第44卷 第7期
2014年7月

中国海洋大学学报
PERIODICAL OF OCEAN UNIVERSITY OF CHINA

44(7):111~116
July, 2014

空间位置信息的多源POI数据融合*

张 巍, 高新院, 李瑞姗

(中国海洋大学信息科学与工程学院,山东 青岛 266100)

摘 要: 为从2个来源不同的POI数据集合中准确找出用于融合的对应对象,在国外研究成果的基础上提出1种改进方案,该方案在空间位置属性的基础上利用非空间属性相似度来提高结果融合集的准确性。技术路线如下:首先对2个POI数据集合实施空间位置方法找出对应对象组成的初步融合集,然后使用低阈值的名称属性相似度方法排除由空间位置方法找出的错误对应对象,最后使用高阈值的名称属性相似度方法找出空间位置方法未能找出的对应对象。用多组POI数据集合测试改进方案,结果表明融合集的准确率、召回率以及F1值都有明显提高。

关键词: POI数据集合; 对应对象; 准确率; 召回率; F1值

中图法分类号: S623.1 **文献标志码:** A **文章编号:** 1672-5174(2014)07-111-06

POI是兴趣点(Point of Interest)的缩写,是一种代表真实地理实体的点状数据,POI一般包含名称、类别、经纬度以及地址等基本信息。伴随着网络地图与基于位置服务(LBS)的快速发展,POI数据也出现了快速增长,然而,不同来源的POI信息完善和丰富程度不同,如何把已存在的POI数据进行集成融合从而实现数据复用,已成为急需解决的问题。由于POI数据不像结构数据和半结构数据那样具有全局标识,要解决POI数据融合问题就比较困难。国外研究者提供给的解决方案有:基于空间位置方法[1-2];基于非空间属性方法[3-6];基于本体(Ontology)的方法[7]。

基于位置方法的优点是它仅仅根据经纬度位置信息就可以找对应对象,而经纬度信息是每个POI都必须具备的,不存在数据缺失问题;缺点是来源不同的POI的经纬度都普遍存在误差与坐标系不统一的问题。基于非空间属性方法的优点是它只使用非空间特征属性不用考虑经纬度中存在的差异,方法也更为成熟,缺点是它要求不同来源的POI之间必须有比较统一的存储模式,另外,非空间特征属性有可能存在信息缺失与标注错误问题。而基于本体(Ontology)方法的优点是它可以为每个POI对象创建一个类似结构化数据的全局标识符,从而使得融合过程变得非常容易;缺点是目前并没有比较成熟的本体库可以使用。

综上所述,单独使用以上方法都不能取得令人满意的融合结果,本文在以上研究成果的基础上提出一种改进方案,该方案是在空间位置的基础上使用名称特征属性相似度来提高POI融合集的准确性。具体实现过程为:首先对2个POI数据集合实施空间位置方法找出对应对象组成的初步融合集,然后使用名称特征属性排除由空间位置方法找出的融合集中的错误对应对象,最后使用高阈值的名称特征属性相似度方法找出剩余的对应对象。

1 POI数据融合方法及其改进方案

1.1 基于空间位置的方法

基于空间位置的方法仅利用POI的位置信息来寻找正确的融合集,Beeri C等提出了几种基于空间位置的POI数据融合方法,其中最重要的有相互最近邻方法、概率方法、标准化权重方法。假设有2个不同来源的POI数据集合A和B,并且有$A=\{a_1,a_2,\cdots,a_m\}$,$B=\{b_1,b_2,\cdots b_n\}$。

相互最近邻方法:

$$\text{confidence}(\{a,b\}) = 1-\frac{\text{distance}(a,b)}{\min\{\text{distance}(a,b2),\text{distance}(a_2,b)\}},$$

式中:$a\in A$,$b\in B$,$\text{confidence}(\{a,b\})$为POI对应对象$(a,b)$的置信度;$a_2$为$b$在$A$中的次近邻对象;$b_2$为$a$在$B$中的次近邻对象。如果置信度大于给定阈值就认为是正确的对应对象。

