引文格式:姜宇荣,高苏,蔡忠亮,等.考虑重合字符位次差异的地名相似性度量方法[J].武汉大学学报(信息科学版),2025,50(7):1425-1434.DOI:10.13203/j.whugis20230066

Citation:JIANG Yurong,GAO Su,CAI Zhongliang,et al.A Toponym Similarity Measure Method Based on Positional Differences of Overlapping Characters[J].Geomatics and Information Science of Wuhan University,2025,50(7):1425-1434.DOI:10.13203/j.whugis20230066

考虑重合字符位次差异的地名相似性度量方法

姜宇荣[1] 高 苏[2] 蔡忠亮[1] 王 巧[3]

1 武汉大学资源与环境科学学院,湖北 武汉,430079

2 云南省地图院,云南 昆明,650034

3 长江水利委员会网络与信息中心,湖北 武汉,430010

摘 要:当前中文地名匹配常见的相似性度量或是只考虑重合字符的数量未考虑位次性,或是体现位次性但未考虑重合字符。通过考虑两个中文地名字符串的重合字符对应关系及其位次差距,构建了一种新的距离度量和相似性度量,能够结合重合字符和位次性两个因素计算两个地名的偏距和偏距相似度。针对重合字符复现的情形确定偏距最小原则,设计全体顺次匹配方案;针对字符片段偏移的情形调整距离度量,从而更符合两个地名相似性的直观认知。该距离度量满足正定性和对称性,但不满足三角不等式。与Jaccard系数和编辑距离相似度的测试对比结果表明,所提偏距算法对相似度刻画更为细致,能够检测到重合字符位次差异但更重视未重合字符的差异;在地名匹配实验中匹配正确率和运行时间分别为63.64%和2 940.56 s,两项指标均优于Jaccard系数和编辑距离相似度。

关键词:相似性度量;近似字符串匹配;自然语言处理;地名匹配;中文地名

中图分类号:P208 **文献标识码**:A **收稿日期**:2023-05-20

DOI:10.13203/j.whugis20230066 **文章编号**:1671-8860(2025)07-1425-10

A Toponym Similarity Measure Method Based on Positional Differences of Overlapping Characters

JIANG Yurong[1] *GAO Su*[2] *CAI Zhongliang*[1] *WANG Qiao*[3]

1 School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China

2 Yunnan Provincial Mapping Institute, Kunming 650034, China

3 Network and Information Center, Changjiang Water Resources Commission, Wuhan 430010, China

Abstract: **Objectives**: Due to the existence of toponym aliases or negligence in data production, the same geographic entity possibly has different names in maps of different scales which may contain spelling errors or not match the standard toponym. It brings many inconveniences to multi-scale map visualization and geospatial entity extraction. Therefore, it is necessary to study toponym similarity measure to achieve toponym matching. However, the commonly used similarity measures for Chinese toponym matching currently either only consider the number of overlapping characters without considering character position, or only reflect positional feature. Therefore, our objective is to construct a new similarity measure that can simultaneously reflect both features. **Methods**: By calculating the positional differences between overlapping characters in two Chinese toponyms, we define a total matching offset representing the degree of positional difference of the overlapping character set. Considering that the impact of non-overlapping characters on the similarity of two toponyms should be greater than that of overlapping characters, we define a total non-matching offset. Then, we define the total offset and offset similarity. For the complex case of overlapping character repetition, we determine the minimum offset principle and design an entire sequential matching scheme. For the

基金项目:国家重点研发计划(2021YFB2501101)。

第一作者:姜宇荣,硕士,主要研究方向为地理信息系统。juineo@whu.edu.cn

通信作者:蔡忠亮,博士,教授。zlcai@whu.edu.cn

complex case of character fragment offset, the sum of offsets for each character in the fragment is replaced by the overall offset of the fragment, making the offset value more reasonable. The total offset satisfies positive definiteness and symmetry, but does not satisfy the triangle inequality, and it is more appropriate to use offset similarity to express toponym similarity. **Results:** Compared with Jaccard similarity and Levenshtein similarity, the results show that the offset similarity can characterize toponym similarity more finely. It attaches greater importance to character differences, and the similarity decreases significantly when the proportion of overlapping characters decreases. It attaches less importance to pure positional differences, but they can also be reflected in slight differences in similarity values. In the toponym matching experiment, the matching accuracy and running time are 63.64% and 2 940.56 s, both of which are better than the Jaccard similarity and Levenshtein similarity. **Conclusions:** The offset similarity has significant advantages in the toponym matching scenario. But like other string similarity measures, it is difficult to understand semantics. Further optimization can be explored in terms of algorithmic solutions for handling complex cases, applicability to other languages, and consideration of the semantic structure of toponyms.

Key words: similarity measure; approximate string matching; natural language processing; toponym matching; Chinese toponym

在中国测绘地理信息技术得到不断发展和广泛应用的背景下,标准地名数据库和多种尺度下的地图要素数据库的数据量大幅上升,显现出巨大的潜在应用价值。但是,由于“一地多名”[1]的现象或数据生产时的疏忽,会出现同一地理实体[2]在不同尺度地图要素数据库中的名称互不相同,出现错别字,或与标准地名不符等问题,给地图的多尺度展示和地理信息数据实体化[3]带来诸多不便。因而有必要研究地名相似性度量方法以实现地名匹配[4],辅助同一地理实体在不同尺度下名称的统一和地理实体库的构建,促进地理信息的精细化管理和应用。

