引文格式：孙晓霞. 融合语义和位置信息的兴趣点实体匹配方法[J]. 地理空间信息, 2023, 21(5): 22-24. doi:10. 3969/j.issn. 1672-4623.2023.05.006

融合语义和位置信息的兴趣点实体匹配方法

孙晓霞[1]
（1. 江苏省测绘工程院，江苏 南京 210013）

摘　要：兴趣点是反映城市建设的重要地理空间框架数据，及时快速更新兴趣点具有重要意义。提出了一种融合语义和位置信息的兴趣点实体匹配方法，可从互联网数据中获取新增的数据，提高了兴趣点实体的时效性，减少了人工检查的工作量。实验结果表明，该方法可快速更新兴趣点实体，降低生产成本，提高生产效率，提升多源数据的应用价值。

关键词：兴趣点实体；城市地理空间框架；Jaccard 距离；Levenshtein 距离
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）05-0022-03

Point of Interest Entity Matching Method Based on Semantic and Location Information

SUN Xiaoxia[1]
(1. Jiangsu Survey and Mapping Engineering Institute, Nanjing 210013, China)

Abstract: Point of interest is important geospatial framework data which can opportunely reflect city construction. Therefore, it is of great significance to update point of interest in time and quickly. In this paper, we proposed a new point of interest entity matching method based on semantic and location information, which could extract new data from Internet data and update points of interest more efficiently then manual operation. The experimental results show that this method can update the points of interest quickly, reduce the production cost, improve the production efficiency and enhance the value of multi-source data.

Key words: point of interest entity, urban geospatial framework, Jaccard distance, Levenshtein distance

兴趣点是重要的数字城市地理空间框架数据，极大地推动了地理信息的发展和应用[1-2]。兴趣点实体既可为用户更方便地规划出行，快速发现感兴趣位置[3]，又可服务于政府部门的管理和决策。随着城市的快速发展，传统测绘数据成果的时效性已无法与城市建设同步，因此对兴趣点的及时快速更新具有重要意义。朱琳[4]等利用兴趣点的名称与点之间的距离来判断兴趣点的相似度，距离在 100 m 以内的点认为是相同的点，需去重；李鹏飞[5]提出了多重属性语义相似度支持下的同名点对识别算法，用于提取新增兴趣点；刘禄艳[6]提出了基于知网的语义相似度算法，将词语语义相似性加入到文本相似度计算中，采用递归的方式提高了算法精度；徐鑫鑫[7]提出了一种改进的 WMD 距离算法，结合关键词计算 WMD-JCS 距离，度量文本的相似度；田红鹏[8]等提出了多特征融合权重的方法，根据数据集中不同类别特征的分布改进信息增益特征选择方法，进行加权融合，以提高文本相似度计算结果的精度。本文主要研究县级市兴趣点实体数据更新方法，模型融合了语义和位置信息，通过对比从互联网获取的靖江市域兴趣点数据与原兴趣点实体库，快速高效地筛选出确为新增的兴趣点数据，并更新到原实体库中。

1　研究区概况和研究方法

1.1　研究区概况

靖江市位于长江下游，江苏省苏北平原南端，东西南三面环江；南与张家港市、江阴市、武进区隔江相望，东北至西北与如皋市、泰兴市毗连，介于 31°56′~32°08′ N、120°01′~120°33′ E 之间，襟江近海；总面积为 655.58 km^2，拥有 52.3 km 的长江岸线，其中深水岸线 35 km，全线获批为国家一类开放口岸。

本文以数字靖江地理空间框架的兴趣点实体库为目标数据集，时效性为 2018 年上半年，共计 18 976 条数据；以高德地图 API 获取的兴趣点数据为参考数据集一，时效性为 2021 年上半年，共计 7 742 条数据；以百度地图 API 获取的兴趣点数据为参考数据集二，时效性为 2021 年下半年，共计 11 128 条数据，局部区域放大如图 1 所示。叠加显示两种数据集

