an improved MD model, we calculated the semantic similarity between entities. Then, we used a priority coverage principle to process semantic information of entities to ensure consistency. Taking residential entities in the Qingshan District, Wuhan City, Hubei Province for example, we verified the effectiveness of this method. Experimental result shows that the method described in this paper helps improve the semantic similarity of entities and ensures semantic consistency of information across geographical data at different scales.

Key words: ontology model, semantic similarity, multi-scale geographical data, consistency processing

多源地理数据融合中，语义一致性处理是关键。现有方法如规则转换、外部控制表映射等，能标准化属性语义并校正属性值[1]。在地理要素类层面，本体建模和语义映射可解决命名及分类冲突；地理实体层面，同名实体匹配和编码转换结合属性相似性计算与本体推理，可解决语义不一致问题[2-13]。但这些研究多集中于相同比例尺数据，跨尺度语义信息处理研究较少[14-17]。本文提出一种面向多尺度地理数据的语义一致性计算与处理方法，通过为本体属性赋予关系信息，引入功能单元与关系概念，构建综合语义相似性模型，计算实体间语义相似度，依据属性类型与功能信息规范化地理数据语义。

1　地理数据语义冲突分析

1.1　语义不一致的原因

应用需求的差异和地理数据的多尺度变化是造成地理数据语义不一致的一个主要原因。由于不同需求的导向，导致了对同一地理实体的语义描述存在冲突。地理数据的多尺度变化也是语义不一致的重要原因。

1.2　语义不一致性表现形式

1.2.1　分类层次差异

由于不同矢量数据采用的分类分级依据、标准的不一致，地理实体在分类和分级上存在显著差异[18-20]。

项目来源：武汉市重点研发计划资助项目（2024050702030122）。
第一作者简介：李雯静（1978—），博士，教授，研究方向为地图综合、空间信息挖掘、智慧矿山、矿山安全等，E-mail：Liwenjing@wust.edu.cn 。

1.2.2 属性描述差异

地理数据尺度变化所引发的简化处理或属性信息采集过程中的差异，使得实体在不同地理数据中呈现出属性信息表达的不一致。这种差异既可能表现为属性值的精度偏差，也可能反映为信息内容的缺失或概化，从而导致同一实体在语义表达上出现逻辑错误。

1.2.3 实体功能差异

尺度变化引发的信息概化或功能细分，使得同一地理实体在不同尺度下可能被赋予不同的功能。这种功能的差异反映了地理数据在表达层次上的语义变化，增加了多尺度数据在语义一致性处理中的复杂性和挑战性。

1.3 语义不一致引起的问题

不同尺度的数据对于同一地理实体的描述上可能存在语义异质的问题，不仅影响了地理数据的融合与共享，更限制了地理信息系统的效能和应用范围[21]。该问题主要体现在以2个方面：①数据的融合与共享困难；②数据集的逻辑不连贯。

本文针对多尺度地理数据的语义冲突问题，以居民地实体为例，结合多尺度地理数据的多义性、歧义性的语义特点，提出了一种基于居民地本体模型的语义相似度计算与处理的方法。

2 居民地语义相似度模型构建

居民地是面状地理数据的重要组成部分之一，建立居民地本体模型是对居民地实体及其专题属性的系统化描述，通过结构化和形式化的方式表达居民地的概念及其语义关系，为语义冲突问题提供了解决路径，并为相似度计算与多尺度地理数据语义处理提供依据。

2.1 居民地本体模型的构建

结合基础地理国情地理要素数据规定[19-20]和天津市地理要素数据字典[22]中对居民地实体的分类与语义描述确定居民地实体包含的语义层次内容，见图1。

2.2 MD模型的原理及改进

MD（Matching-Distance）模型基于本体论，从实体的属性、功能和组成部分3个维度出发，提供了一个有效框架计算实体间的语义相似度[23]。MD模型的核心是通过对实体类的功能相似度、属性相似度和组成部分相似度的计算来衡量不同尺度地理数据实体之间的语义相似性[24-25]，其相似性总值的计算公式如下。

$$S(c_1,\ c_2)=\omega_p\times S_p(c_1,\ c_2)+\omega_f\times S_f(c_1,\ c_2)+\omega_a\times S_a(c_1,\ c_2) \tag{1}$$

式中，c_1，c_2 为不同尺度下的同一实体，即在不同尺度地理数据中对应的同名实体；$S(c_1,\ c_2)$ 为实体间的语义相似度；S_p，S_f，S_a 分别为 c_1，c_2 的属性相似度、功能相似度以及组成部分相似度；ω_p，ω_f，ω_a 分别为 S_p，S_f，S_a 的权重。

