引文格式：符浩军,蒋秉川,陈星宇. 利用多因子融合的多尺度点状地名匹配方法[J]. 地理空间信息, 2024, 22(12): 34-37. doi:10. 3969/j.issn. 1672-4623.2024.12.008

利用多因子融合的多尺度点状地名匹配方法

符浩军[1,2]，蒋秉川[3]，陈星宇[1]
（1. 西安测绘总站，陕西 西安 710054；2. 智慧地球重点实验室，北京 100094；
3. 战略支援部队信息工程大学，河南 郑州 450001）

摘　要：同名实体匹配技术对多尺度环境下地名数据的整合、级联更新等极其重要。首先，对现有尺度下地名库数据的相互关系进行了分析，在此基础上，综合利用点实体地名数据的语义信息和空间信息，选用名称表达、行政隶属关系、类别编码体系和空间关系等地名的稳定特质作为匹配因子，构建了名称表达相似度、类别编码相似度、空间关系相似度的计算方法；然后通过在各匹配因子之间建立相互融合的机制，设计了多尺度环境下基于多因子融合的点实体地名数据匹配方法，实验结果表明，相较于传统的匹配方法，所提出的方法降低了误匹配、漏匹配的概率，且方法具有一定的数据纠错能力，有效提升了点实体地名数据匹配的性能。

关键词：地名数据；点实体；多尺度；多匹配因子；融合机制
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2024）12-0034-04

Multi-scale Point Toponymy Matching Method Based on Multi-factor Fusion

FU Haojun[1,2], JIANG Bingchuan[3], CHEN Xingyu[1]
(1. Xi'an Information Technique Institute of Surveying and Mapping, Xi'an 710054, China;
2. Key Laboratory of Smart Earth, Beijing 100094, China
3. Strategic Support Force Information Engineering University, Zhengzhou 450001, China)

Abstract: The same name entity matching technology is very important to the integration and propagating update of multi-scale toponymy data. Firstly, we analyzed the interrelation of existing toponymy databases in series scale. And then, comprehensive using semantic information and geo-spatial information, taking name expression, administrative subordination, category coding system and spatial relation as matching factors, we constructed the calculation methods of name expression similarity, category coding similarity and spatial relation similarity. Finally, we designed a matching method of point entity toponymy data based on multi-factor fusion in multi-scale environment by establishing the fusion mechanism among the matching factors. The experimental results show that compared with the traditional matching method, the proposed method can reduce the probability of mismatching and missing matching, and have a certain ability of data error correction, effectively improving the performance of point entity toponymy data matching.

Key words: toponymy data, point entity, multi-scale, multi-matching operator, fusion mechanism

同名实体匹配技术是当前地理空间信息融合领域的重点和难点，目前主流的思路包括：①依据属性信息匹配[1-2]；②依据空间信息匹配[3-4]；③依据拓扑信息匹配[5-8]。由于地名所指代的地理实体包括点、线、面等类型，不同的数据情况，需设计不同的匹配策略和方法，本文针对多尺度环境下实体类型为点状的地名数据匹配问题进行研究，以多尺度系列地名库中的点实体地名数据为研究对象，综合利用语义信息和空间信息，通过构建多匹配因子的融合机制，建立一种点实体地名数据匹配方法。

1　匹配因子选择

系列比例尺地名数据库是将各尺度下的基础地理框架数据中各类注记名包括居民地、河流、湖泊、山脉、山峰、岛屿、沙漠、盆地、自然保护区等的名称，同其属性特征如类型、归属、定位点经度、定位点纬度等录入而建成的数据库。由于其来源于基础地理框架数据，因此在语义组成、编码规范、时空参考等方面具有高度一致性。针对系列比例尺地名库中的点实体地名数据而

收稿日期：2023-06-28。
项目来源：国家自然科学基金资助项目（42171456）；国防科技基础加强计划资助项目（2021-JCJQ-JJ-0507）。
第一作者简介：符浩军（1982—），博士，高级工程师，研究方向为多源地名数据获取与整合、 地理数据共享与安全等，E-mail：fhjun121@163.com。

