文章编号：1008-0562(2012)06-0871-04

语义相似性度量技术在地名匹配研究中的应用

王俊超[1]，刘晨帆[1]，徐明世[2]，纪　山[3]，兰　伟[4]
（1.信息工程大学 测绘学院，河南 郑州 450052；2.69027 部队，新疆 乌鲁木齐 830000；
3.61512 部队，北京 100088；4.61773 部队，新疆 乌鲁木齐 830019）

摘　要：完善地名自动匹配更新技术，以地名属性信息为契机，采用语义相似性度量技术进行地名匹配研究.研究结果表明：语义相似性度量技术对于地名的自动化乃至智能化匹配技术有着不可替代的支撑作用，并且可以大大简化以及完善原有的地名匹配方法，值得深入研究并实践采用.该研究成果打破了原有地名匹配研究的局限性，有助于其他学者对地名匹配方法的进一步完善与拓展，展现了良好的地名匹配研究前景.

关键词：地名匹配；编辑距离；生僻字；语义相似性；相似度算法；属性匹配；相似性匹配；地名更新
中图分类号：P 208　　**文献标志码：**A

Application of semantic similarity measurement technology in place name matching

WANG Junchao[1], LIU Chenfan[1], XU Mingshi[2], JI Shan[3], LAN Wei[4]
(1. Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450052, China;
2. No. 69027 Army, Wulumuqi 830000, China; 3. No. 61512 Army, Beijing 100088, China;
4. No. 61773 Army, Wulumuqi 830019, China)

Abstract: To improve the technology of place name automatic matching updating, this study investigates the properties of place name, and introduces the technology of semantic similarity measurement. The study result demonstrates that the semantic similarity measurement technology plays an indispensable role in automatic and intelligent matching of place name. Furthermore, it is able to significantly simplify and improve the traditional matching method. The method of semantic similarity measurement proposed breaks the limitation of traditional matching method, and provides a basis for future study in place name matching.

Key words: place name matching; edit distance; rare characters; semantic similarity; similarity algorithm; attribute matching; matching similarity; place names update

0　引　言

语义相似性研究的主要是两个词语（语句）之间的相似程度，就犹如在初中数学中学习的相似三角形一样，只不过语义相似性属于属性相似性范畴，而三角形相似性属于几何相似性范畴.在自然语言中，词语与词语之间的关系是比较复杂的，尤其是地理名称中.有些地名的字属于生僻字（如邕仔岛、白鸳、藕子头、逯里子、槔庵子、婁埠），有些地名的字是多音字（如长安、宿迁市、海参崴），有些地名的字只在地名应用中读一个特殊的音（如宁波市的邱隘（gà）路），有些地名与所属小范围的地名重名（如吉林省吉林市），有些地名……凡此种种，都是在地名相似性研究中应该特殊处理的情况.而这些地名所对应的实体却是一样的，即它们一般都是指代同一地名本体[1-2].在实际应用中，用户或者计算机不会领会从字面上去理解地名的相似性或者说难以区分本体[3]，这就需要将这种复杂的关系用一种简单的数据来度量.因此，本文将语义相似性研究技术引入地名相似性度量中[4].

1　语义相似性

在讨论语义相似性之前先介绍一下“编辑距离”.“编辑距离”是指为了使两个词语W_1与W_2趋于相同而进行插入、删除、移位、翻转等操作的次数.设W代表词语，a，b，c……代表汉字.当W_1=abcdefg，W_2=aaadddg时，W_1与W_2之间的编辑

收稿日期：2012-05-15
基金项目：国家自然科学基金资助项目（41071297；41201391）
作者简介：王俊超（1987-），男，陕西 渭南人，硕士研究生，主要从事数字地图制图技术方向研究. 本文编校：朱艳华

距离即为 4.然而，“编辑距离”度量方法也存在一定的局限性.比如，当W_1=abcdefg，W_3=aaaabbb，W_4=bbbaaaa，时，W_3与W_1和W_4的“编辑距离”均为 6，但是，显然W_3与W_4相似的多.