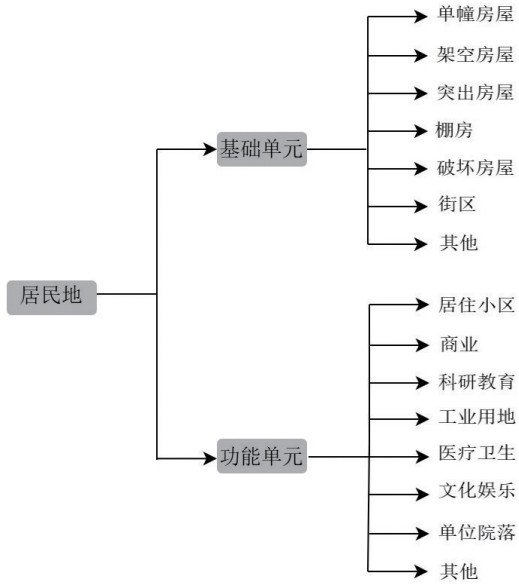


图1 居民地本体结构

本文针对居民地实体组成部分（地基、屋顶、墙）相似性较高，对语义相似度的判断影响有限的问题[26]，去除了MD模型中的组成部分维度。并考虑到MD模型在语义相似度计算中未考虑实体间关系可能导致实体语义关联缺失的问题，在模型中引入实体间关系 $R(c_1,\ c_2)$，以降低语义一致性修复的难度和成本。

改进后的MD计算公式如式（2）、（3），在式（2）中，c_1，c_2 为不同尺度地理数据中的同名实体，$S(c_1,\ c_2)$ 为实体间的语义相似度；R 为实体间的关系值；S_p，S_f 分别为 c_1，c_2 的属性相似度以及功能相似度；ω_p，ω_f 分别为 S_p，S_f 的权重。本研究邀请6位专家对属性与功能的重要性进行评分，基于平均打分结果构建AHP判断矩阵（属性：功能≈1.38:1），计算得出属性权重 ω_p 与功能权重 ω_f 分别为0.58与0.42。式（3）中，$\delta_{\text{part-of}}(c_1,\ c_2)$、$\delta_{\text{equal-to}}(c_1,\ c_2)$、$\delta_{\text{Compound}}(c_1,\ c_2)$ 分别为实体 c_1，c_2 之间的包含关系、同义关系和复合关系的量化值，取值为0或1。

$$S(c_1,c_2)=R(c_1,c_2)\times\left(\omega_p\times S_p(c_1,c_2)+\omega_f\times S_f(c_1,c_2)\right) \tag{2}$$

$$R(c_1,c_2)=\begin{cases}1,\delta_{\text{part-of}}(c_1,c_2)=1or\delta_{\text{equal-to}}(c_1,c_2)=1or\delta_{\text{Compound}}(c_1,c_2)=1\\0,\delta_{\text{part-of}}(c_1,c_2)=0and\delta_{\text{equal-to}}(c_1,c_2)=0and\delta_{\text{Compound}}(c_1,c_2)=0\end{cases} \tag{3}$$

面向基础地理信息领域，利用属性枚举法表达实体的本质语义特征[27]。本体模型中，将属性划分为基础属性和扩展属性，以描述实体特性与功能[28]。基础属性描述实体的核心信息，扩展属性在此基础上细化为专题属性和功能属性，其中专题属性描述实体的语义特性，功能属性则描述其具体用途。针对改进的MD模型的三个维度，描述如下：

1）关系：对于式（2）中的 $R(c_1, c_2)$ 为实体间关系值通过几何匹配的结果确定，在空间位置上识别出大小比例尺地理实体之间的匹配情况后，这些匹配类型在本体中被标记为同义关系、包含关系、上下位关系等[25]。本文仅讨论与语义一致性相关关系：同义关系（$\delta_{\text{part-of}}$）、包含关系（$\delta_{\text{equal-to}}$）与复合关系（δ_{Compound}），任意2个实体间最多只能存在一种关系。

2）属性：式（2）中，$S_p(c_1, c_2)$ 表示实体 c_1，c_2 间的基础属性与专题属性的相似度。基础属性描述实体的基础信息，这些信息是实体的核心标识和最基本的描述特征。专题属性聚焦实体在语义上的特性，描述与其用途相关的详细信息，专题属性与基础属性共同构成了实体的语义属性特征体系。

3）功能：式（2）中，$S_f(c_1, c_2)$ 为实体 c_1，c_2 间的功能相似度。将其与其他属性分开处理，并赋予更高的权重，以更合理地反映实体间的语义关联性。

3　语义相似度计算与一致性处理

基于第2节中构建的本体以及改进的MD模型，本节从关系、属性和功能3个维度出发，计算实体间语义相似度，并设计差异化处理方法对实体语义的不一致进行处理。

3.1　语义相似度计算

根据改进的MD模型，实体间的关系值 $R(c_1, c_2)$、属性相似度 $S_p(c_1, c_2)$、功能相似度 $S_f(c_1, c_2)$ 计算方法如下。