当前有关中文地名相似匹配研究中,采用的文本相似性度量[5-6]多为Jaccard相似系数(又称Jaccard系数)[7-8]、编辑距离[9]和词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)等基于词频向量[10]算法,但是以上相似度算法在实践中均存在各自的缺陷。如果将两地名都包含的字符称为重合字符,那么Jaccard系数定义为重合字符数和总字符数之比。Jaccard系数能够在一定程度上表达地名相似度,但是由于该指标仅考虑了重合字符数而没有考虑位次特征,因此它对相似性的描述是较为粗略的。编辑距离一般指莱文斯坦距离,定义为一个地名字符串变换为另一个地名字符串要经过的最少“操作”次数,其中“操作”包括插入、删除或替换单个字符[11]。综合考量两个地名字符串编辑距离的实际数值和理论最大值可以进一步得到编辑距离相似度。编辑距离算法能够捕捉到两个地名字符串在字符位次排列特征上的相似性,在实际生产中应用也较为广泛。然而该指标容易忽视重合字符特征,在字符片段顺序颠倒等情况中可能对相似性作出不合理的描述[12-13]。另外,也有学者使用TF-IDF等基于词频向量算法[14-15]计算两地名相似度。但是究其原理,TF-IDF算法评估的是某个词语对于语料库中一份文档的“重要程度”,这与“文本相似度”的概念并不能等同。在用于地名地址匹配时,TF-IDF算法会将地名视为诸多关键词的简单集合从而忽略语序,且词频越高,评分越高的逻辑也并不适用于评估地名相似性。

综上,现有的地名相似性度量或是强调重合字符数量,或是强调字符位次特征,却不能将两者很好地结合起来。通过观察可发现,两个地名在字符层面的相似程度关键在于重合字符的匹配情况。因此,本文意在构建能够同时体现重合字符数量及位次特征的新相似性度量指标。由于Jaccard系数和编辑距离相似度同为考虑字符层面的文本相似性度量,在识别字符重合数量和字符位次特征上具有各自的优势且实际中运用广泛[16],故本文将重点与它们进行对比。

1 构建地名相似性度量

1.1 重合字符偏距

考虑两个地名字符串A和B,找到两者所有的重合字符,按在地名A中从左到右的顺序依次记为1、2、3、…、n。对于任一重合字符i,假定它在A和B中均刚好只出现一次。记i在A中的位次是a_i(即i是地名字符串A从左往右数第a_i个字

符);i在B中的位次是b_i,如图1所示,显然每对重合字符恰好构成匹配(一一对应),并且两者的位次越接近(即a_i和b_i差值越小),意味着两个地名字符串A和B差异程度越小、相似度越高。

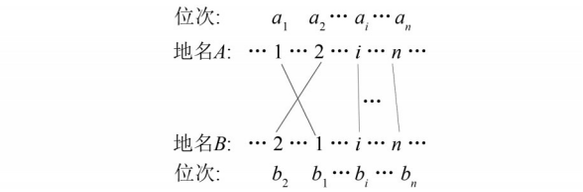


图1　重合字符匹配

Fig. 1　Overlapping Characters Matching

因此,定义偏距$O_i=b_i-a_i$用于表征构成匹配的重合字符的位次差异程度。设所有重合字符的偏距构成偏距向量$\boldsymbol{O}$,那么该向量记录了地名字符串A和B重合字符集的位次差异信息;记O_i的绝对值之和为匹配总偏距$D(\boldsymbol{O})$,该值代表了重合字符集的总体位次差异程度。

$$\boldsymbol{O}=[O_1\ O_2\cdots O_n] \tag{1}$$

$$D(\boldsymbol{O})=|O_1|+|O_2|+\cdots+|O_n| \tag{2}$$

1.2　未重合字符偏距

考虑未重合字符时,可以将偏距视为两个地名文本相似性的距离描述。重合字符的偏距绝对值越大,代表两个地名的文本相似性越小。未重合字符相当于无关字符,对两个地名文本相似性的影响应当比重合字符更大,即未重合字符的偏距绝对值应当更大。设想如下情形:任意找到一对重合字符(i,i),其中地名A的i变为一个A和B中都不曾出现的全新字符j,那么此时i和j两个字符都转变成未重合字符。定义每个未重合字符带来的偏距为O_w。根据直观认知,转变后的两个未重合字符(j,i)带来的偏距$2O_w$理应大于初始两个重合字符(i,i)偏距的绝对值$|a_i-b_i|$。即希望找到一个合适的O_w值,能够保证对所有的i满足:

$$2O_w>|a_i-b_i| \tag{3}$$

设A和B的字符串长度为$l(A)$和$l(B)$。由绝对值不等式得:

$$|a_i-b_i|\leqslant|a_i|+|b_i|\leqslant l(A)+l(B) \tag{4}$$

注意到a_i和b_i均为正整数,所以式(4)中$|a_i-b_i|$和$|a_i|+|b_i|$是无法取到等号的。从而可以得到:

$$|a_i-b_i|<l(A)+l(B) \tag{5}$$

因此将O_w定义为:

$$O_w=\frac{l(A)+l(B)}{2} \tag{6}$$

即能保证式(3)恒成立。

假设A和B之间共有W个未重合字符,那么所有未重合字符带来的偏距之和记为未匹配总偏距D_w,表达式为:

$$D_w=W\times\frac{l(A)+l(B)}{2} \tag{7}$$

需要指出的是,O_w值的定义是灵活的,并非一成不变。在实际应用场景下可以调整O_w的值,使得未重合字符对相似性的影响变大或者变小。

1.3　总偏距与偏距相似度

定义总偏距$D(A,B)$为匹配总偏距$D(\boldsymbol{O})$与未匹配总偏距D_w之和。那么总偏距$D(A,B)$便是两个地名字符串A和B差异程度的距离度量。当A和B没有重合字符时,$W=l(A)+l(B)$,此时给两地名字符串的差异程度最大,总偏距也正好取得最大值$D_{\max}$:

$$D_{\max}=\frac{(l(A)+l(B))^2}{2} \tag{8}$$

进而定义两地名字符串A和B的偏距相似度为$S(A,B)$:

$$S(A,B)=1-\frac{D(A,B)}{D_{\max}}\in[0,1] \tag{9}$$

由定义可知,当且仅当A和B完全相同时,$D(A,B)=0$,$S(A,B)=1$;当且仅当A和B不存在重合字符时,$D(A,B)=D_{\max}$,$S(A,B)=0$。

