地理空间数据相似度计算方法研究与实现

代小亮[1,2]，诸云强[1,3*]，杨　杰[1]，孙　凯[1]，李吉东[4]，宋　佳[1,3]

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；3. 江苏省地理信息协同创新中心，南京 210023；
4. 东营市生态环境局，东营 257091

摘　要：地理空间数据相似度计算是数据智能推荐与发现的关键技术之一。现有地理空间数据相似度计算方法可分为基于文件信息、元数据或数据实体的方法，这些方法各有优缺点，但仅用其中一类信息进行相似度的计算，往往会存在信息项缺失而导致计算结果不准或计算量大等问题，进而影响数据相似度的应用价值。为充分发挥各类方法的优点，本文提出了一种集成文件信息、元数据、数据实体三个层次的地理空间数据相似度计算方法，并开发了相应的软件，可根据地理空间数据的实际情况，有效提升地理空间数据相似度的计算精度和效率。

关键词：地理空间数据；数据相似度；计算方法；软件

DOI: https://doi.org/10.3974/geodp.2022.04.01

CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2022.04.01

数据可用性声明：

本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版，可获取：
https://doi.org/10.3974/geodb.2022.10.02.V1 或 https://cstr.escience.org.cn/CSTR:20146.11.2022.10.02.V1.

1　前言

地理空间数据是地理实体和现象的空间特征和属性特征的数字表达，是地球科学研究的基本要素之一[1]。地理空间数据在自然科学和社会科学研究中发挥着重要作用，已被广泛应用于应急管理[2]、环境监测[3]、自然灾害预测[4]和人口经济研究[5]等诸多领域。地理空间数据相似度计算是基于数据特征计算数据与数据之间的相似程度，是数据智能推荐与发现的关键技术之一，广泛应用于地理空间数据关联网络[6]、地理空间数据智能推荐[7]和地理空间模型自动匹配数据[8]等过程中。此外，类比于文本相似度用于文献查重，地理空间数据相似度计算还可以用于地理空间数据查重。

按照用于相似度计算的数据特征信息来源，现有的地理空间数据相似度计算方法分为三类：（1）基于文件信息的相似度计算方法。该方法主要通过数据文件的信息项，如文件

收稿日期：2022-11-09；**修订日期：**2022-12-19；**出版日期：**2022-12-24

基金项目：国家自然科学基金（42050101），中国科学院（XDA23100100）

***通讯作者：**诸云强L-6116-2016，中国科学院地理科学与资源研究所，zhuyq@lreis.ac.cn

数据引用方式：[1] 代小亮, 诸云强, 杨杰等. 地理空间数据相似度计算方法研究与实现[J]. 全球变化数据学报, 2022, 6(4): 501−512. https://doi.org/10.3974/geodp.2022.04.01. https://cstr.escience.org.cn/CSTR:20146.14.2022.04.01.

[2] 代小亮, 诸云强, 杨杰等. 地理空间数据相似度计算软件（GDSCS V1.0）[J/DB/OL]. 全球变化数据仓储电子杂志, 2022. https://doi.org/10.3974/geodb.2022.10.02.V1. https://cstr.escience.org.cn/CSTR:20146.11.2022.10.02.V1.

名、文件格式等进行数据相似度的计算。例如：孙娟娟等人利用文件名、文件长度、文件类型等文件信息，提出了面向 P2P 文件共享应用的相似度计算模型[9]；Kim 使用函数匹配的方法进行了二进制文件的相似度计算[10]；Kim 等人针对多媒体数据文件相似度计算，提出了部分哈希信息字符串算法[11]。此类方法计算简单、效率高，但能用于计算的数据特征较少，同时可能出现文件信息记录不准确（例如文件名称被修改）等情况。（2）基于元数据的相似度计算方法。该方法主要通过元数据的元数据项，如内容主题、时间范围、空间范围等进行数据相似度的计算。例如：Zhu 等人基于数据主题、类别、空间覆盖、时间覆盖、数据类型和数据格式等八个数据属性计算地理元数据相似度，以定量地互连地理数据[6]；Chen 等人使用人工神经网络基于数据的关键字、类别、空间覆盖和时间覆盖的数据属性来计算地理元数据的相似性，然后根据元数据的相似性推荐地理数据[3]。此类方法能够利用丰富的元数据较全面计算相似度，但有时存在元数据缺失、不完整及记录不精确等诸多的不确定性，例如最小外接矩形不能完全表达空间的数据范围；（3）基于数据实体的相似度计算方法。该方法主要通过数据实体的要素位置和属性信息进行数据相似度的计算。例如：梅耀元等通过研究点群的密度相似、面积相似以及空间方向的相似关系，建立了点状要素相似度的计算模型[12]；朱爽等基于颜色直方图相似关系进行了栅格图像的相似度计算[13]。此类方法利用数据实体的要素位置和属性信息，能够充分反映数据实体内容方面的相似度，但是部分信息项，例如矢量数据的时间信息，并不直接体现在数据实体中，同时由于是逐要素或像元计算，其计算量较大。

