

文章编号:1001-9081(2018)05-1334-05

DOI:10.11772/j.issn.1001-9081.2017102504

基于距离类别的多源兴趣点融合算法

徐爽¹, 张谦², 李琰¹, 刘嘉勇^{1*}

(1. 四川大学电子信息学院, 成都 610042; 2. 中国电子科技集团第二十九研究所, 成都 610036)

(*通信作者电子邮箱 ljjy@scu.edu.cn)

摘要:为了更好地实现多源兴趣点(POI)数据的有效集成与精确融合,提出了一种结合空间与非空间属性的距离类别的兴趣点融合算法(MNMDC)。首先,对空间属性,通过标准化权重算法计算待融合对象的空间相似度得到融合集;其次,利用非空间Jaro-Winkle算法对融合集中类别一致的对象使用低阈值排除,对类别不一致的使用高阈值排除;最后,使用距离约束、类别一致约束和高阈值的非空间Jaro-Winkle算法找出空间算法遗漏的可融合对象。实验结果表明,该方法平均准确率达到93.3%,与空间和非空间算法(COM-NWT)及格网化纠正方法相比,在7组不同重合度的数据下MNMDC方法的平均准确率提高2.7和1.6个百分点、平均召回率提高2.3和1.4个百分点。MNMDC在实际融合过程中能更精确地融合POI数据。

关键词:兴趣点;数据融合;空间属性;非空间属性;距离;类别

中图分类号:TP181 **文献标志码:**A

Multi-source point of interest fusion algorithm based on distance and category

XU Shuang¹, ZHANG Qian², LI Yan¹, LIU Jiayong^{1*}

(1. College of Electronics and Information, Sichuan University, Chengdu Sichuan 610042, China;

2. Southwest China Research Institute of Electronic Equipment, Chengdu Sichuan 610036, China)

Abstract: In order to achieve effective integration and accurate fusion of multi-source Point of Interest (POI) data, a Mutually-Nearest Method considering Distance and Category (MNMDC) was proposed. Firstly, for spatial attributes, standardized weight algorithm was used to calculate the spatial similarity of the object to be fused, and the fusion set was obtained. Secondly, for non-spatial attributes, Jaro-Winkle algorithm was used to eliminate some objects with consistent categories by a low threshold, and exclude some objects with inconsistent categories by a high threshold. Finally, non-spatial Jaro-Winkle algorithm with distance constraint, category consistency constraint and high threshold was used to find out the missing objects in the spatial algorithm. The experimental results show that the average accuracy reaches 93.3%, compared with Combined Normal Weight and Title-similarity algorithm (COM-NWT) and the grid correction methods, the accuracy of MNMDC method in seven different groups of coincidence degree data, the average accuracy increases by 2.7 percentage points and 1.6 percentage points, the average recall increases by 2.3 and 1.4 percentage points. The MNMDC method allows more accurate fusion of POI data during actual fusion.

Key words: Point of Interest (POI); data fusion; spatial attribute; non-spatial attribute; distance; category

0 引言

兴趣点(Point of Interest, POI)是一种代表真实地理实体的点状数据,通常包含多种信息:名称、类别、经纬度等,它可以代表人们感兴趣的地理实体,如商店、景点等。近年来基于位置的社交网络(Location-Based Social Network, LBSN),如Foursquare、Gowalla、Facebook Places等发展迅速^[1],基于LBSN的POI应用的需求和POI数据的规模不断增加^[2-3]。来源于不同网站的POI信息存在位置信息、地址描述及分类属性等方面的一致,如何实现多源POI的有效集成和深度融合,成为空间信息技术面临的一大挑战^[4]。

国内外对POI融合方法提出了三种方案:基于空间位置的方法^[4-6]、基于非空间属性的方法^[7-9]和基于本体的方法

法^[10]。基于空间位置的方法是一种较为简单实用的方法,仅根据经纬度位置信息就可以找到对应对象,但来源不同的POI数据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是仅利用非空间属性信息相似度来寻找融合集,该方法不需要考虑经纬度误差,但要求不同来源的POI之间必须有比较统一的存储模式而且非空间特征属性有可能存在信息缺失与标注错误问题;基于本体的方法可以为每个POI对象创建一个类似结构化数据的全局标识符,从而使融合过程变得非常容易,但目前并没有比较成熟的本体库可以使用,因此不考虑基于本体的方法。