语义相似度是指两个或多个词语（短语）之间的相似程度，主要包括“字数”的相似度和“意思”的相似度两层含义.“字数”的相似度可以用“编辑距离”来度量或测评，而“意思”的相似性度量可以用词义的近似性来度量或测评.

Dekang Lira认为任何两（多）个词语的相似性是由它们的Commonality （共性）和 Differences（个性）所决定的，根据信息论相关知识及理论研究，可以得出两个词语A、B的相似性计算公式[5]

$$Sim(A,B)=\frac{\log p(Common(A,B))}{\log p(description(A,B))}\ , \qquad (1)$$

式中，分子$\log p(Common(A,B))$表示描述 A、B 共性所需要的信息量；分母$\log p(description(A,B))$表示完整地描述 A、B 所需要的信息量.此公式是根据词语的度量的“意思”距离（即“语义”距离）所得出的.

本文对语义相似性做出如下定义：

语义相似性是指多个词语在不同的前后文语境中，可以交叉使用而不改变原有语句的句法与语义结构的特性.

如果多个词语在不同的语句环境中可以交叉使用，而使原有的文本句法或语义发生变化的可能性越小，那么这些词语的相似性就越高，否则，它们的相似性就越低.

记两个词语W_1，W_2的相似性为$Sim(W_1, W_2)$，它们间的语义距离（或者说“编辑距离”）为$Dis(W_1, W_2)$，则根据相关知识可以推导出 又一个相似性计算的公式 [6]

$$Sim(W_1,W_2)=\frac{\alpha}{\alpha+Dis(W_1,W_2)}\ , \qquad (2)$$

式中，α 是一个可变参数.对 α 做如下定义：当两个词语的相似度为 0.5 时这两个词语间的距离值为 α.这个相似性计算公式是基于“编辑距离”信息量得出的，而式（1）则是根据两个词语的信息量的共性与个性所得出的，是有本质区别的.

2 语义相似性度量算法

2.1 “首字符”与“后缀词”方法

在进行地名数据匹配过程中，为了缩短匹配时间、提升匹配效率，也可以采用“首字符”法和“后缀词”法.实际上，这两种方法是互为“倒数”的.即一个是先正向匹配，另一个则是先逆向匹配.

正向匹配是指在对词语进行断字取词时按照从左到右的顺序逐一加字成词的方法，即针对“河南省郑州市”这个地名而言，先取“河”字在地名数据库中比较核对，若是未登录词，则依次加一个字并进行比对核查，即核查“河南”是否为已登录词.若为，则判定其为一个地名，否则，继续逐一加字继续进行匹配核查.若碰见“省”、“市”、“县”、“区”等地名“后缀词”，则将此后缀词归为紧挨着前面的地名.例如，若出现“河南”地名之后又出现了“省”字，则将“省”字归为前面的“河南”地名.部分不规范地名数据中仅有类似于“河南郑州”的地名数据，并不是“河南省郑州市”这样规范的地名.逆向匹配的方法与正向匹配的思想相同，只是取词时的方向不同.逆向匹配是从右往左的顺序进行匹配.实验证明[7-8]，正向最大匹配时误差约为 1/169，逆向最大匹配时的误差约为 1/256.因此，通常选用逆向匹配法.

对于相当一部分地名，它们前面的“冠名”是一致的.例如，“河南省郑州市二七区”与“河南省郑州市中牟县”这两个地名中，“二七区”与“中牟县”前面的“冠名”都是“河南省郑州市”，是一致的.因此，在进行地名匹配时，它们就很有可能具有一定的相似性，至少在地名等级上可以做出一定的判定决策.这就是所谓的“首字符”匹配方法.因此，在对于某些行政区划进行匹配更新时，可以只针对它们发生变化的那一部分进行匹配更新.例如，县市级以上的地名一般不会发生变化.