言，从制图知识及数据处理经验可知，其是否为同一地名主要由名称、所属要素层、空间距离等因素决定。

因此，结合现有属性信息匹配算法和空间信息匹配算法的研究思路、制图知识和多源地名数据融合处理经验，兼顾地名库数据结构特点，本文综合利用点实体地名的语义信息和空间信息，在语义信息方面选用名称表达、行政隶属关系、类别编码体系，空间信息方面选用空间关系，作为匹配方法建立的匹配因子。

1.1　名称表达

对于实体类型为点状的地名数据而言，基于语义信息的匹配至关重要，而名称信息则最为直观有效。然而汉语博大精深，汉语中的常用词就有43 570个，组成词的汉字更是数量繁多[9]，因此地名数据库中的名称表达存在大量的描述不准确、错别字、同音字等情况，导致多尺度环境下同一地理实体的名称表达不一致。如果直接使用名称信息进行匹配，则极可能造成漏匹配情况的出现，使得匹配成功率降低。考虑到汉字拼音只有496个，因此可将名称信息对应的拼音作为多尺度环境下点实体地名数据的匹配因子，相对来说更为稳定、有效。通过上述方式，使得匹配方法具有一定的纠错性能，从而有效降低漏匹配概率。

1.2　行政隶属关系

由于地名表达的多样性，导致地名歧义问题普遍存在[10]，即同一名称对应了多个地名数据，且各地名数据所指代的地理实体都不尽相同。如在1∶25万地名数据库的居民地要素层，有多达45条名称信息为“黄家村”的点实体地名数据，且这些数据基本隶属不同的行政区划。

此类现象可归属于地名歧义问题，此问题将造成地名误匹配情况的出现，使得匹配成功率降低。基于此考虑，可将地名所指代的点实体与政区面数据进行套合，对地名数据建立垂直的行政隶属关系树，并以此作为多尺度环境下点实体地名数据的匹配因子，能有效消除此类地名歧义问题，提高点实体地名数据匹配的效率和准确度。如图1所示，为筛选的部分“黄家村”地名数据对应的行政隶属关系树结构。

1.3　类别编码体系

地名数据对地理实体的描述是建立在一定的语义体系基础之上，一般而言以地名所指代地理实体的类别编码进行表达，表示地理实体的类型、类别、等级等信息。如图2所示，镇代码2152的门、大、中、小类的含义分别为人文地理实体、行政区域及其他区域、四级行政区域、城市型四级行政区域。实际上

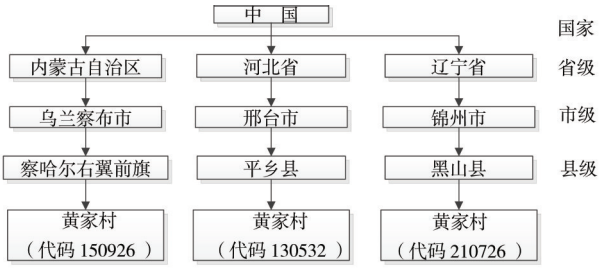


图1　地名对应的行政隶属关系树结构

由于类别编码的不同，尽管诸多地名数据在名称表达、空间位置等方面具有高度相似性，但其所指代的地理实体区别很大，如果忽视类别编码这一因素，很容易造成类别不同的地名数据误匹配情况的出现。

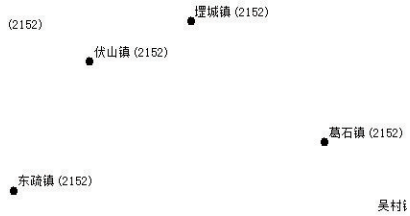


图2　地名所指代地理实体的类别编码

1.4　空间关系

将类别编码作为多尺度环境下点实体地名数据的匹配因子，通过对应层级的类别编码匹配，如大类码的匹配、中类码的匹配、小类码的匹配，以此考察参与匹配的地名数据类别编码体系是否一致，可有效降低点实体地名数据误匹配的概率。