单独使用以上方法都不能取得令人满意的结果,文献[11-12]提出了一种空间位置和非空间属性相结合的方法。该算法准确率和召回率优于单独使用空间和非空间算法。但

收稿日期:2017-10-23;修回日期:2018-01-08;录用日期:2018-01-16。

作者简介:徐爽(1993—),女,山东济宁人,硕士研究生,主要研究方向:机器学习、数据挖掘、大数据可视化;张谦(1990—),男,贵州遵义人,博士,主要研究方向:机器学习、数据挖掘、自然语言处理;李琰(1993—),女,贵州贵阳人,硕士研究生,主要研究方向:机器学习、数据挖掘;刘嘉勇(1962—),男,四川成都人,教授,博士,主要研究方向:信息安全、网络信息处理。

该算法的空间算法只使用经纬度属性,非空间算法只使用名称特征属性,融合结果的质量还有提升空间。

为了解决 POI 融合问题,提高融合结果质量,本文提出了一种基于距离和类别约束的 POI 融合算法。首先通过空间算法初步筛选出融合集;然后利用非空间算法对融合集中类别一致的对象使用低阈值排除,对类别不一致的使用高阈值排除;最后使用距离约束,类别一致约束和非空间算法高阈值找出空间算法遗漏的可融合对象。经实验验证,本文算法的准确率、召回率等各项指标较之文献[11]有明显的提升。

1 POI 融合基本方法

定义 1 数据融合。通过对多个数据源里的信息进行校正与整合,得到一个全面的信息,这个信息比任何一个单一数据源提供的信息都多^[13]。

定义 2 可融合 POI 对象。假设有两个待融合的 POI 数据集合分别为 $A = \{a_1, a_2, \dots, a_n\}$ 和 $B = \{b_1, b_2, \dots, b_n\}$ 。记 $a_b \in A, b_a \in B$ 。如果 a_b 和 b_a 实际上是一个 POI 地理实体,就是可融合 POI 对象。

定义 3 POI 融合。来源不同的 POI 数据通过 POI 融合技术生成信息量更为丰富与完整的 POI 数据,从而实现了 POI 信息的复用与更新,这样就可以节约大量的人力、物力,进而降低 POI 数据更新成本^[13]。

1.1 空间属性算法

空间属性算法是仅利用 POI 的空间位置信息来寻找可融合 POI 对象的算法。

最简单的空间属性算法是通过经纬度计算两个 POI 之间的球面空间距离,距离越小,两点可融合可能性越大。假设点 P_1 的经纬度为 (Lon_1, Lat_1) , P_2 的经纬度为 (Lon_2, Lat_2) , R 为地球的平均半径,根据三角推导可以得到 P_1 和 P_2 之间球面距离的计算公式^[14]:

$$dis(P_1, P_2) = R * \text{Arccos}(C) * \pi / 180 \quad (1)$$

其中: $C = \sin(Lat_1) * \sin(Lat_2) * \cos(Lon_1 - Lon_2) + \cos(Lat_1) * \cos(Lat_2)$, 设地球半径 $R = 6371.004$ km。

除了这种简单的空间算法,Beeri 等^[6]总结了几种地理位置融合算法:片面最邻近、相互最邻近、概率方法和归一化权重算法。

相互最邻近(Mutually-Nearest, MN)算法对两个 POI 数据集合的重合度不敏感^[12],更适合复杂的实际环境,因此本文空间算法选择相互最邻近算法。