3.1.1　关系值计算

为确定实体间的相似度，首先从空间关系上确定实体间的匹配类型，为下一步进行语义相似度计算提供依据，本文采用正反向匹配策略[29]：正向匹配从源数据识别目标实体，反向匹配通过比较面积重叠比例验证一致性，将超过阈值的匹配对认定为有效匹配对，确定为同名实体。

在空间位置上识别出大小比例尺实体之间的匹配类型后，这些类型在本体中分别被标记为同义关系、包含关系、复合关系，当存在以下3种匹配类型时，$R(c_1, c_2)=1$。

当实体间为一对一的匹配类型时，在本体中被定义为同义关系：即 $\delta_{\text{equal-to}}(c_1, c_2)=1$；当实体间为一对多的匹配类型时，在本体中被定义为包含关系：即 $\delta_{\text{part-of}}(c_1, c_2)=1$；当实体间为多对多的匹配类型时，在本体中被定义为复合关系：即 $\delta_{\text{Compound}}(c_1, c_2)=1$。

在上述3种情况中，均有 $R(c_1, c_2)=1$，如若实体间的功能与属性存在不同，则实体的语义相似度取值在0~1之间。若实体间几何匹配为空，则确定实体间不存在任何语义关系，$R(c_1, c_2)=0$，相似度计算无意义，因此直接令 $S(c_1, c_2)=0$。

3.1.2　属性相似度计算

根据属性表现出的数据类型，本文将实体属性划分为字符型、布尔型和数值型，并结合本体结构中概念与属性名称，将实体属性的语义特征划分为累加型和非累加型。累加型属性用于表示可累加的量度如面积、长度和容量等。非累加型属性用于表示不同尺度地理数据之间等价或相似的特性，适用于如使用年限、最高高度和分类类型等不具备累加性质的属性。

针对不同属性类型，采用相应的计算方法，并综合权重因素，计算出实体对的整体属性相似度，具体如式（4）所示，式中 ω_i 为第 i 个属性的权重，$S_{pi}(c_1, c_2)$ 表示实体第 i 个属性的相似度。

$$S_p(c_1, c_2)=\sum_{i=1}^{n}\omega_i \times S_{pi}(c_1, c_2) \tag{4}$$

1）字符型属性相似度。地理数据的属性信息往往呈现为非结构化的文本形式，格式各异，且字符型属性值常存在词语变形与语义模糊等问题，短文本信息（少于140字的文本）在处理实体属性时占据了较大的比例。针对这些文本情况，利用Word2Vec技术能够有效解决此类问题[30]。通过Word2Vec将词语映射为高维语义向量后，采用余弦相似度式（5）计算其语义相似性。

$$S(\boldsymbol{A}_1, \boldsymbol{A}_2)=\frac{\sum_{i=1}^{n}\boldsymbol{A}_1(i)\times\boldsymbol{A}_2(i)}{\sqrt{\sum_{i=1}^{n}\boldsymbol{A}_1(i)^2}\times\sqrt{\sum_{i=1}^{n}\boldsymbol{A}_2(i)^2}} \tag{5}$$

式中，$S(\boldsymbol{A}_1, \boldsymbol{A}_2)$ 为实体属性的相似度；$\boldsymbol{A}_1(i)$、$\boldsymbol{A}_2(i)$ 分别为向量 $\boldsymbol{A}_1$和$\boldsymbol{A}_2$ 在第 i 个分量上的值；$\sum_{i=1}^{n}\boldsymbol{A}_1(i)\times\boldsymbol{A}_2(i)$ 为2个向量的点积；$\sqrt{\sum_{i=1}^{n}\boldsymbol{A}_1(i)^2}$ 为向量 $\boldsymbol{A}_1$

的范数；$\sqrt{\sum_{i=1}^{n}\boldsymbol{A}_2(i)^2}$ 为向量 $\boldsymbol{A}_2$ 的范数。

2）布尔型属性相似度：布尔型属性值可以直接用逻辑比较，式（6）中 $S(\boldsymbol{A}_1,\boldsymbol{A}_2)$ 为实体属性的相似度，$\boldsymbol{A}_1$，$\boldsymbol{A}_2$ 分别为布尔型属性的属性值。

$$S(\boldsymbol{A}_1,\boldsymbol{A}_2)=\begin{cases}0, & \boldsymbol{A}_1=\boldsymbol{A}_2\\ 1, & \boldsymbol{A}_1\neq\boldsymbol{A}_2\end{cases} \tag{6}$$