在中国，地理名称的“后缀词”都具有一定的等级区别和相似性[7]，例如省、市、区（县）、乡（镇）、村、旗、州、府、自治区、自治州、自治县、街道、弄、堡、庄、路、巷、胡同、江、河、湖、海、洋、山、峰、岛等.在进行地名匹配时，也可以从“后缀词”着手进行筛选匹配.如果说某两个地名具有相似性，那么它们的“后缀词”很有可能是一致的.因此，在匹配时也可以根据“后缀词”法进行辅助匹配，从而缩短匹配时间并提升匹配效率.

2.2 相似度计算

假设词语集合为O，$\exists$ 词语M与词语N s. t. $M\subseteq O$并且$N\subseteq O$.设词语$A=a_1a_2\cdots a_i\cdots a_m$，其中$a_1,a_2,\cdots,a_i,\cdots,a_m\in M$，$1\leqslant i\leqslant m$；设词语$B$=

$b_1b_2\cdots b_j\cdots b_n$，其中$b_1,b_2,\cdots,b\cdots,b_n\in N$，$1\leqslant j\leqslant n$.设$q\leqslant\min(m,n)$，$k=0$；如果$a_p=b_p$（$p=1,2,\cdots,q$），那么$k$++.由此，对词语的相似度给出如下定义

$$\rho_{AB}=\frac{k}{m}\times 100\%\ ,\qquad (3)$$

$$\rho_{BA}=\frac{k}{n}\times 100\%\ .\qquad (4)$$

式（3）表示词语A与词语B的相似度，式（4）表示词语B与词语A的相似度.令$\rho=\max(\rho_{AB},\rho_{BA})$，则$\rho$表示$A$、$B$这两个词语的相似度.

2.3　地名匹配中的相似性应用

中国的地名大多数都是按照省、市、区（县）、乡（镇）、村这个等级划分的，因此，这些行政单位也就默认的成为了地名的后缀词.在社会发展变化中，由于种种历史或地理原因，成为地名后缀词的还有旗、州、府、自治区、自治州、自治县、街道、弄、堡、庄、路、巷、胡同等以及成为自然界地名后缀的江、河、湖、海、洋、山、峰、岛等.

针对前两小节的算法设计，结合前文的总结分析，可以得出地名语义匹配的相似性算法[9].

用s代表任意地名字串，a、b、c表示地名中的单个汉字，Σ^*代表地名集合，对于任意给定的地名字串$s\in\Sigma^*$，用$|s|$表示它的长度.令s_i代表地名字串s中第i个字符，用$s_{i,\cdots,j}$表示$s_i,s_{i+1},\cdots,s_j$.下面，将给出近似地名字串匹配的形式化定义[9-10].

设Σ^*是全体地名数据库，其大小为$|\Sigma|=\sigma$，匹配文本$T\in\Sigma^*$长度为$n=|T|$，模式串$P\in\Sigma^*$的长度为$m=|P|$.设$k\in N^+$代表允许的最大误差，$\rho=k/m$表示误差率，ed()表示距离函数.对地名字串的近似匹配定义为给定T、P、k、ed()，求所有满足条件ed（P，$T_{i,\cdots,j}$）$\leqslant k$的字符串$T_{i,\cdots,j}$，或者求解满足匹配条件开始位置的i以及终止位置的j.称函数ed为“编辑距离”（Edit Distance）.

文献[9]规定：若$M[T_j]=0$，那么称T_j为无用汉字；若$M[T_j]\neq 0$，则称T_j为有用汉字.其中$1\leqslant j\leqslant m$，$M[]$的含义与BPM算法[9-11]中的模式匹配向量组$M[]$相同.Bad（s）表示地名字串s中无用字符数，Good（s）表示地名字串s中有用字符数.S_j对应于BPM算法中扫描至T_j后的Score值.一个有用汉字最多可以使得Score减 1，而一个无用汉字将不会使Score减1.因此，若$S_j=m$且T_{j+1}是无用汉字，则$S_{j+1}=S_j=m$.

过滤引理　若$S_j=m$且$Bad(T_{j+1,\cdots j+m})\geq k+1$，则必有$S_{j+i}\geqslant k+1$，其中$1\leqslant i\leqslant m$.