基于空间关系的同名实体匹配是最常见也是最有效的方法，如位置最近的方法、位置相互最近的方法等。对于系列比例尺地名库而言，由于各尺度地名数据之间没有建立纵向关联的关系，且各尺度地名数据所指代的地理实体表达存在误差、不一致等因素，因此单纯采用基于空间关系的匹配方法进行判别处理，很容易造成数据误匹配的情况。基于此考虑，结合点实体地名数据的空间特性，一方面需要设定有效的空间距离阈值，将阈值范围内的点实体地名数据作为候选匹配集，有效提升数据匹配处理的效率，降低误匹配的概率；另一方面还需结合其他匹配因子，建立多因子融合的匹配方法，有效提升点实体地名数据匹配处理的准确度。

2　匹配方法

2.1　方法流程

以多尺度系列地名数据库中的点实体地名数据为对象，以名称表达、行政隶属关系、类别编码体系和空间关系作为匹配因子，构建基于多因子融合的地名数据匹配方法。这里以1∶100万、1∶25万2个尺度

下的点实体地名数据匹配为示例场景，对方法流程进行详细阐述：

1） 获取参与匹配的点实体地名数据，这里为1∶100万地名库和1∶25万地名库中的点实体数据，并以1∶100万点实体地名数据为参考集，1∶25万点实体地名数据为匹配集。一般来说，匹配过程是依据一个数据集与另一个数据集进行匹配处理，依据的数据集即是参考集，参与匹配的数据集即为匹配集，在多尺度环境下，通常原则是将比例尺小的数据作为参考，这样能有效提升数据匹配处理的效率。

2）将参考集中的数据与匹配集中的每条数据进行循环匹配处理，计算每次匹配的相似度，取相似度最大值与设定阈值进行比较：如果最大值大于阈值，则最大值对应的两条地名数据为有效的匹配地名；反之匹配失败。

3）对参考集中的每条地名数据执行上述步骤，直至参考集中的所有地名数据完成匹配处理为止。

2.2 相似度融合机制

点实体地名数据匹配的相似度计算采用多匹配因子融合的方法，由三部分组成：名称表达相似度、类别编码体系相似度、空间关系相似度。其中名称表达相似度是由名称表达和行政隶属关系2个匹配因子组合进行计算，这样设计的原因是点实体地名的名称表达在相同的三级行政区划内地名歧义问题很少，可有效降低因地名歧义问题所引发的地名误匹配概率，类别编码体系相似度、空间关系相似度分别由类别编码体系因子、空间关系因子决定。具体融合机制及方法如下：

1）名称表达相似度：获取参与匹配的2条地名数据对应的行政隶属关系树（此2条数据分属参考集和匹配集），将行政隶属关系树对应的代码与名称对应的拼音组合成匹配因子，采用莱文史特距离算法进行相似度计算[11]。

$$S_N = 1 - \frac{\text{distance}(str1, str2)}{\text{Max}(\text{Len}(str1), \text{Len}(str2))} \tag{1}$$

式中，S_N 为名称表达相似度；str1、str2分别为参与匹配的2个地名数据对应的匹配因子；Max(Len(str1), Len(str2)) 为2个匹配因子的最大长度；distance(str1 , str2) 为示2个匹配因子的编辑距离。

2） 类别编码体系相似度：获取参与匹配的2条地名数据对应的类别编码，并以此为匹配因子，由于类别编码详细地规定了地名所指代地理实体归属的门类、大类、中类、小类，且各类重要性依次递减。因此，可计算门类、大类、中类、小类各自的相似度，并依据重要性程度赋以不同的权重，最后组合得到对应类别编码体系的相似度。具体如下式（2）所示。

$$S_C = \begin{cases} 0; & if(\text{top}_{c1} \neq \text{top}_{c2}) \\ 1 \times w_1; & if(\text{big}_{c1} \neq \text{big}_{c2}) \\ 1 \times w_1 + 1 \times w_2; & if(\text{mid}_{c1} \neq \text{mid}_{c2}) \\ 1 \times w_1 + 1 \times w_2 + 1 \times w_3; & if(\text{sma}_{c1} \neq \text{sma}_{c2}) \\ 1 \times w_1 + 1 \times w_2 + 1 \times w_3 + 1 \times w_4; & if(c1 = c2) \end{cases} \tag{2}$$