3）数值型属性相似度：对于无法确定属性数值上下限的实体来说，可使用归一化相对差异度量的方法，考虑2个数值相对平均值的偏差。式（7）中 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$ 分别为该数值型属性的属性值。

$$S(\boldsymbol{A}_1,\ \boldsymbol{A}_2)=1-\frac{|\boldsymbol{A}_1-\boldsymbol{A}_2|}{\max(\boldsymbol{A}_1,\boldsymbol{A}_2)} \tag{7}$$

对于累加型和非累加型的属性值，$\boldsymbol{A}_1$ 的取值如式（8），$\boldsymbol{A}_{\text{total}}$ 代表所有相关属性的值之和。

$$\boldsymbol{A}_1=\begin{cases}\boldsymbol{A}_1, & \text{非累加型属性}\\ \boldsymbol{A}_{\text{total}}, & \text{累加型属性，}\boldsymbol{A}_{\text{total}}=\sum_{i=1}^{n}\boldsymbol{A}_i\end{cases} \tag{8}$$

3.1.3　功能相似度计算

本文将功能划分为一级功能和二级功能，如图2所示。

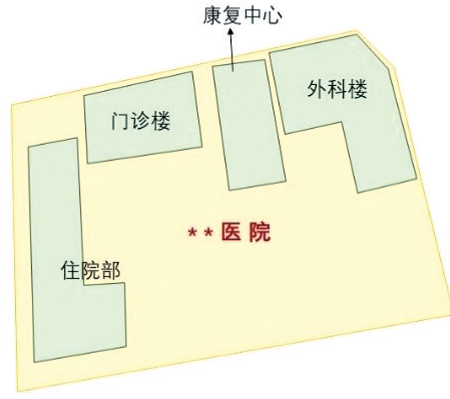


图2　语义功能区内的建筑物

表1展示了部分实体类型及其功能属性的分类结构。

表1　部分实体类功能构成

实体类型	一级功能属性	二级功能属性
医疗卫生	医疗卫生	门诊、预防和社区卫生服务、检（防）疫站、保健疗养、医学化验与药品检验、教学与科学研究、其他
居住小区	居住小区	住宅、社区服务、其他
教育科研	教育科研	幼儿园、小学教育、中学教育、大学、职业技术学院、培训机构、研究所、实验室、图书馆、文化中心、其他

当实体间存在同义关系和包含关系时，实体功能依据分类层次将其分为同属一级功能和不同属一级功能2种类型，并赋予相应的相似度值。复合关系则将多对多的匹配情况转化为“一对多”的关系，即将一个实体与另一实体集合配对，对所有功能的相似度进行加权，得到整体的功能相似度。本文结合构建的居民地本体结构，采用专家经验赋值法设定功能相似度值，具体功能相似度值如表2所示。

表2　实体功能相似度规则

实体关系	一级功能	二级功能	相似度值
同义关系/包含关系	相同	相同	0.9
		不同	0.6
	不同	在功能列表中	0.2
		不在功能列表中	0

3.1.4　整体相似度计算

通过3.1.1、3.1.2、3.1.3节中的方法，分别计算出单个实体间的关系相似度 $R(c_1,\ c_2)$、属性相似度 $S_p(c_1,\ c_2)$、与功能相似度 $S_f(c_1,\ c_2)$，并通过改进的MD模型，以及规定的权重系数计算出单个实体间的语义相似度 $S(c_1,\ c_2)$。

以地理数据的尺度变化为例，尺度变化前后经几何匹配得到的同名实体，其数量存在差异。设小比例尺数据中的实体集合为 C_{i_small}，包含的实体数目为 n；大比例尺数据中的对应实体集合为 C_{i_large}，包含的实体数目为 m，两者实体数目比为 n:m（$n\geqslant 1$，$m\geqslant 1$；$m\geqslant n$）。为计算实体对的整体相似度，本文以尺度变化后某个小比例尺地理实体为目标，并以大比例尺数据中对应的地理实体为参照，依照匹配关系区分以下3种情况：

当 $n=1$，$m=1$ 时实体对间存在同义关系，即单个实体相似度等于整体相似度；

当 $n=1$，$m\geqslant 1$ 时，实体对间存在包含关系；

当 $n\geqslant 1$，$m\geqslant 1$ 时，实体对间存在复合关系。

对于上述3种实体关系，实体对的相似度为 $S(C_{i_\text{small}},C_{i_\text{large}})$ 采用下式（9）计算。

$$S\left(C_{i_\text{small}}, C_{i_\text{large}}\right)=\frac{\sum_{i=1}^{m}\sum_{i=1}^{n}S_i\left(C_{i_\text{small}}, C_{i_\text{large}}\right)}{mn} \tag{9}$$

3.2　多尺度地理数据语义一致性处理

在3.1节语义相似度计算的基础上，本文设定语义相似度阈值Z，当相似度低于该阈值时触发一致性处理机制，具体判断规则如式（10）所示：

$$\begin{cases} S\left(C_{i_\text{small}},\ C_{i_\text{large}}\right)\geqslant Z, \text{语义一致无需处理} \\ S\left(C_{i_\text{small}},\ C_{i_\text{large}}\right)<Z, \text{需进行属性与功能一致性处理} \end{cases} \tag{10}$$