收稿日期：2022-03-02；修回日期：2022-05-12。
项目来源：江苏省自然资源厅指导性科研型项目（2020050）。
作者简介：孙晓霞（1988—），硕士研究生，工程师，研究方向为大数据分析与智慧城市建设，E-mail: 570375304@qq.com。

可知，参考数据集与目标数据集既有重叠部分，又有不重叠部分，它们之中都可能存在新增的兴趣点数据。

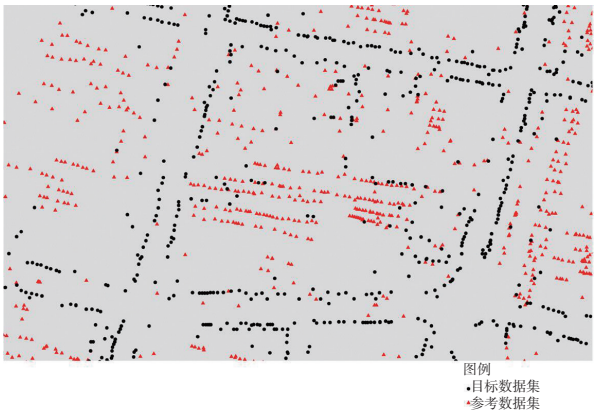


图1　目标数据集与参考数据集

1.2　研究方法

1）数据预处理。对多源兴趣点数据作归一化处理（数据格式、投影转换、拓扑检查、编码等），再将所有数据按照统一标准整合到统一的载体上，使之具有相同的空间参考。

2）名称统一化。对兴趣点的名称属性作统一处理，按照统一的标准规范简化行政地名（中英文括号 、大小写字母、数字、特殊符号等），减少无效匹配，如靖江XXX有限公司，统一为XXX有限公司。

3）初次匹配。筛选坐标和名称完全相同的点，做匹配标记，这些点无需新增到目标数据集中，不参与计算。

4）相似度评价指标。计算参考数据集中剩余点与目标数据集中所有点的距离相似度 $D_{A,B}$ 和语义相似度 $L_{A,B}$。

5）同名实体匹配。采用加权平均值计算综合相似度 $\mathrm{sim}(A,B)$；再根据样本的特征设置距离相似度权重 ω_D 和语义相似度权重 ω_L，且 $\omega_D+\omega_L=1$，最终计算公式为 $\mathrm{sim}(A,B)=\omega_D\times D_{A,B}+\omega_L\times L_{A,B}$。

6）兴趣点实体库更新。根据设定的相似度阈值，标记目标数据集与参考数据集中候选的兴趣点数据，建立两个数据集间的映射关系，提供多源数据融合的依据。

1.3　关键技术

1.3.1　语义相似度

本文中语义单指兴趣点的名称属性，虽然结合地址等其他属性的匹配确实可以提高匹配精度，但实际中获取的数据往往缺乏有效的地址信息等属性内容，而名称属性均为必填项，因此考虑到模型的通用性，在此只用名称属性。常用的衡量两个字符串相似程度的经典方法包括Jaccard距离法和Levenshtein距离法[9-11]。在计算语义相似度时，本文随机抽取了参考数据集中300个兴趣点数据采用上述两种方法进行实验，阈值为最终的相似度计算值，结果见表1，可以看出，阈值为0.8时匹配效果准确，但有较多点被遗漏；阈值为0.5时匹配度最高，但召回的点也较多，增加了工作量。综合考虑准确度和工作效率，最终选用Jaccard距离法，阈值设为0.6。

表1　两种方法结果

阈值	正确个数		召回个数		准确率/%		召回率/%	
	Jaccard距离法	Levenshtein距离法	Jaccard距离法	Levenshtein距离法	Jaccard距离法	Levenshtein距离法	Jaccard距离法	Levenshtein距离法
0.5	52	113	10	78	81.93	88.00	89.78	89.78
0.6	46	88	2	25	85.90	88.18	86.52	90.72
0.7	41	57	2	19	86.18	92.00	86.52	92.52
0.8	42	55	1	15	86.50	92.50	88.01	98.01