假设两个待融合的 POI 数据集合 $A = \{a_1, a_2, \dots, a_n\}$ 和 $B = \{b_1, b_2, \dots, b_n\}$ 。记 $a_b \in A$ 为 b 在数据集 A 中的最近邻对象, $b_a \in B$ 为 a 在数据集 B 中的最近邻对象,那么 a 和 b 成为对应对象的置信度为:

$$\begin{aligned} confidence(a, b) &= 1 - \\ &distance(a, b) / \min(distance(a, b_a), distance(a_b, b)) \end{aligned} \quad (2)$$

其中 $distance(a, b) = [(Lon_a - Lon_b)^2 + (Lat_a - Lat_b)^2]^{1/2}$ 。

如果置信度 $confidence(a, b)$ 大于一定阈值则可以认为 a 和 b 是彼此数据集中的相互最近邻,可以把 a, b 标为可融合 POI 对象。

文献[4]提出了一种基于格网化纠正的多源 POI 位置信息一致性处理方法,通过空间上划分地理格网,对各个地理格

网单元实现局部一致化处理,来进行多源 POI 融合。

1.2 非空间属性算法

非空间属性算法是利用 POI 的非空间属性信息之间的相似度来寻找可融合 POI 对象的算法。

POI 信息中的非空间属性是名称、地址、类别等字符串文本,不同来源的可融合 POI 对象通常都具有极高相似性,可以利用文本相似性算法来分辨。Jaro-Winkler 距离是一种计算字符串之间相似度的方法^[15],它是 Jaro 距离的一个扩展。假设有两个字符串 S_1 和 S_2 ,那么它们之间的 Jaro 距离可以定义为:

$$d_{jaro}(S_1, S_2) = \frac{1}{3} \left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m} \right) \quad (3)$$

式(3)中: m 是匹配的字符数, t 是换位的数目。

Jaro-Winkler 距离:

$$d_{jaro-winkler}(S_1, S_2) = d_{jaro}(S_1, S_2) + (\ell p(1 - d_{jaro}(S_1, S_2))) \quad (4)$$

其中: $d_{jaro}(S_1, S_2)$ 是 S_1 和 S_2 的 Jaro 距离, ℓ 是前缀匹配的长度, p 是前缀匹配的权重。

Jaro-Winkler 距离结果是一个大于 0 小于 1 的数,越接近 1,文本越相似。如果 a 和 b 非空间属性信息的 Jaro-Winkler 距离大于一定阈值则可以认为 a 和 b 是可融合对象。

1.3 结合空间和非空间的算法

文献[11]提出了一种空间和非空间相结合的算法,可以找出不同来源数据集合中可融合的 POI 对象。其思路是初步筛选阶段使用空间算法找出初步融合集,排除阶段用低阈值的非空间算法计算排除初步融合集中错误的对应对象到单集中,补充阶段使用高阈值的名称相似性方法寻找单集遗漏的对应对象并添加到融合集。但是这种算法有不准确的地方,它并未考虑到同名分店的情况,在补充阶段会将其误认为是融合对象;也没有考虑到相近位置相似店名的不同类型 POI 点,在排除阶段无法将其识别,因此其融合结果在这些对象上表现不好。

2 基于距离类别的 POI 融合算法

为了更精确地找出可融合对象,本文提出了一种基于距离类别的 POI 融合算法。

文献[11]提出的传统的 POI 融合算法初步筛选阶段采用了文献[6]中的标准化权重算法计算待融合对象的空间相似度。排除阶段和补充阶段使用 Jaro-Winkler 距离计算待融合对象的名称相似度。本文提出的改进算法中在初步筛选阶段使用了对重合度不敏感、更适合真实环境的相互最邻近算法^[12],排除阶段除了使用 Jaro-Winkler 距离计算待融合对象的名称相似度,还对融合集加入类别判断。补充阶段除了使用 Jaro-Winkler 算法还引用了式(1)的球面距离约束,防止误判。

算法流程如图 1。算法有三个步骤。

第一步 初步筛选阶段。对预处理后的数据集 A, B 使用相互最邻近(MN)算法,将置信度大于阈值 λ 的对应对象放入融合集 ab ,低于阈值的放入单集 AA, BB 。