概率方法:

$$Pa_i(b_j) = \frac{\text{distance}(a_i,b_j)^{-\alpha}}{\sum_{k=1}^{n}\text{distance}(a_i,b_k)^{-\alpha}},$$

$$\text{confidence}(\{a_i,b_j\}) = \sqrt{Pa_i(b_j)\cdot Pb_j(a_i)}。$$

公式中$Pa_i(b_j)\cdot Pb_j(a_i)$为POI对象a_i,b_j互相选择的概率;α为距离衰减因子。

* 基金项目:山东省自然科学基金项目(ZR2010FM002;ZR2012FM016)资助
收稿日期:2012-10-10; 修订日期:2012-12-25

作者简介:张 巍(1975-),男,副教授。E-mail:ihcil.ouc@gmail.com

标准化权重方法：

$$\mu_{ij}=\begin{cases}P_{a_i}(b_j)\cdot P_{b_j}(a_i) & 1\leqslant i\leqslant m,1\leqslant j\leqslant n\\ \prod_{k=1}^{n}(1-P_{b_k}(a_i)) & 1\leqslant i\leqslant m,j=n+1\\ \prod_{k=1}^{m}(1-P_{a_k}(b_j)) & i=m+1,1\leqslant j\leqslant n\\ 0 & i=m+1,j=n+1。\end{cases}$$

定义一个匹配矩阵 $\boldsymbol{M}$，$\boldsymbol{M}$ 为 $(m+1)\times(n+1)$ 的矩阵，μ_{ij} 为 POI 对象 a_i，b_j 互相选择的概率，每一行的最后一列为 a 不被 b 选中的概率，每一列的最后一行为 b 不被 a 选中的概率，然后通过迭代对矩阵归一化，矩阵内的每一个数值代表该位置对应的 POI 成为对应对象的置信度。

1.2 基于名称相似度的 POI 数据融合

一般情况下，成为对应对象的 POI 都具有相似的名称特征属性，而 2 个名称之间的相似度可以用字符串相似度定量表示。常见的字符串相似算法有 Jaro 相似算法、Jaro-Winkler 相似算法和 Levenstein 编辑距离算法等，本文提出的 POI 数据融合的改进方案就用到了这些算法并取得了不错的结果。

事实上，国外的研究者们已经提出了空间位置属性结合名称、地址等非空间特征属性来提高 POI 数据融合质量的可能性。例如 Sehgal V 等人对地理数据集成中实体识别问题的研究[9]，以及 Safra E 对来自网络电子地图数据的融合研究等[10]。

与文献[9]相比本文提出的改进方案不只是对空间位置做简单的距离求值处理，而是通过测试实验数据后选取表现最佳的标准化权重方法；与文献[10]相比改进方案提出非空间属性数据的处理为相似度而不是简单的对比是否相等。改进方案的核心依据是如果来源不同的 POI 数据集合中的对象具有较近的位置和相似的非空间属性信息，那么它们表示同一个真实地理实体的可能性就增大。

1.3 POI 数据融合的改进方案

本文提出的 POI 数据融合的改进方案要解决的主要问题是：如何从不同来源的 POI 数据集合中准确找出对应对象，这些对应对象进一步形成融合集，最终实现 POI 信息的融合。改进方案的详细过程为：首先对两个 POI 数据集合实施空间位置方法找出对应对象组成的初步融合集，然后使用低阈值的名称特征属性相似度方法排除由空间位置方法找出的错误对应对象，最后使用高阈值的名称特征属性相似度方法找出空间位置方法未能找出的对应对象。方案的详细描述如下：

Input：Databases G and B

Output：A set P of pairs and a set S of singletons

1： (GS, BS, P) ← Φ (G, B)

2： for each {g, b} ∈PL such that η (g. title, b. title) < γ do

P← P− {g , b}

(GS, BS) ← {g, b}

3： for each {g} ∈GS and {b} ∈BS if μ (g. title, b. title) > τ do

P ← {g, b}

GS ← GS − {g}

BS ← BS − {b}

4： S ← GS ∪ BS

5： return (P, S)

(1)对 POI 集合 G 和 B 实施空间位置方法，得到一个初步融合集 P 和 2 个无对应项的单集 GS、BS。

(2)对于融合集 P 中的每个对应项{g,b}，如果对象 g 和 b 的名称属性相似度小于给定阈值 γ，就从融合集 P 中删去{g,b}，然后把对象 g 和 b 分别加入到单集 GS 和 BS 中，η、γ 分别为初次过滤函数和给定阈值。

(3)对于余下的每个单集对象{g}∈GS 和{b}∈BS，如果对象 g 和 b 的名称相似度大于给定的阈值 τ，就把{g,b}添加到融合对集 P 中，然后再把 g、b 分别从单集 GS、BS 中去除，μ、τ 分别为二次过滤函数和给定阈值。