通过几何匹配确定的实体关系值 $R(c_1,\ c_2)$ 仅作为计算实体间语义相似度的参考，无需对实体关系本身进行进一步处理。

当地理数据的尺度发生变化时，实体会被赋予不同的功能用途，实体的功能信息表达也会随之改变。这种变化符合现实世界的实际情况，即功能相似度 $S_f(c_1,c_2)\geqslant 0.6$ 时（一级功能相同），无需对其进行语义的一致性处理。当 $S_f(c_1,c_2)<0.6$ 时（一级功能不同），交由人工进行处理。

对于实体的属性信息一致性处理，本文采用优先级覆盖原则，以大比例尺数据作为现势性较高的数据。当不同尺度的地理数据经过几何匹配建立实体对应关系后，依据属性类型进行分类处理。本文引入属性一致性处理阈值 t，用于部分属性的一致性判断与处理。

1）字符型数据。对于非累加型属性，依据属性值的共性采用前缀/后缀提取、词频统计或多数值覆盖方式处理，若无明显共性则交由人工判断；累加型属性通过合并去重后统一赋值，以保障跨尺度属性表达的一致性。

2）布尔型数据。布尔型数据用于表示“是/否”类型的属性值，例如“是否为保护区”“是否有历史遗址”等。对于累加型和非累加型的布尔型属性的一致性处理方法一致：判断大比例属性数据中“是”或“否”的占比是否超过阈值 t，若超过则采用多数结果覆盖小比例尺数据。

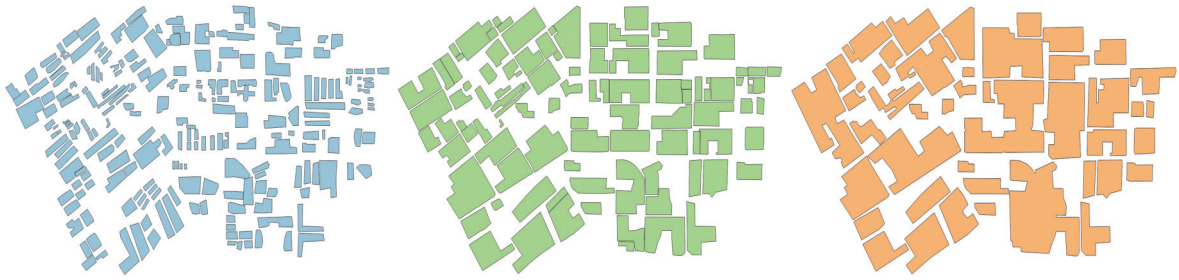
3）数值型数据。对非累加型属性，判断某一属性值在所有大比例尺数据中的占比是否超过70%，超过则用该属性值对小比例尺数据进行覆盖；累加型数值属性通过累加大比例尺属性值，并按比例分配至小比例尺实体。

4　实验与分析

为验证本文所述方法的有效性，实验选取武汉市青山区部分居民地数据，评估本文方法在多尺度场景下的适应性与表现。对于属性处理阈值 t 和语义相似度阈值 Z，开展敏感性分析以确定其最优值。并以“东兴天地”实体为例展示语义相似度计算与一致性处理过程。为系统评估方法效果，采用Jaccard系数和余弦相似度对处理前后全部实体对的语义相似度进行对比。

4.1　实验数据选择与分析

实验数据如图3所示，该数据包含“居住小区”等多个功能单元。其中大比例尺数据信息更丰富、现势性更高，可用于补充和更新小比例尺数据，从而提升数据现势性与语义一致性。



a 1∶1 250居民地数据（图层A）　b 1∶5 000居民地数据（图层B）　c 1∶20 000居民地数据（图层C）

图3　相同实验区域不同尺度的居民地数据

采用3.1.1节中提到的正反向匹配方法，对实验区域进行几何匹配。表3对各比例尺组合下的匹配类型数量进行了统计。

表3　不同比例尺组合下地理实体匹配关系统计表

组别	比例尺组合	一对一实体对数量	一对多实体对数量	多对多实体对数量	同名实体对数量	未匹配实体数
组Ⅰ	1∶1 250与1∶5 000	6	63	5	74	7
组Ⅱ	1∶1 250与1∶20 000	2	41	0	43	5
组Ⅲ	1∶5 000与1∶20 000	20	19	2	41	2

4.2　地理数据语义相似性计算

对于匹配到的实体对进行语义相似度的计算与一致性处理，未匹配到的实体交由人工进行处理，判断是否保留其现有的语义信息。“居住小区”功能单元的属性名称及其属性类型见表4，各属性权重由专家