以上三类方法各有优缺点，目前的研究大多只采用其中的一类方法，这些方法要么准确率不高，或者计算量大、时间长；此外，由于缺乏必要的特征因子，而导致相似度无法计算的问题也会发生。本文提出了一种集成文件信息、元数据、数据实体三个层次的地理空间数据相似度计算方法，可综合利用各方法的优点，并根据数据信息及实际应用需求，灵活选择不同的方法及其组合，进而实现既快又准的数据相似度计算。通过对比三个层次相似度计算结果，作者认为它将有利于发现数据间的差异。作者为实现地理空间相似度的计算，配套开发了对应的计算机软件。

2 地理空间数据相似度计算

2.1 相似度因子选择

选择合适的相似度因子是相似度计算的前提。作者通过分析三个层次的特点，选择了对应的相似度因子。

用于文件信息相似度计算的信息项选取文件名称、文件格式、文件大小、文件数量等。文件名称是标识一个数据区别另一个数据的名称。不同的数据通常具有不同的数据格式。相同的数据其文件大小和数量通常是一致的。

地理空间元数据是描述地理空间数据的数据，主要包括：地理空间数据的内容主题、时间范围、空间范围等信息。目前已有许多不同的元数据标准，例如 ISO19115 地理信息元数据标准[14]、数字地理空间元数据内容标准（CSDGM）[15]、《地理信息元数据》(GB/T 19710——2005)[16]等。大部分的元数据标准都包括有内容主题、空间范围、时间范围、空间精度、时间粒度等元数据项，所以本文选择上述元数据项用于元数据相似度的计算。

地理空间数据实体主要包括：实体要素位置和属性两类数据特征。实体要素位置特征是地理空间数据实体在像元或图形上的位置信息，实体要素属性特征是指属性表中所蕴含的属性项和属性值。地理空间数据实体格式主要包括栅格和矢量两种类型。栅格数据实体要素位置相似度主要基于像元进行计算；由于矢量数据又可进一步分为点要素、线要素和面要素实体，其相似度计算需要分别按点、线、面要素展开。其中：点要素的实体要素位置相似度主要基于点群拓扑、分布范围方向关系和距离进行相似度计算，线要素和面要素的实体要素位置相似度主要基于线群或面群的拓扑关系、距离关系、方向关系和几何特征进行计算。

2.2　总体计算流程

地理空间数据相似度集成计算方法的总体流程如图 1 所示。算法的基本步骤是：（1）首先进行文件信息相似度计算。利用该方法进行数据相似度计算，简单、快速但得到的是粗粒度的相似度，不够精准；（2）然后进一步计算元数据相似度。该方法的前提是必须有元数据，计算量适中，得到的是中粒度的相似度，但依赖于元数据的质量，可能有元数据缺失或描述不准确的问题；（3）最后，进一步计算数据实体相似度。数据实体相似度精度高，得到的是细粒度的相似度，但计算量大。

2.3　相似度计算方法

2.3.1　文件信息相似度计算方法

（1）文件哈希值相似度计算方法

文件哈希值是用哈希算法根据文件名称、大小、格式等文件信息生成的唯一字符串，文件信息完全相同的文件在同一哈希算法下会生成相同的哈希值，但文件信息有微小的差异，其哈希值会相差很大。所以，文件哈希值（ S_{hash} ）可以快速判断文件信息是否相同。本文用 sha256 哈希算法[17]生成数据文件哈希值，当两个哈希值相同时，则文件哈希值相似度为 1，否则哈希相似度为空（None），需要继续计算文件名称、格式、大小和数量等相似度。

（2）文件名称相似度计算方法

文件名称的文本较短，通常采用基于字符串的语义文本相似度算法进行文件名称一致性的判断。通过分析常用的基于字符串的文本相似度计算方法的特点，本文选择编辑距离（Minimum Edit Distance，MED）算法[18]用于文件名称相似度计算。