第二步 排除阶段。对初步筛选后的融合集 ab 的对应数据 $AmBm$ 使用 Jaro-Winkler 算法计算名称相似度。将类别一致时名称相似度低于低阈值 γ 和类别不同时名称相似度低于高阈值 δ 的对应对象排除到单集。

第三步 补充阶段。对初始单集数据 AA, BB 进行使用

Jaro-Winkler 算法和球面距离算法, 将其中名称相似度高于阈值 τ 且类别一致、空间距离低于距离阈值 φ 的对象补充到融合集内, 得到单集和融合集。

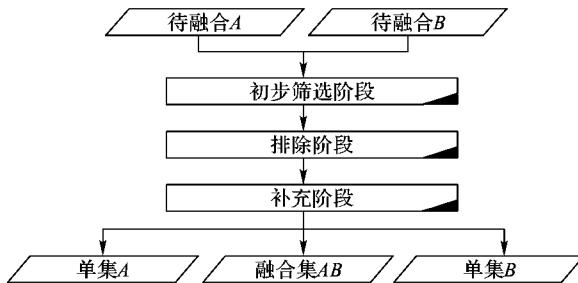


Fig. 1 Flow chart of fusion algorithm

算法伪代码如下:

```

输入: 数据集 A 和 B
1) for each  $A_i \in A, B_j \in B$ 
    if ( $conf(A_i, B_j) > \lambda$ ) then
        do:  $A_i, B_j \rightarrow ab$ 
    else
        do:  $A_i \rightarrow AA, B_j \rightarrow BB$ 
    endif
2) for each  $A_m B_m \in ab$ 
    if ( $category(A_m) = category(B_m)$  and
         $d_{jaro-winkle}(A_m, B_m) < \gamma$ ) then
        do:  $A_m \rightarrow A, B_m \rightarrow B$ 
    elseif ( $category(A_m) \neq category(B_m)$  and
             $d_{jaro-winkle}(A_m, B_m) < \delta$ ) then
        do:  $A_m \rightarrow A, B_m \rightarrow B$ 
    else
        do:  $(A_m, B_m) \rightarrow AB$ 
    endif
3) for each  $A_i \in AA, B_j \in BB$ 
    if ( $d_{jaro-winkle}(A_i, B_j) > \tau$  and  $category(A_i) = category(B_j)$ 
        and  $distance(A_i, B_j) < \varphi$ ) then
        do:  $(A_i, B_j) \rightarrow AB$ 
    else
        do:  $A_i \rightarrow A, B_j \rightarrow B$ 
输出单集 A, B 和融合集 AB

```

伪代码中 $conf(A_i, B_j)$ 是式(2)计算得到的置信度, 其中 $category(A), category(B)$ 是 a 的类别, $d_{jaro-winkle}(A_i, B_j)$ 是使用式(3)和式(4)计算的 Jaro-Winkler 距离, $distance(A, B)$ 是由式(1)计算得来的地球直线距离。

本文针对文献[11]在排除阶段相似名称相近位置 POI 数据误判问题, 对初步筛选阶段后的融合集内位置相近的数据, 加入类别判断。考虑到分类结果不可能完全准确, 单纯将类别不同的数据排除到单集会造成误判, 结合名称相似性方法, 将类别不同但名称相似度低于高阈值 δ 和类别不同名称相似度低于稍低阈值 γ 的数据对排除到单集。这种改动会降低在排除阶段相似名称、相近位置、不同类别 POI 数据误判。

在补充阶段使用严格的距离、类别和名称相似度约束, 在补充初步筛选阶段遗漏的可融合对象的同时, 可以减少同名异地 POI 数据被错误补充进融合集的可能性。

3 实验与结果

3.1 数据采集与预处理

本文使用爬虫在百度地图、谷歌地图采集眉山市的 POI

数据, 然后对数据进行清洗、去重等预处理操作, 将数据的坐标统一转换到百度坐标, 并将 POI 数据中类别缺失的数据按照谷歌分类体系^[12]进行分类。

处理后的数据约 3 万条, 每一条 POI 数据代表一个真实的地体实体, 它由 7 个字段组成, 分别是 ID、名称、地址, 经度、纬度、类别和 POI 编号。ID 是 POI 标识, 名称字段表征 POI 的名字, 类别字段表示 POI 所属类别, 经度、纬度可用来标识 POI 的地理位置, 地址表示 POI 所在的位置(街道、门牌号等), POI 编号是人工标注的, 值是它对应另一个来源的可融合 POI 对象的 ID, 如果没有为空。