(4)最后，剩余单集对象 GS、BS 的并集作为最终单集元素。算法最终结果为一个融合对集 P 与一个最终单集集合 S。

2 POI 实验数据的获取及预处理

2.1 POI 数据简介

每个 POI 数据包含 4 方面信息，名称、类别、经度、纬度。POI 数据是一种代表真实地理实体的点状数据，它可以代表建筑物、商店甚至是占有一定面积的地理存在。本文中所研究的 POI 数据除了以上给出的属性特征外，还可以有门牌号、邮编、地址、电话号码等更多丰富的属性信息。

2.2 如何从网络电子地图上获取 POI 数据

目前比较流行的网络电子地图有谷歌地图、百度地图、搜狗地图、Mapabc、Mapbar、51 地图等，本文选取谷歌地图和百度地图的 POI 数据作为实验数据。由于来自谷歌地图和百度地图的 POI 数据坐标系统不统一而导致空间位置信息误差非常大，如果不对所获取的 POI 数据进行相应处理其空间坐标信息就不能用来找出融合集，这将在 POI 数据预处理中给出具体的解决办法。

百度地图和谷歌地图都能提供全国各个城市诸如道路、店铺、学校、酒店等信息，这些信息可以用来组织

POI 实验数据。POI 数据抽取过程如下：用百度地图和谷歌地图提供的本地搜索功能对地图实施可视范围内的搜索，对所需 POI 数据的范围逐一移动可视区域进行搜索并保存搜索结果，这里搜索的范围选定为山东省青岛市市南区、市北区、崂山区和李沧区，所涉及的行业有书店、学校、餐饮等服务行业。

2.3 POI 数据预处理

现在流行的网络电子地图的地理坐标除了系统误差外还有坐标系统不统一问题，坐标系统不统一导致的偏差远远大于系统误差，所以首先要解决的问题就是统一坐标系统。统一网络电子地图地理坐标系统一般有 3 种方法：建立并查找偏移数据库得出相应坐标的偏移数据；获取电子地图间坐标转换接口；拟合坐标转换公式。

本文采用的方法是获取电子地图间坐标转换接口，大多数网络电子地图服务商都有开放的应用程序编程接口(API)，并提供地图间坐标转换接口，这就为实施不同地图间 POI 数据地理坐标转换统一提供了可能。本文是把来自谷歌地图 POI 数据坐标统一到百度地图坐标系统下。坐标统一后，对从两家地图上抽取的 POI 数据集合经过人工判定，找出对应对象。判定的标准为先观察名称再观察地址等，确认为对应对象后再人工分配 ID 号，作为最终实验数据。

3　寻找最佳阈值

通过比较空间位置方法和名称属性相似度方法得出的结果，选择最佳空间位置算法和名称属性相似度算法及最佳阈值参数。

3.1　实验数据集

数据集 G 表示来自谷歌地图的 POI 数据集合，B 表示来自百度地图的 POI 数据集合，其中每个对象都有名称、经纬度坐标以及地址等特征信息。两实验数据集合各有 500 个对象，其总共代表地理实体 724 个，其中反例 448 个，正例 275 个，正例为 2 个 POI 集合中有对应项的对象，反例为无对应项的对象，2 个 POI 数据集中的每个对象都唯一地代表一个真实的地理实体。

表 1　来自谷歌地图和百度地图的 POI 测试数据

Table 1　The POI test data from Google map and Baidu map

数据集来源 Dataset source	实体数 Entities	正例 Positive instance	反例 Negative instance
谷歌地图	500	275	225
百度地图	500	275	225

3.2 结果质量评价标准

本文用国际上比较权威且通用的准确率、召回率和 F1 值作为衡量算法结果质量的评价标准。由于在数据预处理过程中，2 个 POI 数据集合中的对应对象分配了相同的 ID 号，这样就预先知道了来源不同的 2 个 POI 是否代表同一个地理实体，进而就可以确定融合集中的对应对象是否正确。这些先验知识可以用来定义评价标准，并求出评价标准的具体值。