设 U_A,U_B 为两个数据的文件名称，l_{U_A},l_{U_B} 分别为两个文件名称的长度，即字符串包含字符的个数，$D(U_A,U_B)$ 为文件名称 A,B 的编辑距离，则地理空间数据的文件名称相似度用 S_{name} 用公式(1) [19]。

$$S_{name}=1-\frac{D(U_A,U_B)}{\max(l_{U_A},l_{U_B})} \tag{1}$$

（3）文件格式相似度计算方法

地理空间数据格式主要包括栅格和矢量两类格式。数据格式的相似度取决于两种数据之间的转换难度。数据转换越容易，则两种数据的数据格式相似度越高[8]。根据数据转换

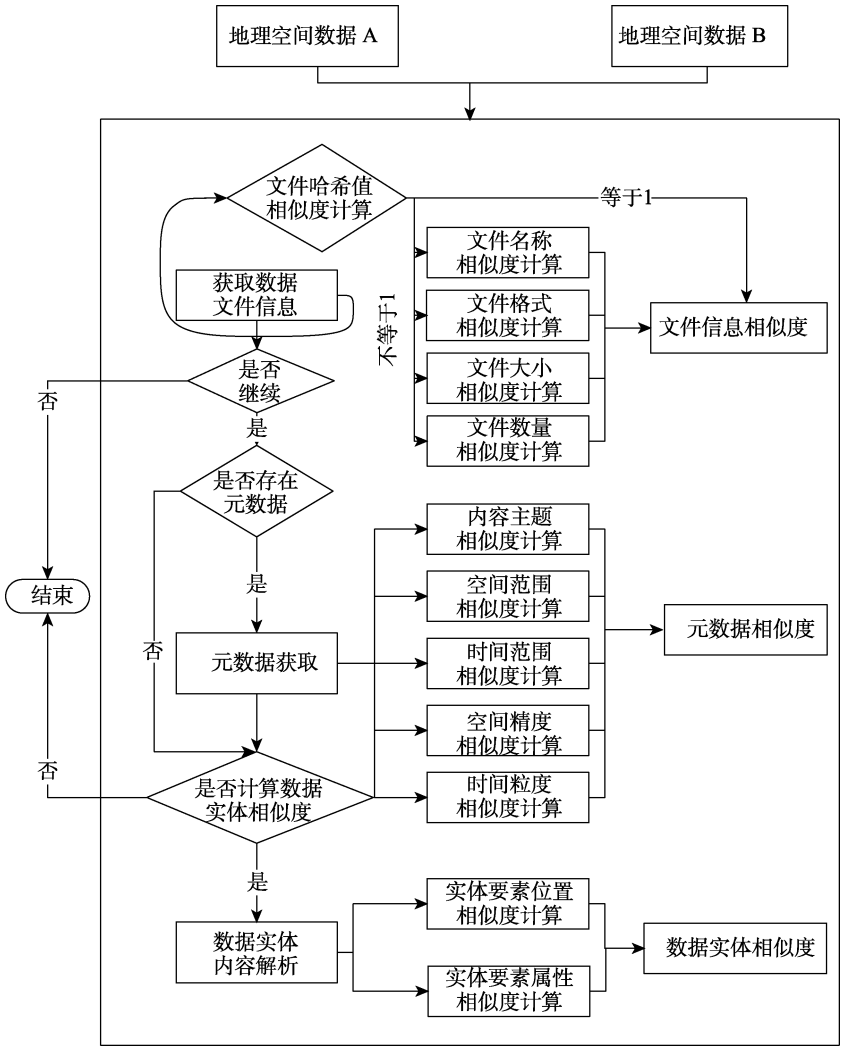

图 1 地理空间数据相似度集成计算方法总体流程

的难易程度，本文将数据格式的相似度计算分为三种情况：相同的数据格式、相同族的数据格式和不同族的数据格式。

若两种数据的数据格式相同，不需要进行数据转换，所以数据格式相似度为 1；相同族的数据格式是指同一企业的软件产品可支持的系列格式（如 ArcGIS 产品支持的数据格式，Shp、E00 等即为同族数据格式），其数据转换通过现有的数据处理工具能够比较容易实现。已有研究，将相同族的数据格式的数据相似度设置为 0.85[8]。

不同家族的数据格式转换难度相对于前述两类格式，则复杂得多，其转换难度由数据格式的开放性决定，具体的相似度计算方法可参考 Zhu *et al*. (2017)[8]。

（4）文件数量相似度计算方法

数据的文件数量是一个大于 0 的整数型数值，计算地理空间数据的文件数量的相似度只需要比较两个地理空间数据的文件数量的大小。因此，地理空间数据的数据数量的相似度用公式(2)计算：

$$S_{num}=1-\frac{|N_A-N_B|}{N_A+N_B} \tag{2}$$

式中，S_{num} 是数据集 A 和数据 B 文件数量的相似度，N_A 和 N_B 分别为数据集 A 和数据 B 的文件数量。

（5）文件大小相似度计算方法

文件大小是数据在计算机中的存储量，是组成地理空间数据所有文件的大小之和。由于不同的压缩方式会改变文件的大小，因此，本文的文件大小是文件未压缩状态的文件大小。

计算地理空间数据的文件大小相似度需要将两个地理空间数据的所有文件的大小都转换为相同单位下的数值，如统一为 KB 或 MB。地理空间数据的存储量大小相似度可用公式(3)计算：

$$S_{size}=\begin{cases}\dfrac{S_A}{S_B}, & S_A\le S_B\\ \dfrac{S_B}{S_A}, & S_A>S_B\end{cases} \tag{3}$$