表4 居民小区属性类型及权重

实体属性	基础属性					扩展属性				
	名称	地址	面积	所属区域	容积率	产权年限	总户数	所属商圈	权属类别	停车位
属性类型	字符型	字符型	数值型	字符型	数值型	数值型	数值型	字符型	字符型	布尔型
累加 非累加	非 累加	非 累加	累加	非 累加	非 累加	非 累加	累加	累加	非 累加	非 累加
属性权重	0.076	0.084	0.111	0.148	0.101	0.063	0.121	0.092	0.135	0.069

打分并通过层次分析法确定。

本文对于组Ⅰ、组Ⅱ、组Ⅲ中的同名实体对，分别计算其平均属性相似度、平均功能相似度、平均语义相似度与标准差如表5所示。

表5 不同比例尺组合相似度计算结果

比例尺组别	同名实体对数量	平均属性相似度	平均功能相识度	平均相似度	标准差
组Ⅰ	74	0.642	0.575	0.614	0.041
组Ⅱ	43	0.451	0.408	0.433	0.085
组Ⅲ	44	0.511	0.519	0.514	0.059

4.3 语义一致性处理与分析

4.3.1 属性处理阈值 t 的确定

为了确保属性一致性处理阈值 t 的合理性与适应性，本文设计了基于多尺度实体样本的对比实验，评估不同处理强度下语义相似度的提升效果，确定最优参数 t^*，作为后续一致性处理的基础参数。

考虑到地理数据的多尺度特征，实验选取组Ⅰ、组Ⅱ、组Ⅲ3个比例尺匹配场景，每种匹配场景人工选取10对属性相似度较低的同名实体，作为参数评估样本集。本文将处理阈值 t 设置在50%~80%区间，并以10%为步长依次测试。该区间覆盖了语义相似度的模糊过渡带：t 低于50%时属性差异过大，不具备有效处理基础；高于80%则对属性一致性要求过高，实际应用中难以满足，反而限制了处理效果。

实验结果表明随着处理强度的增加，语义相似度先升至峰值后逐渐降低。适度的处理阈值可有效缓解异构属性所导致的语义分歧，提升实体间语义一致性；当处理阈值超过一定范围后，属性差异被过度统一，削弱了实体语义表达能力，从而使相似度提升趋缓或出现回落。在3组数据中，当 $t=60\%$ 时均实现了较为显著且稳定的语义相似度提升，其中2组数据相似度达到最高值，另1组略低于最高值，具有较强的跨尺度适应能力。因此，本实验将一致性处理阈值 t^* 设为60%，作为在多比例尺匹配场景下的属性一致性处理的最优值。

4.3.2 语义相似度判断阈值 Z 的确定

基于处理强度参数 t^*，本文进一步开展语义相似度阈值 Z 的敏感性分析。实验针对3组比例尺组合，设置 Z 分别为0.5、0.6、0.7、0.8，对所有相似度低于阈值的实体对进行属性与功能一致性处理，并计算处理前后的平均语义相似度（见表6）。其中，0.5为语义相似度可接受下限，0.8虽理想但处理成本较高，为平衡效果与效率，将步长设为0.1，结果如表6所示。

表6 阈值 Z 对语义相似度的影响

比例尺组别	相似度阈值Z	处理前相似度	处理后相似度	变化率/%
组Ⅰ	0.5	0.614	0.623	1.5
	0.6		0.641	4.4
	0.7		0.729	18.7
	0.8		0.737	20.0
组Ⅱ	0.5	0.433	0.576	33.0
	0.6		0.584	34.9
	0.7		0.602	39.0
	0.8		0.613	41.6
组Ⅲ	0.5	0.514	0.582	13.2
	0.6		0.673	30.9
	0.7		0.685	33.7
	0.8		0.708	37.7

实验结果表明实体语义相似度在处理阈值 Z 提高的过程中呈现先上升后趋于平稳的趋势。当阈值 Z 略高于该组合的初始语义平均相似度时，提升幅度最大且较为稳定。

4.4 实验结果分析

基于上述实验结果，本文统一设定属性一致性处理强度参数 $t^*=60\%$，并结合各比例尺组合初始相似度均值情况，分别设置语义相似度判断阈值：ZⅠ为0.7，ZⅡ为0.5，ZⅢ设为0.6。

4.4.1 语义一致性处理实例分析

本文选取组Ⅰ中的“东兴天地”实体对为例，进行语义相似度计算与一致性处理。

1）实体关系值：通过几何匹配后确认，东兴天地小区与东兴天地A区、东兴天地B区和东兴天地C区建立1：3的匹配关系，确定实体间的关系为包含关

系，$\delta_{\text{part-of}}(c_1, c_2)=1$。

2）实体属性相似度：表7为“东兴天地”实体的属性详细信息，根据前文中公式（4）~（8）中的属性相似度计算方法，并分别计算实体的字符型属性、数值型属性和布尔型属性的相似度，最终综合为总体属性相似度 $S_p(c_1, c_2)=0.581$。