跳跃引理　若$S_j=m$且Bad$(T_{j+1,\cdots,}T_{j+m})\geqslant k+1$，$T_{j+i}$是从$T_{j+m}$开始从右往左数的第$k+1$个无用字符，则将$T_{j+1,\cdots,j+i-1}$全改为无用字符不影响$P$与$T$从$T_j$开始的后续匹配.

在跳跃引理前提下，可直接令$S_{j+i}=m$.实际上若能记住T_{j+i}右边第一个有用字符$T_{j+\text{last}}$，必有$S_{j+\text{last-1}}=\ldots=S_{j+\text{i}}=m$，故可令$S_{j+\text{last-1}}=m$，然后继续扫描进行匹配计算.

依据前文的分析总结，给出地名匹配框架流程图，见图 1.

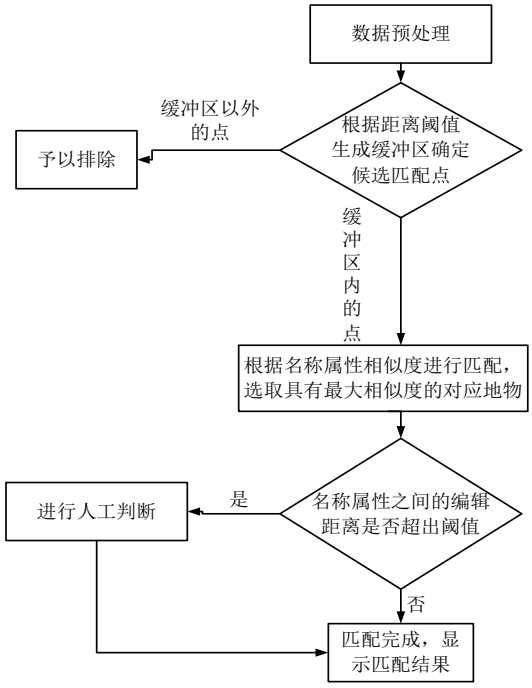

图 1　一致性匹配处理流程

Fig.1　consistency matching process

因此，上下文知识和各自实体所对应的属性项信息就成为一个重要的评判标准和识别依据.例如，假设原有地名为“河南省郑州市科学大道第六十二号”，在地名工作普查中发现其与“河南省郑州市科学大道 62 号”所指是同一地方，因此，对其进行地名匹配并更新.旧地名“河南省郑州市科学大道第六十二号”与新地名“河南省郑州市科学大道 62 号”具有极高的相似性，它们的编辑距离为 5，占总字符数的 1/3.从地图数据中查询新旧地名各自所指的具体地物属性信息时发现它们的备注皆为“信息工程大学”.因此，进一步确认这两个地名确实指的是同一个地方.故此，有必要对其所对应的地理数据进行更新.再如，原有地名数据中，旧地名“河南省郑州市中牟县卢店镇”和“河南省郑州市中牟县大冶镇”在新地名数据中有所缺失，但在新的地理

数据相似的地理坐标范围内出现了“河南省郑州市中牟县大卢镇”.经过匹配分析，发现是由于前些年国家出台扯乡并镇政策时地理名称发生了变化，原有的卢店镇与大冶镇合并为一个新的镇——大卢镇.因此，原有的卢店镇和大冶镇确实与现有的大卢镇在地图数据中指的是同一个地方.

3 结 论

无论是基于语义距离（编辑距离）还是利用BPM-BM[9]算法来计算地名的相似性并进行匹配，其结果都不可能达到100 %的正确率.因此，必须借助其他匹配方法（例如几何相似性匹配[12]、首字符筛选法、后缀词语过滤法等）首先对待匹配地物进行粗匹配过滤，之后再借助计算机利用提前设计好的相似性度量匹配算法进行匹配，以减少工作量并提升匹配的准确率.