在预处理后 POI 数据集中随机抽取 70% 数据作为训练集, 30% 作为测试集。在融合实验中, 为了验证算法是否容易导致过拟合现象, 而且真实环境中, 不同数据源之间的重合度不是一个确定值, 因此本文在测试集中不同数据源分别随机抽取了重合度不同的 7 组数据作为测试数据, 如表 1 所示。

表 1 不同重合度测试数据集

Tab. 1 Different coincidence test data sets

序号	实体数	正例数	负例总数	正例比例
1	5 000	1 000	4 000	0.2
2	5 000	1 500	3 500	0.3
3	5 000	2 000	3 000	0.4
4	5 000	2 500	2 500	0.5
5	5 000	3 000	2 000	0.6
6	5 000	2 500	1 500	0.7
7	5 000	4 000	1 000	0.8

其中: 正例是在数据集中有对应 POI 的数据中随机取样得到的, 负例是随机抽取的相应数量的没有可融合 POI 对象的实体, 正例比例是正例数占实体总数的比例, 实体数代表两个测试集中 POI 的数量。

3.2 融合结果质量评价指标

本文实验部分采用国际上权威且通用的准确率(Precision)、召回率(Recall)和 F1 值作为衡量 POI 融合结果质量的评价指标。

准确率是指结果融合集中正例占融合集比例, 它表示算法计算出来的融合对象是真正可融合的对象的概率:

$$precision = \text{融合集中正例数} / \text{融合集总数} \quad (5)$$

召回率是指结果融合集中正例占样本实际正例的比例, 它表示的是样本中的可融合对象有多少被算法正确融合了:

$$recall = \text{融合集中正例数} / \text{实验数据集正例总数} \quad (6)$$

实际应用时, 需要平衡准确率和召回率, 使用 F1 值作为算法评价指标。计算方法如下:

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (7)$$

3.3 阈值选取

补充阶段的距离使用球面距离式(1)来计算训练集样本中随机选择的 200 个已经标记的融合对象的距离, 结果如图 2 所示, 并对每对可融合 POI 计算距离后按距离排序。

由图 2 可以看出 90% 以上的融合对象都集中在 30 m 以内, 为了提高准确率防止误判选取 30 m 作为距离阈值 φ 。

本文中初步筛选阶段只使用相互最邻近算法, 根据文献[11]选定阈值 λ 为 0。

三个阶段中使用的算法有空间算法:相互最近邻算法和球面距离,非空间算法:Jaro-Winkler 算法。其中空间算法阈值都已经确定,Jaro-Winkler 算法因为涉及到排除阶段和补充阶段两个阶段,有三个阈值,因此更加难以确定阈值。

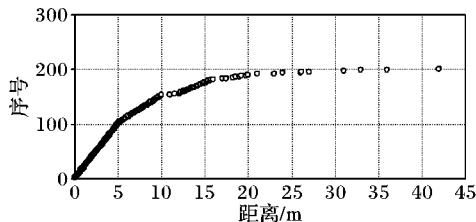


图 2 可融合对象的距离分布

Fig. 2 Distance distribution of object to be fused

排除阶段和补充阶段使用的 Jaro-Winkler 算法有三个阈值 γ, δ, τ 待定。排除阶段 Jaro-Winkler 算法的两个阈值初始值 γ, δ 之间是相互影响的,无法单独确定。文献[11] Jaro-Winkler 算法选定阈值为 0.86,这个阈值是不在类别影响下得到的,因此根据控制变量法原理本文设类别不同时采用的阈值 δ 等于常量 0.86。在训练集上做实验调试阈值,设 γ 为 0.1 并依次增加直至 1,由式(7)计算排除阶段的 F1 值。此时可以得到 γ 的最佳阈值:F1 为最大值时的值。将 γ 设为最佳阈值,设 δ 为 0.1 并依次增加直至 1,计算排除阶段的 F1 值。记 F1 为最大值时的 δ 为最佳阈值 δ 。