表7　不同比例尺“东兴天地”实体属性

比例尺	名称	地址	权属类别	所属区域	所属商圈	产权年限/年	总户数/户	面积/㎡	容积率	停车位
1∶5 000	东兴天地	青山区随州街与工业四路交叉口东60 m	普通住宅	青山区	Null	70	1 813	26 665.6	1.82	是
	东兴天地A区	青山区冶金大道随州街	公寓住宅	青山区	冶金大道	70	671	14 136	1.82	是
1∶1 250	东兴天地B区	青山区友谊大道随州街	公寓住宅	青山区	友谊大道	70	435	6 987	1.82	是
	东兴天地C区	青山区友谊大道随州街30号	公寓住宅	青山区	友谊大道	70	740	7 877	1.82	是

3）实体功能相似度：在小比例尺数据中，东兴天地小区的一级功能为“居住小区”，二级功能为空；在大比例尺数据中，东兴天地A区、B区、C区一级功能为“居住小区”，二级功能为“住宅”。实体的一级功能相同，二级功能不同，$S_f(c_1, c_2)=0.6$。

综上所述，该同名实体的 $S_p(c_1, c_2)=0.581$，$S_f(c_1, c_2)=0.6$，根据2.2节中给定的阈值，根据式（2），该实体对的语义相似度 $S(c_1, c_2)=0.589$，低于 $ZI=0.7$，需要进行语义一致性处理。

该实体的功能相似度≥0.6时，无需进行处理；实体的属性信息，根据本文所述方法进行一致性处理。表8展示了该实例小比例尺地理数据处理前后属性值的对比，以说明语义一致性处理的效果。

表8　“东兴天地”属性一致性处理结果

属性名称	原属性值	一致性处理后的属性值	处理方式
名称	东兴天地	东兴天地	—
地址	青山区随州街与工业四路交叉口东60 m	青山区友谊大道随州街	字符非累加型属性，采用词频法提取出公共信息覆盖小比例尺数据
面积	26 665.6 m²	29 000 m²	数值累加型属性，对大比例尺数据加权覆盖小比例尺数据
所属区域	青山区	青山区	—
容积率	1.82	1.82	—
产权年限	70年	70年	—
总户数	1 813	1 846	数值累加型属性，对大比例尺数据加权覆盖小比例尺数据
所属商圈	Null	友谊大道、冶金大道	字符累加型属性，大比例尺数据整合去重后覆盖小比例尺的空数据
权属类别	普通住宅	公寓住宅	字符非累加型属性，大比例尺数据一致覆盖小比例尺数据
停车位	是	是	—

4.4.2　语义一致性处理效果评估

为验证本文提出方法的有效性，本文选取Jaccard系数、余弦相似度2种语义相似度计算方法，评价处理前后的实体语义相似度。

Jaccard系数是一种用于衡量2个集合相似度的统计量，通常用于比较样本集之间的相似性和差异性，数值范围为0到1，0表示2个集合完全没有交集，1表示2个集合完全相同。余弦相似度是一种常用于衡量2个向量在多维空间中的方向相似性的方法，广泛应用于文本挖掘和自然语言处理中，余弦相似度通过测量两个向量间角度的余弦值来评估它们的相似性，其值范围从–1（完全相反）到1（完全相同），其中0表示两者独立，没有相似性，表9展示了处理前后Jaccard系数、余弦相似度变化。

表9　处理前后语义相似性结果

组别	Jaccard系数（处理前）	Jaccard系数（处理后）	余弦相似度（处理前）	余弦相似度（处理后）
组Ⅰ	0.538	0.611	0.694	0.743
组Ⅱ	0.421	0.469	0.587	0.623
组Ⅲ	0.467	0.519	0.612	0.654

通过实验分析，本文提出的语义一致性处理方法有效提高了地理实体数据的语义相似度。结果显示，经过此方法处理的数据，在Jaccard系数和余弦相似度2个指标上均有提升。表明该方法能够有效增强数据的语义一致性，深化地理实体之间的语义关联，从而为地理信息系统中的地理数据分析提供了更为准确和可靠的语义支持。

5 结 语

本文提出了一种面向不同比例尺地理数据的语义一致性处理方法，通过词频分析、聚合多数决策和数值属性合并等策略，为比例尺差异引起的语义不一致问题提供了解决思路。实验结果表明，该方法能够提高数据的现势性，同时保证信息在不同尺度间的语义一致性，为地理信息科学中数据一致性的研究提供了新的技术支持与研究视角。