2　地名相似性识别的复杂情形

2.1　重合字符复现

通常情况下,重合字符在A和B中均刚好只出现一次;然而对于某些相对复杂的地名,重合字符是有可能在A和B中多次出现的。此时重合字符匹配不再是两个字符的简单对应,而是在多个字符中确定合适的匹配关系。任取一个重合字符i,假设它在A中出现x次,在B中出现y次,将A、B间的重合字符i尽可能地进行一一匹配,剩余$|x-y|$个无法匹配上的i则视为未重合字符处理。

直观来看,A(或B)中的每一个重合字符i只需要就近和B(或A)中的i匹配起来即可,不然匹配关系与直观认知不符(见图2),也会带来一些多余且无谓的偏距。例如A和B为同一个地名“尖尖山”时,如果刻意地不就近匹配重合字符“尖”而是选择交叉匹配,那么$D(A,B)=1+1=2$,与直观认知中A和B偏距应当为0是不符的。因此,将重合字符的匹配原则确定为“就近”,严格来讲即偏距最小原则。

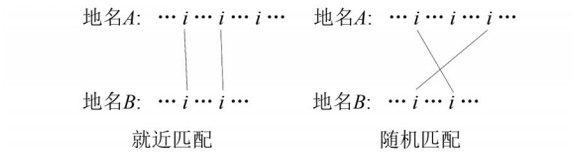


图2 就近匹配与随机匹配

Fig. 2 Nearest Matching and Random Matching

根据偏距最小原则,容易得到以下结论:如果A中位次为m、n($m<n$)的重合字符i与B中位次为r、s($r<s$)的重合字符i进行匹配,那么顺次匹配(m匹配r、n匹配s)的偏距必然小于等于逆次匹配(m匹配s,n匹配r)的偏距,如图3所示。更为形象的描述是:A、B间重合字符的匹配连线不存在交叉即可保证偏距最小。

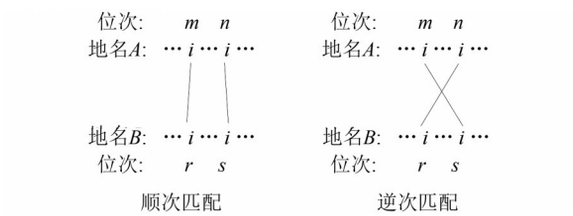


图3 顺次匹配与逆次匹配

Fig. 3 Forward Matching and Backward Matching

证明过程如下:记顺次匹配的偏距为$D_1=|m-r|+|n-s|$,逆次匹配的偏距为$D_2=|m-s|+|n-r|$。按表1分类讨论计算D_1-D_2即可。并且D_1-D_2有可能取到0,意味着有时顺次匹配和逆次匹配的偏距相等。

表1 分类讨论计算D_1-D_2

Table 1 Calculate D_1-D_2 by Listing Cases

r的范围	s的范围		
	$s\leqslant m$	$m<s\leqslant n$	$s>n$
$r\leqslant m$	0	$2(m-s)<0$	$2(m-n)<0$
$m<r\leqslant n$		$2(r-s)<0$	$2(r-n)\leqslant 0$
$r>n$			0

当$x=y$(重合字符i在A和B中出现次数相等)时,根据顺次匹配小于或等于逆次匹配的结论,对A、B间的所有重合字符i进行全体顺次匹配(如图4所示,A的第1个i匹配B的第1个i,A的第2个i匹配B的第2个i,依次下去,重合字符的匹配连线不存在任何交叉),就能满足偏距最小原则。

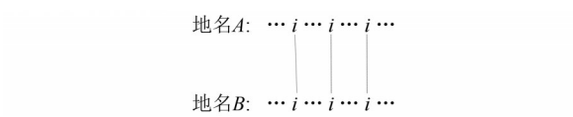


图4 全体顺次匹配

Fig. 4 Complete Forward Matching

当然,偏距取得最小值的匹配方案可能不止一种,本文只是选用了形式上最简洁的。例如,图5“杉杉”和“北京杉市杉”的两种匹配方案中,匹配总偏距分别为2+3=5、1+4=5,均满足偏距最小原则,而默认选用左侧全体顺次匹配的形式。

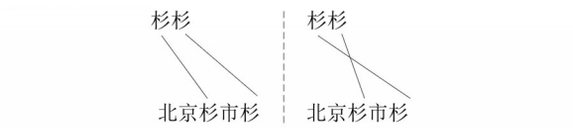


图5 偏距最小匹配

Fig. 5 Minimum-Offset Matching

当$x\neq y$(重合字符i在A和B中出现次数不等)时,如果$x>y$,先从A的x个重合字符i中取出y个,一共有C_x^y种方法,再将取出的y个i与B进行全体顺次匹配,便得到了C_x^y个偏距,最后选择其中偏距最小的一种匹配方案。如果$x<y$,同理从B中y个重合字符中取出x个即可。