式中，S_{size} 是数据集 A 和数据 B 文件大小的相似度，S_A 和 S_B 分别为数据集 A 和数据 B 的文件大小。

2.3.2　元数据相似度计算方法

（1）内容主题相似度计算方法

地理空间数据的内容主题特征主要由元数据标题、关键词和摘要组成。内容主题相似度可基于这些元数据项的主题词进行计算。因此，地理空间数据内容主题相似度可用公式(4)计算：

$$S_{cont}=W_{title}\times S_{title}+W_{abs}\times S_{abs}+W_{key}\times S_{key} \tag{4}$$

式中，S_{cont} 为内容主题相似度，S_{title} 为标题相似度，S_{abs} 为摘要相似度，S_{key} 为关键词相似度，W_{title}、W_{abs} W_{key} 分别为标题、摘要和关键词的权重。根据 Zhu *et al*. (2017)[8]，标题、关键词和摘要的权重可分别设置为 0.529、0.309 和 0.162。

标题、关键词和摘要的相似度分别由对应的主题词相似度来衡量。本文主题词的相似度的计算流程如下：（1）通过分词工具将提取到的标题和摘要中连续的文本分割为多个词语；（2）去除对相似度计算无意义的停用词，如中文中的“的”、“得”，英文中的冠词等，然后得到主题词；（3）通过词频-逆向文件频率（term frequency–inverse document frequency，TF-IDF）的特征权重计算方法分别对主题词进行向量化表示；（4）利用余弦相似度计算主题词的相似度。

（2）空间范围相似度计算方法

地理空间数据的空间范围通常由数据集的最小外包矩形（Minimum Bounding Rectangle，MBR）来表示。因此，本文利用最小外包矩形代表地理空间数据空间范围。首先计算数据空间范围拓扑关系，然后进一步计算其空间度量关系，具体计算公式如(5)所示：

$$S_{stp}=W_{sbs}\times S_{sbs}+W_{sds}\times S_{sds} \tag{5}$$

式中，S_{stp}为空间范围相似度，S_{sbs}和S_{sds}分别为空间拓扑关系相似度和空间度量关系相似度，W_{sbs}、W_{sds}为对应的权重，具体相似度计算方法和权重确定方法参考 Zhu *et al*. (2017)[6]。

（3）时间范围相似度计算方法

地理空间数据的时间通常有两种形式：瞬时时间（时间点）和时间段（时间范围）。时间范围通常由开始时间和结束时间两个时间点组成。瞬时时间和时间段是相对的，在不同的时间尺度下可以进行相互转换。

根据前面的分析，地理空间数据时间范围相似度计算会出现三种情况：（1）两个时间都是瞬时时间；（2）一个时间是瞬时时间，另一个是时间段；（3）两个时间都是时间段。对于第两种情况，可以通过时间降尺度方法将瞬时时间转换为时间间隔，并将两个时间间隔统一为最小的时间尺度。例如一个地理空间数据的时间尺度为“年”（2020 年），而另一个地理空间数据的时间尺度为“月”（2020 年 3 月至 2021 年 3 月），则需要将时间尺度为“年”（2020 年）的点时间转换为时间尺度为“月”的时间段（2020 年 1 月至 2020 年 12 月），以保持两个地理空间数据具有统一的时间尺度，从而可以计算他们的时间范围的相似度。所以，两种地理空间数据的时间范围相似度计算的三种情况都可以转换为同一时间尺度下的时间间隔进行计算。

地理空间数据的时间范围相似度可以结合时间拓扑关系、顺序关系和度量关系进行计算[20]。时间拓扑关系是地理现象在时间上的关系。这种拓扑关系表示一种地理空间数据在另一种地理空间数据之间、之后或者同时的时间。对于时间同时（相等或相交）的情况，还考虑两个时间范围的先后顺序。时间度量关系包含时间重叠比例和时间距离两个指标。此外，时间顺序也是计算时间范围相似度需要考虑的重要指标。一般假设新数据比旧数据好。因此，本文的地理空间数据的时间范围相似度用公式(6)计算：

$$S_{tc} = W_{tt} \times S_{tt} + W_{td} \times W_{ts} \times S_{td} \tag{6}$$

式中，S_{tc}为时间范围相似度，S_{tt}、S_{td}分别为时间拓扑相似度和时间距离相似度，W_{tt}、W_{td}分别为时间拓扑关系和时间度量关系的权重，W_{ts}为时间顺序度，其具体相似度计算方法和权重确定方法参考 Chen *et al*. (2018)文献[20]。