式中，S_C 为类别编码体系相似度；$c1$、$c2$ 为参与匹配的2条地名数据对应的类别编码；w_1、w_2、w_3、w_4 为门类 top_c、大类 big_c、中类 mid_c、小类 sma_c 对应的权重，权重值满足 $w_1 + w_2 + w_3 + w_4 = 1$。

3）空间关系相似度：由于不同尺度下点实体地名数据的点位精度并不一致，依据误差传播定律，同名点实体在不同尺度下的距离精度可定义为：

$$M = \sqrt{M_1^2 + M_2^2} \tag{3}$$

式中，M_1、M_2 为参考集、匹配集在对应尺度下的点位精度，由比例尺和最小误差共同确定；M 为不同尺度下同名点实体的距离精度。这里可将不同尺度下同名点实体的距离精度扩展为 $3M$，以进一步降低漏匹配的概率。在此基础上，依据参与匹配的2条地名数据对应的点实体信息，采用欧氏距离计算其空间关系相似度，具体如下式（4）所示。

$$S_G = \begin{cases} 0; & if(\text{dis}(p1, p2) > 3M) \\ 1 - \dfrac{\text{dis}(p1, p2)}{3M}; & if(\text{dis}(p1, p2) \leq 3M) \end{cases} \tag{4}$$

式中，S_G 为空间关系相似度；$p1$、$p2$ 为参与匹配的2条地名数据；$\text{dis}(p1, p2)$ 为2条地名数据对应点实体的欧氏距离。

4） 多因子融合的相似度：对各匹配因子计算的相似度进行加权处理，得到参与匹配的2条点实体地名数据最终的相似度，具体如下式（5）所示。

$$S = \alpha \times S_N + \beta \times S_C + \gamma \times S_G \tag{5}$$

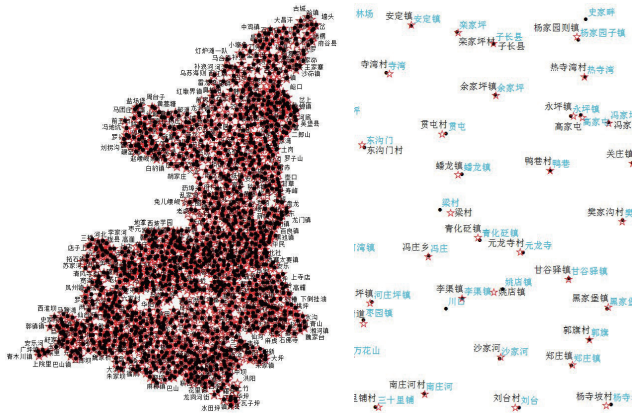
其中权重满足 $\alpha + \beta + \gamma = 1$，至此得到此次匹配最终的融合相似度 S。

3 实验及分析

3.1 实验算例

依据所设计的方法，以vs2012为工具，实现了方法原型系统。以1∶100万、1∶25万2个尺度下的地名库中某区域部分点实体数据为实验对象，对多尺度环境下基于多因子融合的点实体地名数据匹配方法进行实验与验证，实验数据中1∶100万点实体地名数据