2.2 字符片段偏移

以上偏距计算都是基于字符而非词或词组。这样做的优点在于不会因为分词问题而捕捉不到全称与简称之间的相似性;缺点在于存在词或词组偏移时会产生不合理的偏距值。例如,A=“北京南湖路”,B=“南湖路北京”,C=“湖北南京路”,其中B相较于A虽然“北京”和“南湖路”两词的顺序颠倒,但所指代的地理实体没有发生改变,C相较于A字符顺序完全改变,导致所指代的地理实体也发生改变,因此$D(A,C)$应当大于$D(A,B)$。但是计算结果为$D(A,C)=1+2+0+3+0=6<D(A,B)=3+3+2+2+2=12$。此时偏距值不合理的关键就在于计算$A$、$B$间的偏距时应当视为“北京”和“南湖路”两个词的整体偏移(即等于每个词的首字符偏移),而非5个字符的偏移之和。“北京”和“南湖路”两词的偏距分别是3和2,从而$D(A,B)=3+2=5$是较为合理的,也满足直观认知$D(A,C)>D(A,B)$。因此,需要进一步识别出词或词组的偏移,即连续字符片段的整体偏移后,再去计算偏距。识别过程如图6所示,此时偏距的概念也得到推广——从仅限于字符之间的偏距,变为字符片段(包含单个字符)之间的偏距。

2.3 相关定义更新

由于§2.1和§2.2对复杂情形的处理,前文关于$\boldsymbol{O}$、$D(\boldsymbol{O})$、D_w等定义的描述不再完全适用。相关定义更新如下:

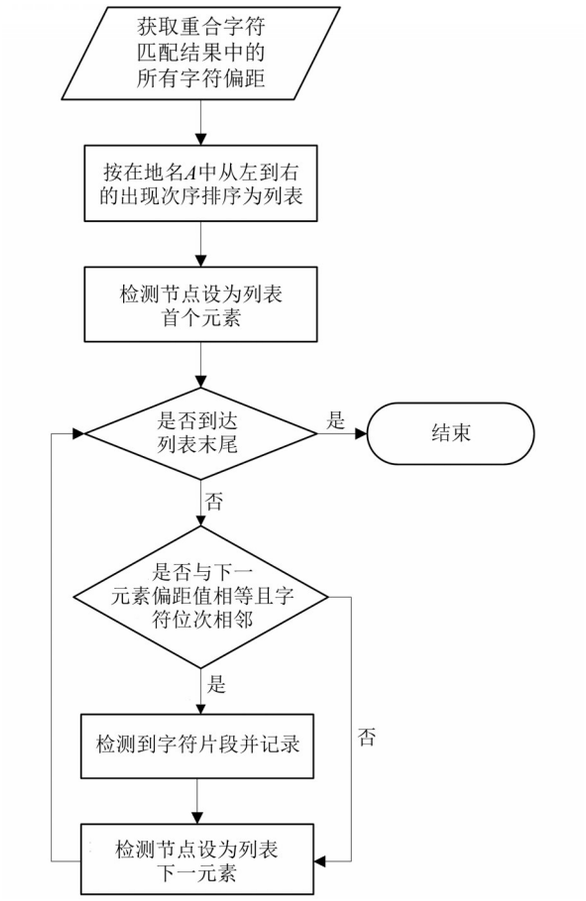


图6　字符片段识别

Fig. 6　Character Fragment Recognition

经过重合字符匹配和字符片段识别后,所有匹配成功的重合字符片段(包含单个重合字符)的首字符偏距O_i构成偏距向量$\boldsymbol{O}$。所有O_i的绝对值之和构成匹配总偏距$D(\boldsymbol{O})$。每个未匹配的字符(包括匹配失败的重合字符和未重合字符)带来的偏距均为O_w,它们的和构成未匹配总偏距D_w。此时总偏距$D(A,B)$的定义保持不变,仍为$D(\boldsymbol{O})$与D_w两者之和。相似度$S(A,B)$的定义保持不变。

其中,总偏距$D(A,B)$作为新提出的距离度量有以下特点:

1)满足正定性。显然$D(A,B)\geqslant 0$恒成立,满足正定性。

2)满足对称性。只要$D(A,B)$和$D(B,A)$的重合字符匹配都遵循偏距最小原则,那么计算结果必然是同一个最小值,即满足对称性$D(A,B)=D(B,A)$。

3)不满足三角不等式。例如A=“北京大学”,B=“南湖”,C=“南湖西园”,$D(A,B)=D_w=(4+2)\times(4+2)/2=18$,$D(B,C)=D_w=2\times(4+2)/2=6$,$D(A,C)=D_w=(4+4)\times(4+4)/2=32$,此时$D(A,C)>D(A,B)+D(B,C)$不满足三角不等式。

事实上,总偏距更像是两个地名间相似性推导过程中的一个分值,且满分值$D_{\max}$本身也会随地名而变化,所以不宜直接用它表达地名相似性而应当使用偏距相似度。当考虑3个及以上地名间的相似性关系时,可计算它们两两之间的偏距相似度。

3　相似度对比实验

为观察本文偏距算法的相似度表达能力与特性,先设计8对不同差异程度的地名和9对相似地名计算偏距相似度,并与经典的Jaccard系数和编辑距离相似度进行对比。如表2和表3所示。其中打钩(√)表示两地名指代的为同一个地理实体。

表2　辨别不同差异程度的地名

Table 2　Identify Toponyms with Distinct Degrees of Difference

差异程度	地名		Jaccard系数	编辑距离相似度	偏距相似度
完全相同	南湖路	南湖路(√)	1.00	1.00	1.00
仅位次不同	南湖路与双河南路交叉口	双河南路与南湖路交叉口(√)	1.00	0.45	0.95
	北京南湖路	南湖路北京(√)	1.00	0.20	0.90
部分字符相同	南湖路	北京南湖路(√)	0.60	0.60	0.69
	北京大学	北大(√)	0.50	0.50	0.61
	北京师范大学	首都师范大学	0.44	0.57	0.60
	北京师范大学	北京新青年学校	0.30	0.43	0.46
完全不同	南湖路	北京大学	0.00	0.00	0.00