（4）空间精度相似度计算方法

地理空间数据的空间精度通常由空间比例尺（矢量）/分辨率（栅格）和空间粒度来表示。地理空间数据的空间精度相似度的用公式(7)计算：

$$S_{spr} = W_{ssc} \times S_{ssc} + W_{sgr} \times S_{sgr} \tag{7}$$

式中，S_{spr}表示空间精度相似性；S_{ssc}和 S_{sgr}分别表示空间比例尺度（分辨率）和空间粒度的相似性；W_{ssc}和W_{sgr}为对应的权重其具体相似度计算方法和权重确定方法参考 Zhu *et al*.（2017）文献[6]。

（5）时间粒度相似度计算方法

地理空间数据的时间粒度一般由数据更新频率来表示。如土地覆盖数据 GlobeLand30 的数据更新间隔为 10 年。时间粒度一般根据转换难易程度来衡量。不同的时间粒度可以通

过时间尺度上推或下推方法进行转换。上推是指将较精细的时间粒度变成较粗糙的时间粒度，使地理实体和现象的表达过程更加概略；下推则相反。当两个地理空间数据的时间粒度为相同、细-粗、粗-细时，时间粒度的相似度分别为 1、0.875 和 0.125[7]。

2.3.3　数据实体相似度方法

（1）实体要素位置相似度计算方法

数据实体相似度是在数据实体要素层面进行的相似度计算，其前提是待计算数据集的数据格式相同，如同为矢量或栅格数据。栅格数据的实体要素位置相似度主要基于两个栅格数据重叠区域进行计算，在计算相似度前先统一两个栅格数据的坐标系和分辨率，然后按公式(8)计算：

$$S_{raster} = S_{ncc} \times S_{\mathrm{cov}} \tag{8}$$

式中，S_{raster} 是栅格数据的实体要素位置相似度，S_{ncc} 和 S_{cov} 分别为两个栅格数据的重叠区域的归一化相关系数和重叠比例。

对于栅格数据重叠区域的归一化系数，用公式(9)计算[21]：

$$S_{ncc} = \frac{\sum_{i}^{N}\sum_{j}^{N}\left(a_{ij}-\overline{a}\right)\left(b_{ij}-\overline{b}\right)}{\sqrt{\sum_{i}^{N}\sum_{j}^{N}\left(a_{ij}-\overline{a}\right)^{2}\sum_{i}^{N}\sum_{j}^{N}\left(b_{ij}-\overline{b}\right)^{2}}} \tag{9}$$

式中，a_{ij} 和 b_{ij} 分别为栅格数据 A 和 B 在行列数分别为 i 和 j 的像元值，$\overline{a}$ 和 $\overline{b}$ 分别为栅格数据 A 和 B 的重叠区域的像元平均值，N 为该区域的像元数。

对于两个栅格数据的重叠比例，用公式(10)计算：

$$S_{\mathrm{cov}} = \frac{Area\left(E_A \cap E_B\right)}{\max(Area(E_A), Area(E_B)} \tag{10}$$

式中，$Area\left(E_A \cap E_B\right)$ 为栅格数据 A 和 B 的重叠面积，$Area(E_A)$ 和 $Area(E_B)$ 分别为栅格数据 A 和 B 的面积。

矢量要素的实体要素位置相似度分为点要素、线要素和面要素三种情况进行计算。

点要素数据的实体要素位置相似度由点群间的拓扑关系、距离关系、方向关系、分布范围和密度确定，用公式(11)计算：

$$S_{poi} = W_{topo} \times S_{topo} + W_{fb} \times S_{fb} + W_{dir} \times S_{dir} + W_{dis} \times S_{dis} + W_{den} \times S_{den} \tag{11}$$

式中，S_{poi} 为点要素数据的实体要素位置相似度，S_{topo}、S_{fb}、S_{dir}、S_{dis} 和 S_{den} 分别为点要素数据的拓扑相似度、分布范围相似度、方向关系相似度、距离关系相似度和分布密度相似度，W_{topo}、W_{fb}、W_{dir}、W_{dis} 和 W_{den} 为其对应的权重，上述点要素的各个相似度具体计算方法和对应权重参考段晓旗的计算方法（2016）[22]。

线要素数据的实体要素位置相似度由线群间的拓扑关系、距离关系、方向关系和几何特征确定，用公式(12)计算：

$$S_{line} = W_{topo} \times S_{topo} + W_{dir} \times S_{dir} + W_{dis} \times S_{dis} + W_G \times S_G \quad (12)$$

式中，S_{line}为线要素数据的实体要素位置相似度，S_{topo}、S_{dir}、S_{dis}和S_G分别为线要素数据的拓扑相似度、方向关系相似度、距离关系相似度和几何特征相似度，W_{topo}、W_{dir}、W_{dis}和W_G为其对应的权重，上述线要素的各个相似度具体计算方法和对应权重参考文献[12]。