此时已经得到最佳阈值 $\gamma, \delta, \lambda, \varphi$ 。将其他算法阈值设置为最佳阈值,补充阶段的 Jaro-Winkler 算法的 τ 设为 0.1 并依次增加直至 1。综合计算三个阶段算法得到融合结果 F1 值,观察结果可得最佳阈值 τ 。

测试结果如图 3。

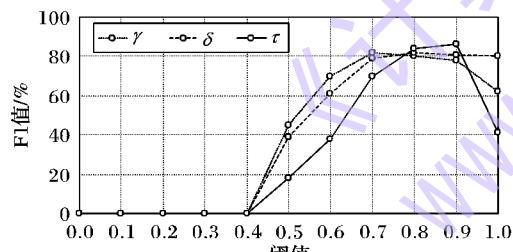


图 3 阈值测试结果

Fig. 3 Results of threshold testing

根据测试结果,Jaro-Winkler 距离方法在 POI 融合过程中各个阶段参数的阈值 γ, δ, τ 分别为 0.7, 0.8 和 0.9。

3.4 融合实验

对本文方案和文献[11]、文献[4]进行 POI 融合对比实验,使用式(5)、式(6)、式(7)评估这些方案在本文测试集实验数据上的性能,方案详情如表 2。

表 2 POI 融合方案所用算法和属性对比

Tab. 2 Algorithms and attributes used in different POI fusion schemes

方案名称	空间算法	非空间算法	涉及属性
空间和非空间算法 ^[11]	标准化权重	Jaro-Winkler	名称, 经纬度
格网化纠正 ^[4]	地理格网	无	经纬度
距离类别的兴趣点融合算法(本文)	相互最近邻、球面距离	Jaro-Winkler	名称,类别, 经纬度

其中,空间和非空间算法(Combined Normal Weight and Title-similarity algorithm, COM-NWT)是文献[11]的方案,格

网化纠正(简称为“网格化”)是文献[4]的方案,距离类别的兴趣点融合算法(Mutually-Nearest Method considering Distance and Category, MNMDC)是本文提出的距离、类别改进的方案。测试融合方案在不同重合度下的性能,融合结果如图 4 所示。