在辨别不同差异程度的地名时,本文的偏距算法对相似度的刻画更为细致合理,符合设计的初衷。遇到两个地名完全相同、仅位次不同、部分字符相同和完全不同等情况时,偏距相似度从1到0逐渐下降,且下降过程较为连续、平滑。与之相较,Jaccard系数整体呈下降趋势但在仅位次不同时恒为1,编辑距离相似度呈曲折变化,两者均不如偏距相似度数值变化均匀。

表3 辨别相似地名

Table 3 Identify Similar Toponyms

类型	地名		Jaccard系数	编辑距离相似度	偏距相似度
位次差异	北京南湖路	南湖路北京(√)	1.00	0.20	0.90
	北京南湖路	湖北南京路	1.00	0.40	0.88
	北京南湖路	南京路湖北	1.00	0.40	0.84
	北京南湖路	湖南北京路	1.00	0.20	0.88
字符差异	北京南湖路	南湖路(√)	0.60	0.60	0.69
	北京南湖路	北京市南湖路(√)	0.83	0.83	0.89
	北京南湖路	南河路	0.33	0.40	0.38
	北京南湖路	北京市湖光南街	0.50	0.43	0.63
	北京南湖路	北京市团结湖南路	0.63	0.50	0.66

在辨别相似地名时,可进一步观察到偏距算法的特性:对字符差异重视程度较大,重合字符的比例减小时相似度显著降低;对单纯的位次差异重视程度较小,但也能体现在相似度数值的细微差距中。与之相较,Jaccard系数无法识别到位次差异;编辑距离相似度对位次差异和字符差异的重视程度没有显著区别。偏距算法的该特性决定了其会更适用于地名匹配等应用场景。例如位次差异的4个地名中,仅有偏距算法能通过最大相似度分辨出“北京南湖路”与“南湖路北京”构成匹配;字符差异的5个地名中,偏距算法也能通过寻找相似度最大的两项判别出“北京南湖路”与“南湖路”“北京市南湖路”构成匹配。

为测试本文偏距算法在地名匹配应用场景下的实际效果,通过数据预处理和随机筛选,获得了5万条中国北京市范围内的高德地图兴趣点(point of interest,POI)数据(见图7,图7的底图来源是天地图,行政界线未修改),其中每条数据均满足:(1)恰含一个地名和一个别名;(2)地名和别名均已去除特殊符号;(3)若地名或别名有英文字母则统一为小写;(4)别名和地名不等同[17-18]。

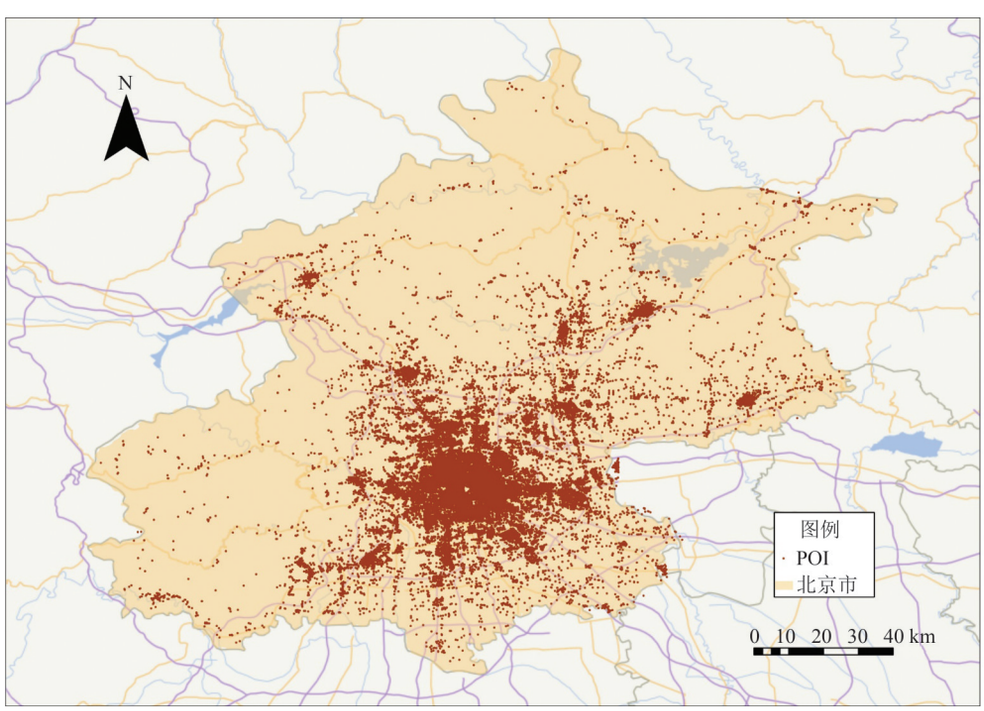


图7 POI位置分布

Fig. 7 POI Location Distribution

测试方案为:对于任意一个别名,计算它和每一个地名的相似度,寻找该别名的最大相似度地名;若最大相似度地名仅有一个且和别名恰好对应同一条POI数据,则为匹配正确;统计5万个别名中匹配正确的数量,并记录算法运行时间。分别采用Jaccard系数、编辑距离相似度和偏距相

似度进行上述测试,匹配结果如表4所示。部分别名的最大相似度地名如表5所示,如果该别名的最大相似度地名存在多个,则取其中一个作为示例,打钩(√)表示该别名匹配正确。