1 601条、1∶25万点实体地名数据26 602条。针对权重的设定，对于类别编码而言，门类、大类、中类、小类对地名匹配的影响程度依次递减，在此基础上经实验 w_1 、 w_2 、 w_3 、 w_4 权重分别取值0.35、0.30、0.20、0.15为佳；在相似度融合机制中，名称表达相似度最为重要，类别编码体系相似度、空间关系相似度相对次之，经试验 α 、 β 、 γ 权重分别取值0.4、0.30、0.3为佳，综合相似度 S 的阈值设置为0.81。如图3所示为采用本文所提方法得到的匹配结果总体示意图，标注红色五角星为1∶100万地名数据，标注黑色圆点为匹配成功的1∶25万地名数据，标注为绿色圆点为匹配失败的1∶25万地名数据。



a 匹配结果总体示意图　　b 匹配结果局部示意图

图3　匹配结果总体/局部示意图

对匹配结果进行统计，1∶100万地名数据累计完成匹配1 446条，匹配失败155条。为进一步验证算法的实用性，采用本文提出的方法在相同实验区域，继续以1∶25万、1∶5万2个尺度的点实体地名数据为实验对象开展验证，其中1∶25万点实体地名数据为26 602条、1∶5万点实体地名数据为86 913条。通过本文方法进行数据匹配后对结果统计，1∶25万点实体地名数据累计完成匹配22 751条，匹配失败3 851条。在工程化实际中，人工综合判别匹配的效率约为800条/d，而采用本文方法，匹配成功率较好，且仅需对匹配失败的地名数据进一步人工综合判别，极大地减少了工作量，提高了数据融合处理效率。

对实验结果进行分析，匹配失败的主要原因如下：其一是不同尺度下同名实体类型的不一致造成的，由于本文仅考虑不同尺度下实体类型均为点状的地名数据，对于由于尺度变化所造成的地名实体类型变化这一情况，不在本文研究范围；其二从制图综合的角度及常规认知原则上理解，在小尺度中存在的数据，其大尺度中一定存在，然而实际数据建设并非如此，由于相应尺度下的地名数据并没有建立上述严谨的纵向关联关系，直接导致了相应尺度下匹配数据的不存在，这也是造成数据匹配失败的主要因素。

3.2　方法比较

在相同实验场景下，将本文方法与传统的匹配方法相比较，这里选用基于空间信息的匹配方法、基于属性信息的匹配方法，其中基于空间信息的匹配方法以地名实体的空间距离作为匹配指标，采用相互最近的方式并加以阈值控制进行匹配处理；基于属性信息的匹配方法选用名称、行政区划、类别编码作为匹配指标，采用逐层匹配筛选的方式进行匹配处理。如表1所示为匹配结果。

表1　与传统方法横向比较

实验方法	匹配地名实际对数	算法匹配地名对数	误匹配对数	漏匹配对数	有效匹配对数
本文方法	1 443	1 446	3	0	1 443
基于空间信息的匹配方法	1 443	1 585	223	81	1 362
基于属性信息的匹配方法	1 443	1 104	30	369	1 074

由比较结果可知，基于属性信息的方法漏匹配概率较高，究其原因有二：一方面是数据本身不规范所造成的，现实中地名库数据的属性信息表达普遍存在缺失行政后缀、同音字、名称不一致等情况，而匹配方法又没有对数据自动纠错的能力；另一方面是匹配方法本身固有的缺陷，由于采用了逐层匹配筛选的方法流程，造成了大量的漏匹配情况出现。而基于空间信息的方法仅仅从空间相互关系的角度进行匹配处理，对数据匹配结果正确与否缺乏可控手段，造成了大量的误匹配情况出现。相较于此2种方法，本文提出的方法采用名称表达、行政隶属关系、类别编码体系和空间关系作为匹配因子，通过建立各因子相互融合的机制，有效控制了误匹配、漏匹配概率，且方法对数据具有较好的纠错能力。从匹配结果可以看出，本文方法的匹配正确率要优于传统方法。

4　结　语

多尺度环境下的地名数据整合、级联更新过程中，同名实体的识别与匹配是关键一环。传统的方法一般采用单个匹配因子判别或多个匹配因子逐层判别，很容易造成误匹配、漏匹配的情况。本文从语义信息和空间信息着手，通过在各匹配因子之间建立相互融合的机制，提出了基于多因子融合的点实体地名数据匹配方法，实验证明了所提出方法的有效性和实

（下转第91页）