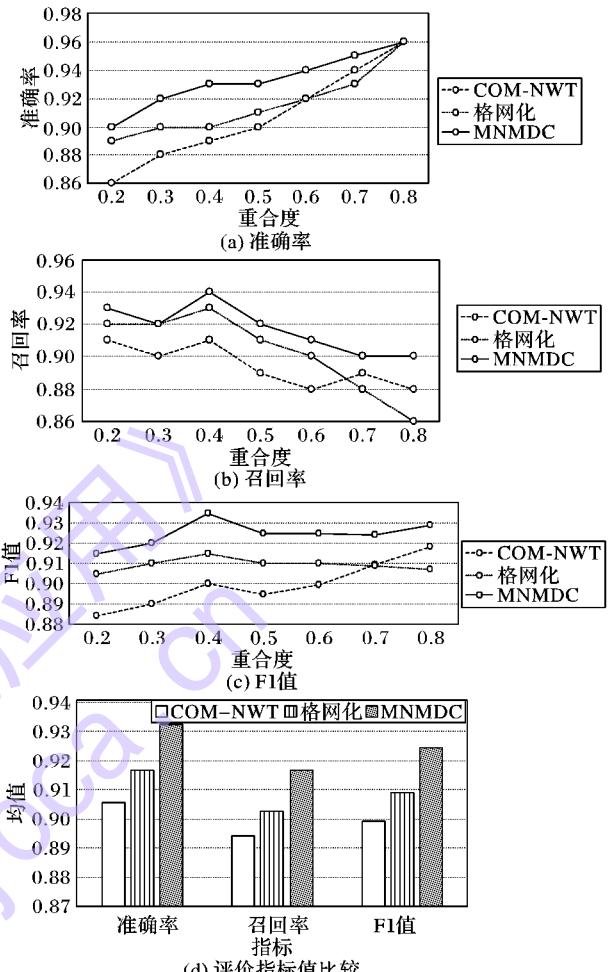


图 4 三种 POI 融合方案准确率、召回率、F1 值和均值对比

Fig. 4 Precision, recall, F1 and average comparison of three POI fusion schemes

从实验结果中可以看出,当数据集中的正例较少时,MNMDC 方法因为引入了距离的判断,能够比 COM-NWT 方法更有效地排除数据集中的负例,并且 MNMDC 方法通过对类别的判断,进一步提升了融合过程中的准确率,从而获得了明显优于 COM-NWT 方法的和格网化方法的融合效果。

随着正例比的增加,数据集中存在类别缺失的数量逐渐增多,由于这些 POI 的类别来自人工分类,POI 分类过程产生的误差会导致融合时类别判断的错误,再加上距离判断中为了提高准确度而选取的严格的距离阈值,会使一些可融合的对应对象被划分到单集,所以 MNMDC 方法的召回率出现了明显的下降,而 COM-NWT 方法不受分类的影响,因此表现较为平稳。但是正例比的增加,造成 POI 密度的增加,格网化方法出现误判,召回率下降。在三个方案中,本文提出的 MNMDC 方案召回率仍最高。在多组测试集中进行测试,实验结果均表现良好且相差不大,采用公开数据集中百度和谷歌的 POI 数据做对比实验,准确率为 91%,召回率为 90.5%,和在自行爬取数据集中相差不大,对比实验证明本文采用的算法具有普适性且不易产生过拟合现象。

4 结语

在数据时代,伴随着网络电子地图与 LBSN 的快速发展,POI 数据的需求与日俱增,单独来源的 POI 数据已经不能满足这种需求。为了将不同来源的 POI 数据融合到一起,组成一个更完整的 POI 库,本文在国内外研究成果基础上,提出的 MNMDC 方法在 COM-NWT 的基础上引入了对距离和类别的判断,在 POI 数据集中可融合对象比例较低的情况下,准确率、召回率和 F1 值获得了明显的提升,更适用于本文中多个数据源的 POI 融合方案。但本文的数据仅仅是一个城区的数据,下一步研究方向应该是大数据下的数据融合方法。

参考文献 (References)