表4　匹配结果

Table 4　Matching Results

匹配算法	匹配正确数/个	匹配正确率/%	运行时间/s
Jaccard系数	30 650	61.30	4 005.43
编辑距离相似度	25 079	50.16	6 936.29
偏距相似度	31 821	63.64	2 940.56

偏距相似度的匹配正确率为63.64%,相较于Jaccard系数提升2.34%,相较于编辑距离相似度提升13.48%。参考表5实例可发现,偏距相似度无论遇到错别字(例如“万恒万泽中心”-“方恒万泽中心”)、字符片段颠倒(例如“北环西路大发路路口”-“大发路北环西路路口”)还是各类全称或简称(例如“央央生物”-“北京央央生物技术有限公司”),都能凭借对重合字符片段的敏感性而更容易匹配正确,而Jaccard系数没有考虑字符顺次结构,导致“龙跃苑4区12号楼”错误匹配到“龙跃苑三区14号楼”,编辑距离相似度对片段颠倒和简称辨识能力较差,因此“北环西路大发路路口”和“建壮咨询”等别名均没有匹配正确。

表5　最大相似度地名

Table 5　Toponyms with Maximum Similarity

别名	地名	Jaccard系数	编辑距离相似度	偏距相似度
盛世宾馆	鹏翔盛世宾馆	鹏翔盛世宾馆 0.67(√)	鹏翔盛世宾馆 0.67(√)	鹏翔盛世宾馆 0.76(√)
万恒万泽中心	方恒万泽中心	方恒万泽中心 0.83(√)	方恒万泽中心 0.83(√)	方恒万泽中心 0.83(√)
北环西路大发路路口	大发路北环西路路口	大发路北环西路路口 1(√)	北西路西韩路路口等 0.67	大发路北环西路路口 0.95(√)
阜三社区	阜石路第三社区	阜石路第三社区 0.57(√)	建北三社区 0.60	阜石路第三社区 0.68(√)
建壮咨询	北京市建壮咨询有限公司	北京市建壮咨询有限公司 0.36(√)	信索咨询等 0.50	北京市建壮咨询有限公司 0.51(√)
幽幽独语	幽幽独语北京文化发展有限公司	茶语 0.25	幽幽独语北京文化发展有限公司 0.29(√)	幽幽独语北京文化发展有限公司 0.44(√)
龙跃苑4区12号楼	龙跃苑四区12号楼	龙跃苑三区14号楼等 0.80	龙跃苑四区12号楼 0.89(√)	龙跃苑四区12号楼 0.89(√)
奇迹美昔健康俱乐部	奇迹美昔	华人健康俱乐部 0.45	华人健康俱乐部 0.56	奇迹美昔 0.62(√)
央央生物	北京央央生物技术有限公司	三生万物 0.40	昱欣生物科技等 0.33	北京央央生物技术有限公司 0.48(√)
郦城工作区西门	郦城西门	郦城西门 0.57(√)	郦城二区西南门等 0.57	郦城二区西南门 0.70
达懿集团	山西达懿电子集团有限公司	雅世集团等 0.33	雅世集团等 0.50	雅世集团等 0.50
大溪地1区16号楼	大溪地一区16号楼	大溪地1区26号楼等 0.89	大溪地1区26号楼等 0.89	大溪地1区26号楼等 0.89

虽然偏距相似度的构建初衷是提升地名相似度评价的合理性,但在实验过程中它的运行时间指标同样展现出明显优势。其耗时2 940.56 s,仅为Jaccard系数的73.41%,编辑距离相似度的42.39%。时间效率优势的原因主要在于偏距相似度算法的第一步便是定位所有重合字符,作为后续处理的基础。一般只有当别名遇到对应地名或少数相似地名时才会存在较多重合字符,大多数无重合字符或重合字符很少的地名都将被快速处理,所以运行时间更短。此外,Jaccard系数算法虽然也是基于寻找重合字符,但它运用了集合去重复和取交集等复杂操作,所以运行时间相对较长。

综上,偏距相似度是地名匹配测试中总体表

现最佳的,在匹配正确率和运行时间两项指标上都更为优秀。但是也需承认在5万条数据量的情形下,偏距相似度算法的匹配正确率相比于其他两种算法提升幅度有限。这一方面是因为数据量较大时,同类型的相似地名的频繁出现容易导致匹配错误。例如,统计的5万个别名中有2 402个包含“停车场”,5万个地名中亦有2 449个包含“停车场”;随着同类型的相似地名的频繁出现,它们在匹配过程中产生混淆的概率也会增加。另一方面则是因为偏距相似度本质仍为文本相似性度量,只能关注字符层面的相似性而难以顾及复杂的语义。当附加过多通名字符[19-20]或地址信息字符[21-22]时,偏距相似度会显著降低从而难以正确匹配,例如“达懿集团”未能正确匹配到“山西达懿电子集团有限公司”;当出现“1”和“一”等语义一致的字符时,偏距相似度同样不能理解从而导致失败,例如“大溪地1区16号楼”没有正确匹配到“大溪地一区16号楼”。

4 结 语

本文通过计算重合字符偏距定义未重合字符偏距,以及处理重合字符复现和字符片段偏移等复杂情形,最终构建出新的中文地名相似性度量。本文提出的偏距相似度具有数值刻画细致的优点和重视字符差异且能识别位次差异的特性,在匹配正确率和运行时间上的表现均优于Jaccard系数和编辑距离相似度。未来可在处理复杂情形的算法设计、对其他语种适用性[23]和考虑地名语义结构[24]上进一步优化。需要指出的是,考虑地名语义结构通常需要引入较为全面的通名词典[25-26]或借助各类复杂的自然语言模型[27-28],此类方法在提升匹配正确率的同时也会牺牲时间效率[29]。

参 考 文 献

[1] 姜松言,宋国民,贾奋励,等. 结合地名知识和位置信息的中文地名匹配方法[J]. 测绘通报,2022(5):120-125.
JIANG Songyan, SONG Guomin, JIA Fenli, et al. Combining Toponymy Knowledge and Location Information for Chinese Place Names Matching[J]. *Bulletin of Surveying and Mapping*, 2022(5): 120-125.