准确率是指结果融合集中正例所占的比例：

$$\text{准确率}=\frac{\text{融合集中正例数}}{\text{融合集对象总数}}。$$

召回率定义为结果融合集中正例数占实验数据集中总正例数的比例：

$$\text{召回率}=\frac{\text{融合集中正例数}}{\text{实验数据集正例总数}}。$$

F1 值为召回率和准确率的调和平均数，定义如下

$$\text{F1}=\frac{2\times\text{召回率}\times\text{准确率}}{\text{召回率}+\text{准确率}}。$$

对于特定算法，阈值降低，召回率增大，准确率减小；反之召回率减小，准确率增大。一般情况下，准确率和召回率之间是种反向关系，最佳阈值可以使算法得到较高的召回率和准确率。

3.3 空间位置方法的最佳阈值

用空间位置方法测试 POI 实验数据，在其过程中算法所涉及的参数有误差上限、距离衰减因子和给定阈值，前 2 个参数依次为 100 和 2，具体过程参见文献[1-2,9]，阈值采用 0.05 为起始值然后依次增大 0.1 直到 0.85 观察其准确率、召回率和 F1 值。根据实验结果最终选取的最佳阈值如下：

表 2　空间位置方法的最佳阈值

Table 2　The optimal threshold value of the location-based algorithms

位置方法 Location-based algorithms	选定阈值 Selected threshold
相互最近方法(NN) Nearest Neighbor	0.0
概率方法(PRO) Porbit Method	0.57
标准化权重方法(NW) Normal Weight	0.57

3.4 名称属性相似度算法最佳阈值

用名称相似度方法测试 POI 实验数据，主要目的是为名称相似度方法找出最佳阈值。测试过程与空间位置方法类似，阈值采用 0.1 为起始值然后依次增大 0.1 直到 1 观察其准确率、召回率和 F1 值，根据实验结果找出的最佳阈值如下：

表 3 名称属性相似度方法的最佳阈值

Table 3 The optimal threshold value of the title-similarity algorithms

名称相似度方法 Title-similatity algorithms	选定阈值 Selected threshold
getJaro	0.8
getJaroWinkler	0.86
getLevenshtein	0.64

3.5 单组 POI 数据集合测试改进方案

用已找出的最佳阈值对各个方法进行测试，通过比较各自测试结果，选择表现最佳的方法用于改进方案，实验结果对比如下：

从图 1 可以看出，名称相似度方法的表现优于空间位置方法，而名称相似度表现最好的是 getJaroWinkler 方法，空间位置方法中表现最好的则是标准化权重方法。所以本文提出的改进方案选择标准化权重方法和 getJaroWinkler 方法。

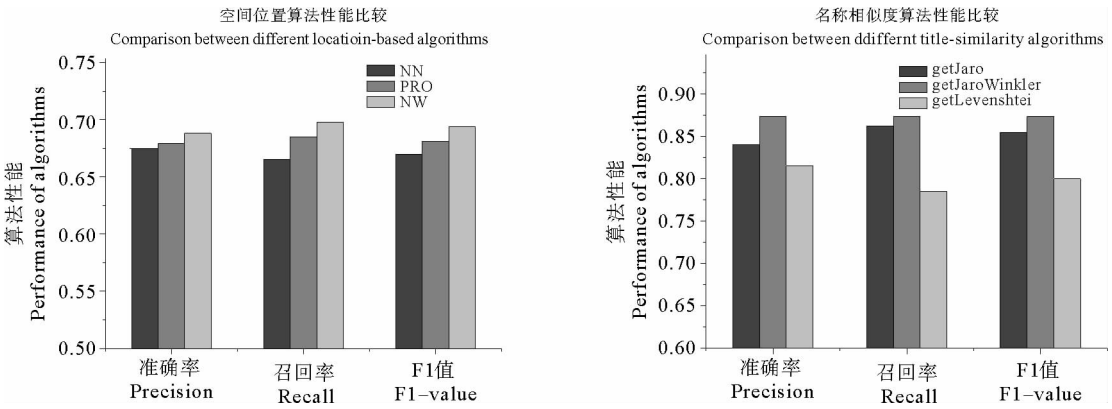

图 1 空间位置算法与名称相似度算法性能比较

Fig. 1 The comparison of the performance of location-based algorithms and title-similarity algorithms