参考文献

[1] 孙群,温伯威,陈欣. 多源地理空间数据一致性处理研究进展[J]. 测绘学报,2022,51(7):1561-1574

[2] 李勤超.基于本体的地理信息语义转换模型与方法研究[D]. 郑州:信息工程大学,2011

[3] 赵云鹏,孙群,朱蕊,等. 一种顾及描述知识的地理要素分类映射方法[J]. 测绘科学技术学报,2020,37(5):531-536

[4] Kuai X, Li L, Luo H, et al. Geospatial Information Categories Mapping in a Cross-lingual Environment:A Case Study of “Surface Water” Categories in Chinese and American Topographic Maps[J]. ISPRS International Journal of Geo-information,2016,5(6):90-111

[5] Li W, Raskin R, Goodchild M F. Semantic Similarity Measurement Based on Knowledge Mining:An Artificial Neural Net Approach[J]. International Journal of Geographical Information Science,2012,26(8):1415-1435

[6] 赵云鹏.多源矢量数据语义一致性处理关键技术研究[D]. 郑州:信息工程大学,2021

[7] 郭黎.多源地理空间矢量数据融合理论与方法研究[D]. 郑州:信息工程大学,2008

[8] 王少一,苏绣,于海波.增强语义的地理信息服务匹配相似性算法探讨[J].测绘科学,2015,40(3):119-122

[9] 吴玉婷,肖提荣,何照攀. 多源数据间线状地物属性信息传递方法[J]. 测绘技术装备,2018,20(4):59-61

[10] 张彭涛.面向多源矢量道路和建筑物数据的关联匹配与属性融合研究[D].南京:南京师范大学,2019

[11] Ding Y, Guo Y H, Lu W, et al. Context-aware Semantic Type Identification for Relational Attributes[J]. Journal of Computer Science and Technology,2023,38(4):927-946

[12] 靳滢. 基于深度学习的地理关系抽取系统研究与实现[D]. 北京:北京邮电大学,2019

[13] 蒋秉川,万刚,许剑,等. 多源异构数据的大规模地理知识图谱构建[J]. 测绘学报,2018,47(8):1051-1061

[14] Schwering A. Approaches to Semantic Similarity Measurement for Feo-spatial Data:A Survey[J]. Transactions in GIS,2008,12(1):5-29

[15] 杜萌,赵萍,张浩,等.多源时空数据的地表分类信息挖掘[J]. 测绘通报,2021(5):10-14

[16] 唐忠成,何元禹,张允. 空间数据融合技术在编制数字现状图中的应用[J]. 地理空间信息,2021,19(6):101-103

[17] 陈换新,肖强,李明,等. 国外空间数据的语义差异及转换方法研究[J]. 测绘科学,2014,39(7):83-86

[18] 朱蕊.多源空间矢量数据一致性处理技术研究[D].郑州:信息工程大学,2012

[19] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会.基础地理信息要素数据字典第1部分1∶500 1∶1 000 1∶2 000基础地理信息要素数据字典:GB/T20258.1-2007[S].

[20] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. 基础地理信息1∶10 000地形要素数据规范:GB/T33462-2016[S].

[21] Sun K, Zhu Y, Pan P, et al. Geospatial Data Ontology:The Semantic Foundation of Geospatial Data Integration and Sharing[J]. Big Earth Data,2019,3(3):269-296

[22] 天津市市场监督管理委员会. 天津市基础地理信息要素数据字典 第1部分:1∶500 1∶2 000 基础地理信息要素数据字典:DB 12/T 473.1-2021[S].

[23] Rodriguez M A, Egenhofer M J, Rugg R D. Assessing Semantic Similarities Among Geospatial Feature Class Definitions[C]//Interoperating Geographic Information Systems:Berlin, Heidelberg:Springer Berlin Heidelberg, 1999, 189-202

[24] Tversky A. Features of Similarity[J]. Psychological Review,1977,84(4):327-352

[25] 丁虹. 空间相似性理论与计算模型的研究[D]. 武汉:武汉大学,2004

[26] 高晓蓉,闫浩文,禄小敏. 多尺度地图空间居民地语义相似度计算方法[J]. 测绘学报,2022,51(1):95-103

[27] 谭永滨,李霖,王伟,等. 本体属性的基础地理信息概念语义相似性计算模型[J]. 测绘学报,2013,42(5):782-789

[28] Li W, Ma Q, Tang Y, et al. Entity-driven New Paradigm of Mine Data:Model Construction and Application[J]. ISPRS International Journal of Geo-information, 2024, 13(12): 442-464

[29] 殷甲伟. 基于地理实体匹配的多尺度空间数据联动更新研究[D]. 武汉:武汉大学,2018

[30] 李晓,解辉,李立杰. 基于Word2vec的句子语义相似度计算研究[J]. 计算机科学,2017,44(9):256-260