[2] 张亮,覃俊华,王爱华,等. 一种基于地理实体的时空数据模型设计[J]. 地理空间信息,2022,20(10):110-112.
ZHANG Liang, QIN Junhua, WANG Aihua, et al. Spatio-temporal Data Model Design Based on Geographical Entity[J]. *Geospatial Information*, 2022, 20(10): 110-112.

[3] 赵云鹏,孙群,刘新贵,等. 面向地理实体的语义相似性度量方法及其在道路匹配中的应用[J]. 武汉大学学报(信息科学版),2020,45(5):728-735.
ZHAO Yunpeng, SUN Qun, LIU Xingui, et al. Geographical Entity-Oriented Semantic Similarity Measurement Method and Its Application in Road Matching[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(5): 728-735.

[4] RECCHIA G, LOUWERSE M. A Comparison of String Similarity Measures for Toponym Matching[C]// The First ACM SIGSPATIAL International Workshop on Computational Models of Place, Orlando, FL, USA, 2013.

[5] 王俊超,孙群,肖计划,等. 相似性理论技术在地名匹配中的应用[J]. 测绘科学技术学报,2013,30(6):643-646.
WANG Junchao, SUN Qun, XIAO Jihua, et al. Research on Theory of Similarity in Place Names Matching[J]. *Journal of Geomatics Science and Technology*, 2013, 30(6): 643-646.

[6] 徐流畅. 预训练深度学习架构下的语义地址匹配与语义空间融合模型研究[D]. 杭州:浙江大学,2020.
XU Liuchang. Research on Semantic Address Matching and Semantic Space Fusion Model Under Pre-training Deep Learning Framework[D]. Hangzhou: Zhejiang University, 2020.

[7] 陈俊月,郝文宁,张紫萱,等. 基于改进句子相似度算法的释义识别研究[J]. 计算机工程,2020,46(9):76-82.
CHEN Junyue, HAO Wenning, ZHANG Zixuan, et al. Research on Paraphrase Identification Based on Improved Sentence Similarity Algorithm[J]. *Computer Engineering*, 2020, 46(9): 76-82.

[8] 周艳平,李金鹏. 一种基于词向量及位置编码的Jaccard相似度算法[J]. 青岛科技大学学报(自然科学版),2020,41(6):93-98.
ZHOU Yanping, LI Jinpeng. Jaccard Similarity Algorithm Based on Word Embedding and Position Encoding[J]. *Journal of Qingdao University of Science and Technology (Natural Science Edition)*, 2020, 41(6): 93-98.

[9] 张琛,陈张建,刘江涛,等. Lucene自适应分词的地址匹配方法改进与实现[J]. 测绘科学,2021,46(10):185-193.
ZHANG Chen, CHEN Zhangjian, LIU Jiangtao, et

al. Improvement and Realization of Address Matching Based on Adaptive Word Segmentation in Lucene [J]. *Science of Surveying and Mapping*, 2021, 46(10): 185-193.

[10] 徐嘉康, 张晨, 王柳静, 等. 非均权-动态规划地址匹配算法设计与实现[J]. 小型微型计算机系统, 2022, 43(3): 530-535.
XU Jiakang, ZHANG Chen, WANG Liujing, et al. Design and Implementation of Non-equal Power-Dynamic Programming Address Matching Algorithm [J]. *Journal of Chinese Computer Systems*, 2022, 43(3): 530-535.

[11] LEVENSHTEIN V I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals[J]. *Soviet Physics Doklady*, 1966, 10: 707.

[12] 王俊超, 刘晨帆, 徐明世, 等. 语义相似性度量技术在地名匹配研究中的应用[J]. 辽宁工程技术大学学报(自然科学版), 2012, 31(6): 871-874.
WANG Junchao, LIU Chenfan, XU Mingshi, et al. Application of Semantic Similarity Measurement Technology in Place Name Matching[J]. *Journal of Liaoning Technical University* (*Natural Science*), 2012, 31(6): 871-874.

[13] 叶鸥, 张璟, 李军怀. 中文地名数据清洗中的重复字段匹配[J]. 应用科学学报, 2013, 31(2): 212-220.
YE Ou, ZHANG Jing, LI Junhuai. Duplicate Field Matching for Data Cleaning of Chinese Placenames [J]. *Journal of Applied Sciences*, 2013, 31(2): 212-220.

[14] 刘斐, 贺向阳, 邹志云. 基于全文索引知识图谱的危化品运输地址匹配研究[J]. 计算机应用研究, 2022, 39(2): 407-410.
LIU Fei, HE Xiangyang, ZOU Zhiyun. Address Matching Based on Full-Text Indexed Knowledge Graph for Hazardous Materials Transportation[J]. *Application Research of Computers*, 2022, 39(2): 407-410.

[15] 吴睿, 龙华, 熊新, 等. 一种多策略结合的地址匹配算法[J]. 河南理工大学学报(自然科学版), 2019, 38(5): 124-129.
WU Rui, LONG Hua, XIONG Xin, et al. A Multi-strategy Combined Address Matching Algorithm[J]. *Journal of Henan Polytechnic University* (*Natural Science*), 2019, 38(5): 124-129.