面要素数据的实体要素位置相似度计算方法与线要素一样，面要素的各个相似度具体计算方法和对应权重参考刘涛的计算方法（2013）[23]。

（2） 实体要素属性相似度计算方法

地理空间数据的属性通常由图层属性表中的属性项名称和属性值来表示，因此属性相似度通常由属性项名称和属性值来确定，具体计算方法如公式(13)所示：

$$S_{att} = W_{item} \times S_{item} + W_{value} \times S_{value} \quad (13)$$

式中，S_{item}和S_{value}分别为属性项名称和属性值的相似度，W_{item}和W_{value}是其对应的权重。根据谭永滨等（2017）[24]，W_{item}和W_{value}的值分别为0.4和0.6。属性项名称和属性值的相似度都通过编辑距离算法进行整体度量。

2.3.4 相似度聚合方法与应用策略

（1）相似度聚合方法

基于上述单一特征或元数据项的相似度，利用层次分析法可以计算出各层次的复合相似度（文件相似度、元数据相似度和数据实体相似度），具体的方法如公式(14)所示。

$$S = \sum_{i=1}^{n} (W_i \times S_i) \quad (14)$$

式中，S是各层次的复合相似度，S_i和W_i是第i个单一特征相似度和对应的权重，n代表单一特征相似度的个数，W_i利用层次分析法计算得出（表 1）。当某个单一相似度缺失时，该层次其他的相似度按照局部权重的相似大小重新分配局部权重，用于该层次的复合相似度。

表 1 各层次复合相似度计算的权重表

复合相似度	单一相似度	局部权重
文件信息相似度	文件名称相似度	0.500
	文件格式相似度	0.200
	文件大小相似度	0.200
	文件数量相似度	0.100
元数据相似度	内容主题相似度	0.550
	空间范围相似度	0.100
	时间范围相似度	0.150
	空间精度相似度	0.100
	时间粒度相似度	0.100
数据实体相似度	实体要素位置相似度	0.625
	实体要素属性相似度	0.375

（2）应用策略

首先通过文件信息相似度对两个数据的相似度进行粗算，然后计算元数据相似度，如果元数据缺失或想进一步得到更加精确的相似度，可进一步计算数据实体相似度，以此形成满足应用需求的不同层次的数据相似度计算，实现地理空间数据相似度计算既保证精度又提高效率的目标。

3　系统软件

3.1　地理空间数据相似度计算软件

为了便于用户理解和应用地理空间数据相似度的计算，本文基于 Python 语言开发了地理空间数据相似度计算软件（Geospatial Data Similarity Calculation Software），简称 GDSCS V1.0[25]。由于地理空间数据的复杂性和多样性，本软件仅以国家青藏高原科学数据中心[1]作为数据源进行了计算实现。

3.2　GDSCS 软件功能与测试结果

GDSCS 是基于地理空间数据相似度计算方法开发的集数据特征解析、相似度计算和结果可视化于一体的地理空间相似度计算软件。该软件功能包括：地理空间数据输入、数据特征信息提取、数据相似度计算、计算结果可视化和导出等功能。从国家青藏高原科学数据中心随机选择的“川藏交通廊道植被覆盖度（1985–2020）[26]”（以下简称数据 A）和“‘一带一路’沿线国家植被覆盖状况恢复力数据集（2000–2020）[27]”（以下简称数据 B）为实验数据，测试 GDSCS 系统的运行和结果。数据 A 和数据 B 的文件信息如表 2 所示。

利用 GDSCS 软件，生成文件测试、元数据测试和实体数据测试结果如图 2、图 3 和图 4 所示。需要说明的是元数据输入格式为 json，实体数据输入格式为矢量.shp、栅格.tif。

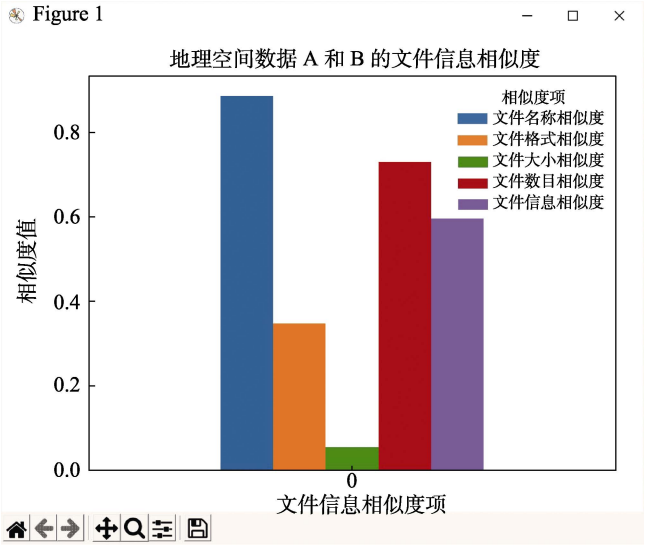

图 2　文件信息相似度计算结果对比图

[1] https://data.tpdc.ac.cn/.