改进方案的具体过程为：首先用空间位置方法最大限度找出 2 个 POI 数据集中所有可能的对应对象，也就是尽可能地增大召回率，这就要求空间位置方法的阈值不能太高，经过多次试验得出将阈值设置 0.4 为最佳；然后对找出的融合集进行名称相似度的首次过滤，在这个过程中，要尽可能保留名称属性相似的对象，因为距离相近且名称相似的 POI 成为对应对象的可能性较大，经过测试，首次过滤名称相似度方法选择 getLevenshtein 方法最佳，其最佳阈值为 0.3 左右；然后对剩下的 POI 集合做名称相似度的二次过滤，由于没有了距离约束其方法选择要严格且阈值要高，本文选取 getJaroWinkler 方法，阈值设置为 0.9。改进后的方法用 COM-NWT 表示。

表 4 不同方法之间比较

Table 4 The comparison of different methods

方法 Methods	准确率 Precision	召回率 Recall	F1 值 F1-value
NW	0.68	0.7	0.69
GetJaroWinkler	0.87	0.87	0.87
COM-NWT	0.94	0.94	0.94

上表可以看出本文提出的改进方案表现明显优于上面 2 种方法，这是仅用单组 POI 数据集合测试得出的实验结果，接下来将给出多组 POI 数据集合的测试结果来验证该方案的有效性。

4 实验结果

4.1 组织多组 POI 数据集合

来源于谷歌地图和百度地图的每个 POI 测试集合各有 450 个对象，每个对象都唯一地代表一个地理实体，所有对象没有重复项。总共有 7 组 POI 数据集合参与测试，这些 POI 集合具有不同的重合度，实验数据集合组织如下：

表 5 多组 POI 测试数据集合

Table 5 The multiple test sets of POI data collection

实体总数 Enfities	正例数 Posifive instance	反例数 Negative instance	正例比例 Positive instance of ratio
800	100	700	0.2
750	150	600	0.3
700	200	500	0.4
650	250	400	0.5
600	300	300	0.6
550	350	200	0.7
500	400	100	0.8