[16] 盖森, 刘建忠, 熊伟, 等. 自然语言空间查询中应用编辑距离的规则匹配模型[J]. 测绘科学技术学报, 2015, 32(4): 416-421.
GE Sen, LIU Jianzhong, XIONG Wei, et al. Research on Rule Matching in Natural Language Spatial Query Based on Levenshtein Distance[J]. *Journal of Geomatics Science and Technology*, 2015, 32(4): 416-421.

[17] KıLıNÇ D. An Accurate Toponym-Matching Measure Based on Approximate String Matching[J]. *Journal of Information Science*, 2016, 42(2): 138-149.

[18] SANTOS R, MURRIETA-FLORES P, MARTINS B. Learning to Combine Multiple String Similarity Metrics for Effective Toponym Matching[J]. *International Journal of Digital Earth*, 2018, 11(9): 913-938.

[19] 张剑, 叶远智, 翁宝凤. 面向智慧城市的高精度地名地址匹配方法[J]. 测绘与空间地理信息, 2019, 42(11): 166-169.
ZHANG Jian, YE Yuanzhi, WENG Baofeng. High Precision Place Name and Address Matching Method for Smart City[J]. *Geomatics & Spatial Information Technology*, 2019, 42(11): 166-169.

[20] 曹存相. 基于多数据源的地名匹配及地名翻译方法研究[D]. 北京: 北京建筑大学, 2019.
CAO Cunxiang. Research on Place Name Matching and Place Name Translation Based on Multi-data Sources[D]. Beijing: Beijing University of Civil Engineering and Architecture, 2019.

[21] 应申, 李威阳, 贺彪, 等. 基于城市地址树的地址文本匹配方法[J]. 地理信息世界, 2017, 24(6): 81-86.
YING Shen, LI Weiyang, HE Biao, et al. Address Text Matching Method Based on City Address Tree [J]. *Geomatics World*, 2017, 24(6): 81-86.

[22] 吕欢欢. 基于地理信息公共服务平台的语义地名地址匹配方法研究[D]. 阜新: 辽宁工程技术大学, 2014.
LÜ Huanhuan. Research on Semantic Place Name Address Matching Method Based on Geographic Information Public Service Platform[D]. Fuxin: Liaoning Technical University, 2014.

[23] 占斌斌, 赵英, 赵婷婷, 等. 归类识别地名匹配算法[J]. 北京测绘, 2018, 32(4): 484-487.
ZHAN Binbin, ZHAO Ying, ZHAO Tingting, et al. Classification and Recognition of Place Names Matching Algorithm [J]. *Beijing Surveying and Mapping*, 2018, 32(4): 484-487.

[24] KAFFES V, GIANNOPOULOS G, KARAGIANNAKIS N, et al. Learning Domain Specific Models for Toponym Interlinking [C]//The 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, 2019.

[25] 唐旭日,陈小荷,张雪英.中文文本的地名解析方法研究[J].武汉大学学报(信息科学版),2010,35(8):930-935.
TANG Xuri, CHEN Xiaohe, ZHANG Xueying. Research on Toponym Resolution in Chinese Text[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(8): 930-935.

[26] 张春菊,张雪英,吉雷静,等.地名通名与地理要素类型的关系映射[J].武汉大学学报(信息科学版),2011,36(7):857-861.
ZHANG Chunju, ZHANG Xueying, JI Leijing, et al. Relation Mapping Between Generic Terms of Place Names and Geographical Feature Types[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(7): 857-861.

[27] 邬伦,刘磊,李浩然,等.基于条件随机场的中文地名识别方法[J].武汉大学学报(信息科学版),2017,42(2):150-156.
WU Lun, LIU Lei, LI Haoran, et al. A Chinese Toponym Recognition Method Based on Conditional Random Field[J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(2): 150-156.

[28] 魏勇,李鸿飞,胡丹露,等.一种基于复合特征的中文地名识别方法[J].武汉大学学报(信息科学版),2018,43(1):17-23.
WEI Yong, LI Hongfei, HU Danlu, et al. A Method of Chinese Place Name Recognition Based on Composite Features[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(1): 17-23.

[29] SANTOS R, MURRIETA-FLORES P, CALADO P, et al. Toponym Matching Through Deep Neural Networks[J]. *International Journal of Geographical Information Science*, 2018, 32(2): 324-348.

(上接第1403页)

HU Liuru, TANG Xinming, ZHANG Zhi, et al. Accuracy Optimization and Assessment of GF-7 Satellite Multi-source Remote Sensing Data[J]. *Infrared and Laser Engineering*, 2021, 50(3): 384-395.

[18] XIE J F, HANG G H, LIU R, et al. Design and Data Processing of China's First Spaceborne Laser Altimeter System for Earth Observation: Gaofen-7[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, 13: 1034-1044.

[19] 李国元,高小明,陈继溢,等.资源三号02星激光测高数据质量分析[J].遥感学报,2019,23(6):1159-1166.
LI Guoyan, GAO Xiaoming, CHEN Jiyi, et al. Data Quality Analysis of ZY-3 02 Satellite Laser Altimeter[J]. *Journal of Remote Sensing*, 2019, 23(6): 1159-1166.

[20] 王密,韦钰,杨博,等.ICESat2/ATLAS全球高程控制点提取与分析[J].武汉大学学报(信息科学版),2021,46(2):184-192.
WANG Mi, WEI Yu, YANG Bo, et al. Extraction and Analysis of Global Elevation Control Points from ICESat2/ATLAS Data[J]. *Geomatics and Information Science of Wuhan University*, 2021, 46(2): 184-192.