表 2 数据 A 和数据 B 的文件信息表

文件信息项	数据 A	数据 B
文件名称	CZLD_VFC_1000m_2016-2020	vegetation_country
文件格式	.tif	.shp
文件大小	37.6 MB	616 MB
文件数量	8	10

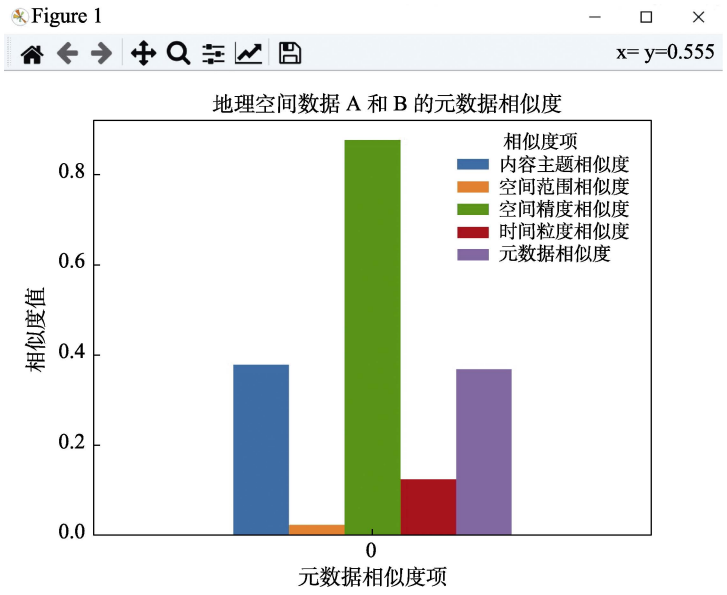

图 3 元数据相似度计算结果对比图

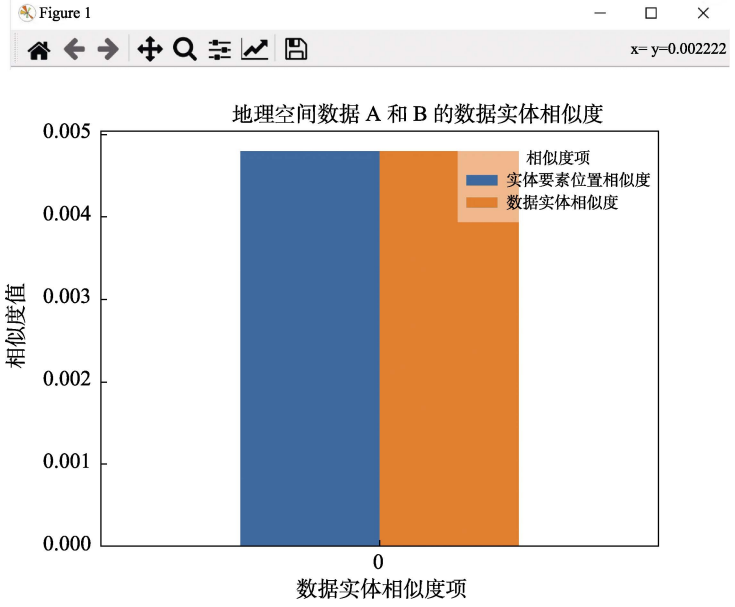

图 4 数据实体相似度计算结果对比图

4 讨论和总结

地理空间数据相似度计算与应用，对地理空间数据共享与智能发现具有重要的价值和意义。本文提出的 GDSCS 方法为科学数据的相似度计算和数据集查重提供可操作的工具，并为进一步完善奠定了基础。GDSCSV1 无论在理论方法还是在工具实现方面还都有进一步完善的方面。例如：在文件信息相似度计算中，文件信息的关联程度、元数据格式的多样性、实体数据集的数据格式多样性、空间数据内容关联性、空间位置系统性偏移现象的相似度计算特殊性等都需要进一步改进和完善。

作者分工：诸云强负责方法的总体设计与论文修改；代小亮进行了方法实践与写作；杨杰进行了方法的研究；孙凯负责方法的研究；李吉东参与了方法的设计和论文修改；宋佳对论文进行了修改。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

[1] Zhu, Y., Yang, J. Automatic data matching for geospatial models: a new paradigm for geospatial data and models sharing [J]. *Annals of GIS*, 2019, 25(4): 283–298.

[2] Chen, Z., Yang, Y. Semantic relatedness algorithm for keyword sets of geographic metadata [J]. *Cartography and Geographic Information Science*, 2020, 47(2): 125–140.