由前文的研究分析，可以得出如下结论：

（1）实践证明，将语义相似性技术引入地名匹配确实有必要进行研究；

（2）为了使得地名匹配结果更加接近实际并且提升地名匹配的准确率，必须综合考虑各种因素与技术方法，比如：同时借助属性相似性匹配与几何相似性匹配技术；

（3）实验证明，尽管在对地名进行匹配研究时综合考虑了其他种种因素，但匹配结果仍然达不到完全的准确，距离实际还有一定的差距.因此，要想提高匹配正确率，只能避免错误并且尽可能的减少误差.

无论是几何数据匹配还是属性数据匹配，都是地理数据匹配研究中的一个重要课题，而且其正确率一直都有待于进一步提高.

参考文献：

[1] 李淑霞,安敏,李宏伟,等.常识空间认知研究与地名本体设计[J].测绘科学技术学报,2011,28(6):450-453.
Li Shuxia,An Min,Li Hongwei,etal.Design of the ontology of place based on commonsense spatial cognition[J].Journal of Geomatics Science and Technology,2011,28(6):450-453.

[2] 陈健,李宏伟,张斌,等.基于地名本体的地名演变分析[J].测绘科学技术学报,2011,28(6):446-449.
Chen Jian,Li Hongwei,Zhang Bin,etal.Toponym evolvement analysis based on the toponym ontology[J]. Journal of Geomatics Science and Technology,2011,28(6): 446-449.

[3] 李淑霞.地名本体及其在地理空间数据组织中的应用研究[D].郑州:信息工程大学,2009.
Li Shuxia.Research on Ontology of Place and its Applications in Geospatial Data Organization[D].Zhengzhou: Information Engineering University,2009.

[4] Konstantinos A Nedas,Max J Egenhofer. Spatial-Scene similarity queries[J].Transactions in GIS, 2008,12(6): 661-681.

[5] 魏凯斌,冉延平,余牛.语义相似度的计算方法研究与分析[J].计算机技术与发展,2010,20(7) :102-105.
Wei Kaibin,Ran Yanping,Yu Niu. The research and analysis of computing methodson sem antic similarit[J].Computer Technology and Development,2010,20(7):102-105.

[6] 刘青宝,金燕,邓苏.基于模糊聚类的属性匹配算法[J].模糊系统与数学,2006,20(6):96-102.
Liu Qingbao,Jin Yan,Deng su.Based on the fuzzy clustering attribute matching algorithms[J]. Fuzzy System and Math,2006,20(6):96-102.

[7] 郭岚,赵亚宁,杨永崇.基于界址点的地籍时态数据模型[J].辽宁工程技术大学学报:自然科学版,2011,28(3):367-369.
Guo Lan,Zhao Yaning,Yang Yongchong.Cadastral temporal data model based on boundary points[J].Journal of Liaoning Technical University: Natural Science,2011,28(3):367-369.

[8] 赵彬彬,邓敏.多尺度地图面目标匹配的统一规则研究[J].武汉大学学报:信息科学版,2011,36(8):991-994.
Zhao Binbin,Deng Min.Multi-scale map object matching face of uniform rules research[J].Wuhan University of Technology: Information Science 2011,36(8):991-994.

[9] 范立新.改进的中文近似字符串匹配算法[J].计算机工程与应用,2006(34):172-174.
Fan Lixin. Improved algorithm of approximate Chinese string matching[J]. Computer Engineering and Application, 2006(34):172-174.

[10] 陈开渠,赵洁,彭志威.快速中文字符串模糊弧配算法[J].中文信息学报,2004,18(2):58-65.
Chen Kaiqu,Zhao Jie,Peng Zhiwei. Rapid Chinese string fuzzy arc match algorithm[J]. Journal of Chinese Information Processing, 2004,18(2):58-65.

[11] Myers G.A fast bitvector algorithm for approximate string matching based on dynamic programming[J].Journal of the ACM,1999,46(3):395-415.

[12] 安晓亚.空间数据几何相似性度量理论方法与应用研究[D].郑州:信息工程大学研究生学院,2011.
An xiaoya.Research on theory, methods and applications of geometry similarity measurement for spatial data[D].Zhengzhou: Information Engineering University,2011.