4.2 多个数据集测试改进方法

用标准化权重方法、名称相似度方法和本文给出的改进方案对上面多组 POI 实验集合进行测试，用准确率、召回率和 F1 值衡量算法测试结果。其结果对比图示如下：

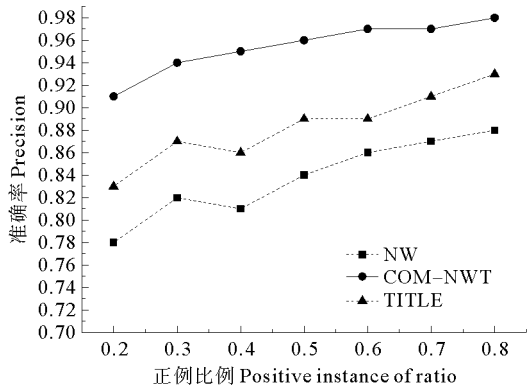

图 2　多组 POI 数据集合准确率比较

Fig. 2　The comparison of the precision of multiple sets of POI

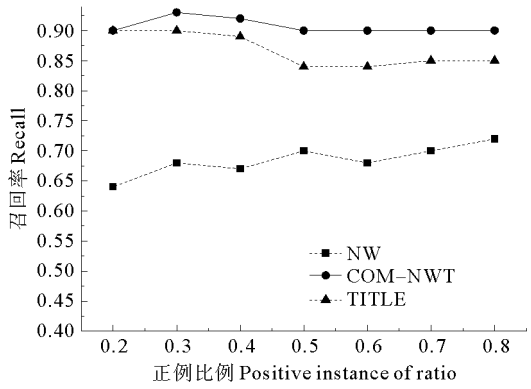

图 3　多组 POI 数据集合召回率比较

Fig. 3　The comparison of the recall of multiple sets of POI data collection

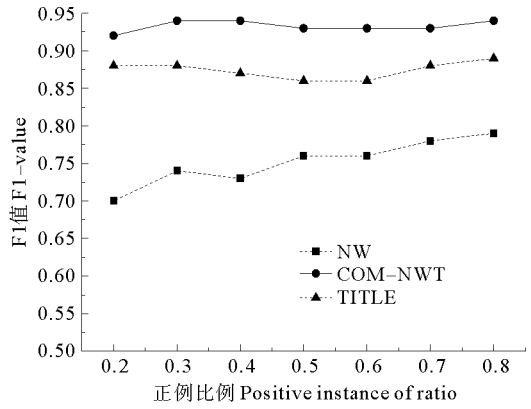

图 4　多组 POI 数据集合 F1 值比较

Fig. 4　The comparison of the F1-value of multiple sets of POI

图 2～4 中 NW 表示仅用空间位置的标准化权重方法，TITLE 表示仅用名称属性相似度的 getJaroWinkler 方法，COM-NWT 为本文提出的改进方案，图 2～4 中结果依次为多组 POI 数据集合测试的准确率、召回率和 F1 值。从图 2～4 可以看出，改进方案的性能表现明显优于单独使用空间位置方法和名称属性相似度方法。本文所使用的 POI 实验数据来自密度大、范围小的繁华区域，加上地图间坐标误差比较大等因素导致了单独使用空间位置方法性能表现弱于单独使用名称属性相似度方法。如果用大范围的 POI 数据，情况就会好些，但这不会影响本文提出的改进方案的性能。另外，由于实验数据和现实中具体的数据特性有所不同，改进方案中的某些参数可能要做相应调整以达到最好效果。

5　结语

为从不同来源的 POI 数据集合中准确找出用于融合的对应对象，本文提出了在空间位置的基础上引入名称特征属性相似度的解决方案，该方案对空间位置方法找出的对应对象使用名称属性相似度进行二次处理，这就实现了在空间位置属性的基础上结合其他非空间属性的多源 POI 数据融合的设想。然后分别用单组 POI 数据集合和多组 POI 数据集合对空间位置方法、名称特征属性相似度方法和本文提出的解决方案作对比试验，实验结果表明，本文提出的解决方案找出的 POI 对应对象的准确性明显优于单独使用空间位置方法和名称特征属性相似度方法。本文除了 POI 的空间位置属性仅使用了名称特征属性，如果结合名称以外的其他非空间属性信息（比如地址、邮编等），POI 对应对象的准确性还会有提升空间。

参考文献：

[1] Beer C, Kanaza Y, Safra E, et al. Object fusion in geographic information system[C]. Toronto: Proceedings of the 30th VLDB Conference, 2004: 816-827.

[2] Beer C, Doytsher Y, Kanaza Y, et al. Finding Corresponding Objects when Integrating Several Geo-spatial Datasets[C]. Bremen: Proceedings of the 13th annual ACM international workshop on Geographio information systems, 2005:87-96.

[3] Samal A, Seth S, Cueto K. A feature based approach to conflation of geospatial sources [J]. International Journal of Geographical Information Science, 2004, 18(5): 459-489.

[4] Sester M, Anders S K H, Walter V. Linking objects of different spatial data sets by integration and aggregation [J]. GeoInformatica, 1998, 2(4):335-338.

[5] Wiederhold G. Mediators in the architecture of features information system [J]. Computer, 1992, 25(3):38-49.

[6] Water V, Fritsch D. Matching spatial data sets: a statistical approach [J]. International Journal of Geo-graphical Information Science, 1999, 13(5): 445-473.

[7] Fonseca F T, Egenhofer M J, Agouris P. Using ontologies for integrated geographic information system [J]. Transaction on GIS, 2002, 6(3): 231-257.

[8] Sehgal V, Getoor L, Viechniki P D. Entity Resolution in Geospatial data Integration [C]. New York: ACM Press, 2006: 83-90.

[9] Safra E, Kanaza Y, Sagiv Y, et al. Integrating Data from Maps on the World-Wide Web [C]. Lausanne: Web and Wireless Geographical information Systems Lecture Notes in Computer Science Volume 4295, 2006:180-191.

[10] 彭煜玮，彭智勇. 空间数据融合技术的研究[J]. 计算机工程，2007, 33(18): 51-52.

Multi-Source POI Data Fusion Based on the Spatial Location Information

ZHANG Wei, GAO Xin-Yuan, LI Rui-Shan

(College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

Abstract: To find out the corresponding objects from distinct original POI datasets, we propose a novel approach on the bases of the abroad researching results. This approach increases the accuracy of the fusion set by using non-spatial properties similarity based on the spatial location. Firstly, apply location-based algorithms to two POI datasets to find out initial fusion set consisting of corresponding objects, and then use the methods based on name information with low threshold to exclude wrong corresponding objects obtained using location-based method. At last, the remain corresponding objects not found will be searched out by using the methods based on name information with high threshold. This modified technique was tested on the different POI datasets. It has been demonstrated by our experiments that the precision, recall and F1-value of the fusion set was improved to a large part.

Key words: POI datasets; corresponding objects; precision; recall; F1 values

责任编辑 陈呈超