[3] Chen, Z., Song, J., Yang, Y. An approach to measuring semantic relatedness of geographic terminologies using a thesaurus and lexical database sources [J]. *ISPRS International Journal of Geo-Information*, 2018, 7(3): 98.

[4] 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.

[5] Guo, H., Liu, Z., Jiang, H., *et al.* Big earth data: a new challenge and opportunity for digital earth's development [J]. *International Journal of Digital Earth*, 2017, 10(1): 1–12.

[6] Zhu, Y., Zhu, A. X., Song, J., *et al.* Multidimensional and quantitative interlinking approach for linked geospatial data [J]. *International Journal of Digital Earth*, 2017, 10(9): 923–943.

[7] Boubenia, M., Belkhir, A., Bouyakoub, F. M. Combining linked open data similarity and relatedness for cross OSN recommendation [J]. *International Journal on Semantic Web and Information Systems* (*IJSWIS*), 2020, 16(2): 59–90.

[8] Zhu, Y., Zhu, A. X., Feng, M., *et al.* A similarity-based automatic data recommendation approach for geographic models [J]. *International Journal of Geographical Information Science*, 2017, 31(7): 1403–1424.

[9] 孙娟娟, 禹继国, 刘祥涛. 面向 P2P 文件共享应用的相似度计算模型[J]. 计算机工程与应用, 2012, 48(4): 111–114.

[10] Kim, T. G., Lee, Y. R., Kang, B. J., *et al.* Binary executable file similarity calculation using function matching [J]. *The Journal of Supercomputing*, 2019, 75(2): 607–622.

[11] Kim, B. K., Oh, S. J., Jang, S. B., *et al.* File similarity evaluation scheme for multimedia data using partial hash information [J]. *Multimedia Tools and Applications*, 2017, 76(19): 19649–19663.

[12] 刘涛, 杜清运, 毛海辰. 空间线群目标相似度计算模型研究[J]. 武汉大学学报·信息科学版, 2012, 37(8): 992–995.

[13] 朱爽. 用直方图面积法进行图像相似度计算[J]. 测绘通报, 2018(12): 96–100.

[14] Karschnick, O., Kruse, F. A, Töpker, S., *et al.* The UDK and ISO 19115 standard [C]. EnviroInfo, 2003: 475–481.

[15] Authority, T. V. Content standard for digital geospatial metadata [D]. National Aeronautics and Space Administration, 1998.

[16] Jiang, J., Liu, R. China geographic information–metadata GB/T 19710—2005 [S]. World Spatial Metadata Standards: Elsevier Science, 2005.

[17] Rachmawati, D., Tarigan, J. T., Ginting, A. B. C. A comparative study of Message Digest 5 (MD5) and SHA256 algorithm [J]. *Journal of Physics*: *Conference Series*, 2018, 978(1): 012116.

[18] Strube, M., Rapp, S., Müller, C. The influence of minimum edit distance on reference resolution [C]. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), 2002: 312–319.

[19] 赵作鹏, 尹志民, 王潜平等. 一种改进的编辑距离算法及其在数据处理中的应用[J]. 计算机应用, 2009(2): 424–426.

[20] Chen, Z., Song, J., Yang, Y. Similarity measurement of metadata of geospatial data: an artificial neural network approach [J]. *ISPRS International Journal of Geo-Information*, 2018, 7(3): 90.

[21] Rao, Y. R., Prathapani, N., Nagabhooshanam, E. Application of normalized cross correlation to image registration [J]. *International Journal of Research in Engineering and Technology*, 2014, 3(5): 12–16.

[22] 段晓旗, 刘涛, 武丹. 基于层次分析法的多尺度点群目标相似度计算[J]. 地球信息科学学报, 2016, 18(10): 1312–1321.

[23] 刘涛, 闫浩文. 空间面群目标几何相似度计算模型[J]. 地球信息科学学报, 2013, 15(5): 635–642.

[24] 谭永滨, 唐瑶, 李小龙等. 语义支持的地理要素属性相似性计算模型[J]. 遥感信息, 2017, 32(1): 126–133.

[25] 代小亮，诸云强，杨杰等. 地理空间数据相似度计算软件 V1.0 [J/DB/OL]. 全球变化数据仓储电子杂志, 2022. https://doi.org/10.3974/geodb.2022.10.02.V1. https://cstr.escience.org.cn/CSTR:20146.11.2022.10.02.V1.

[26] 眭天波.川藏交通廊道植被覆盖度(1985–2020)[OL]. 国家青藏高原科学数据中心, 2021. https://doi.org/10.11888/Soil.tpdc.271618.

[27] 徐新良. “一带一路”沿线国家植被覆盖状况恢复力数据集(2000–2020)[OL]. 国家青藏高原科学数据中心, 2022. https://doi.org/10.11888/HumanNat.tpdc.272282.