- [1] YUAN Q, CONG G, MA Z, et al. Time-aware point-of-interest recommendation [C]// Proceedings of the 36th ACM International Conference on Research and Development in Information Retrieval. New York: ACM, 2013: 363 – 372.
- [2] CHENG C, YANG H, LYU M R, et al. Where you like to go next: successive point-of-interest recommendation [C]// Proceedings of the 23rd International Joint Conference on Qualitative and Quantitative Practical Reasoning. Menlo Park, CA: AAAI Press, 2013: 2605 – 2611.
- [3] LIU B, FU Y, YAO Z, et al. Learning geographical preferences for point-of-interest recommendation [C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1043 – 1051.
- [4] 王勇. 深网 POI 信息获取与一致性处理方法研究 [J]. 测绘学报, 2017, 46(3): 399 – 399. (WANG Y. Research on crawling and consistency processing of POIs from deep Web [J]. Acta Geodaetica et Cartographica Sinica, 2017, 46(3): 399 – 399.)
- [5] BEERI C, DOYTSHER Y, KANZA Y, et al. Finding corresponding objects when integrating several geospatial datasets [J]// Proceedings of the 13th Annual ACM International Workshop on Geographic Information Systems. New York: ACM, 2005: 87 – 96.
- [6] BEERI C, KANZA Y, SAFRA E, et al. Object fusion in geographic information system [C]// Proceedings of the 30th International Conference on Very Large Data Bases, 2004, 30(10): 816 – 827.
- [7] SAMAL A, SETH S, CUETO K. Feature based approach to conflation of geospatial sources [J]. International Journal of Geographical Information Science, 2004, 18(5): 459 – 489.
- [8] 陈瑞. 基于多源 POI 数据的匹配融合方法研究 [D]. 兰州: 兰州交通大学, 2014. (CHEN R. Study on the method of matching and fusion based on the multi-source POI data [D]. Lanzhou: Lanzhou Jiaotong University, 2014.)
- [9] 李瑞姗. 基于自然语言处理的多源 POI 数据融合的研究 [D]. 青岛: 中国海洋大学, 2013. (LI R S. Multi-source POI information fusion based on natural language processing [D]. Qingdao: Ocean University of China, 2013.)
- [10] FONSECA F T, EGENHOFER M J, AGOURIS P, et al. Using ontologies for integrated geographic information system [J]. Transaction in GIS, 2002, 6(3): 231 – 257.
- [11] 张巍, 高新院, 李瑞姗. 空间位置信息的多源 POI 数据融合 [J]. 中国海洋大学学报: 自然科学版, 2014, 44(7): 111 – 116. (ZHANG W, GAO X Y, LI R S. Multi-source POI data fusion based on the spatial location information [J]. Periodical of Ocean University of China: Natural Science Edition, 2014, 44(7): 111 – 116.)
- [12] 高新院. 基于空间位置信息的多源 POI 数据融合问题的研究 [D]. 青岛: 中国海洋大学, 2013. (GAO X Y. Study on fusion of multi-source POI based on the spatial location information [D]. Qingdao: Ocean University of China, 2013.)
- [13] 王婷婷. 基于位置与属性的多源 POI 数据融合的研究 [D]. 青岛: 中国海洋大学, 2014. (WANG T T. Multi-source POI fusion based on geospatial and natural [D]. Qingdao: Ocean University of China, 2014.)
- [14] 汪和平, 韩正友. 球面距离公式的推导及应用 [J]. 数理化解题研究: 高中版, 2006(5): 8. (WANG H P, HAN Z Y. Derivation and application of spherical distance formula [J]. Research on Mathematical Physics Problem Solving, 2006(5): 8.)
- [15] KEVIN DREBLER, AXEL-CYRILLE NGONGA NCOMO. Time-efficient execution of bounded Jaro-Winkler distances [C]// Proceedings of the 9th International Conference on Ontology Matching. Aachen, Germany: CEUR-WS.org, 2014: 37 – 48.

XU Shuang, born in 1993, M. S. candidate. Her research interests include machine learning, data mining, big data visualization.

ZHANG Qian, born in 1990, Ph. D.. His research interests include machine learning, data mining, natural language processing.

LI Yan, born in 1993, M. S. candidate. Her research interests include machine learning, data mining.

LIU Jiayong, born in 1962, Ph. D., professor. His research interests include information security, network information processing.

(上接第 1319 页)

- [13] DONG X, NIU Z, SHI X, et al. Mining both positive and negative association rules from frequent and infrequent itemsets [C]// Proceedings of the 3rd International Conference on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2007: 122 – 133.
- [14] SWESI I M A O, BAKAR A A, KADIR A S A. Mining positive and negative association rules from interesting frequent and infrequent itemsets [C]// Proceedings of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE, 2012: 650 – 655.
- [15] ANTONIE M L. Mining positive and negative association rules: an approach for confined rules [C]// Proceedings of the 8th European

Conference on Principles and Practice of Knowledge Discovery in Databases. New York: Springer-Verlag, 2004: 27 – 38.

This work is partially supported by the National Natural Science Foundation of China (61673285), the Natural Science Foundation of Sichuan Education Department (15ZB0029), the Sichuan Youth Science and Technology Foundation (2017JQ0046).

CHEN Liu, born in 1991, M. S. candidate. Her research interests include data mining.

FENG Shan, born in 1967, Ph. D., professor. His research interests include intelligent platform for education software, data mining, realtime database system.