此外，在核查相似度匹配结果时发现一些桥梁由于用村名等行政名称命名，导致与村名的相似度较高，易被判断为高相似性，因此本文对Jaccard距离做了优化，当匹配的两个兴趣点中有一个是“桥”时，在计算距离时增加其权重，便于区分。

1.3.2　距离相似度

常用的空间距离计算方法包括欧氏距离、曼哈顿距离、Hausdorff距离和Fréchet距离等[12-15]。本文选用的数据都统一为投影后的平面坐标，综合考虑模型的计算复杂度和时间复杂度，最终采用欧氏距离，距离越小，相似度越高。结合影像，测量同一建筑物的覆盖范围，选取市中心的大型商场、地标建筑、周边村镇的公园、工业园区等建筑物，实验距离为100~500 m；依据实验的精度和时间复杂度，最终以322 m为最大判断距离，即同名点之间的距离大于322 m则认为是不同的点。

2　实验结果分析

匹配结果与精度如图2所示，精度均在85%以上。为进一步研究不同空间位置的精度差异，本文选

取500 m的网格，计算每个网格中兴趣点匹配的精度[16]，结果如图3所示，可以看出，精度较低的区域多分布在城镇聚集区周边或兴趣点密集区。分析错误匹配的兴趣点发现，大多数是自然村的点，如“春及庄”周围有“春及垒”或“春及圩”等，通过地理坐标结合影像核查，这类点中有的是正确的，确实有两个自然村落，少数代表同一个自然村，可能是名称更改或属性录入错误，该情况可在匹配结果中标注说明，后期需人工核查，但已大大减少了传统方法中需进行人工核查的数量。

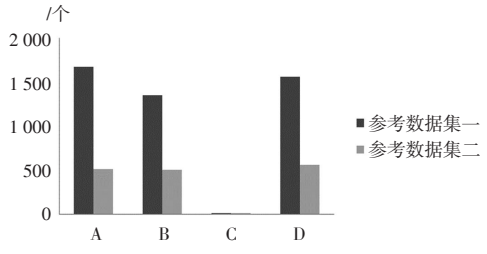
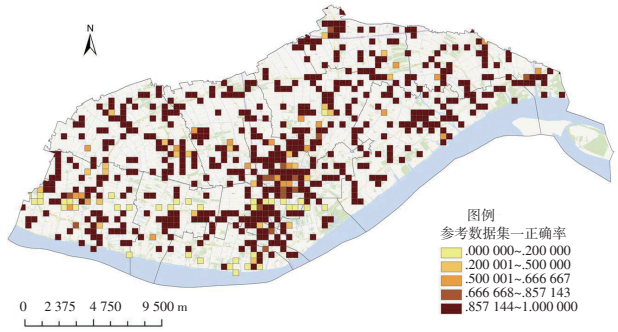
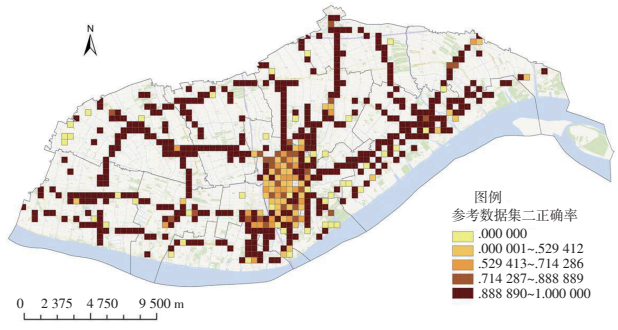


图2 匹配结果

注：A为需新增的点数，B为需新增点中正确的点数，C为相同的点中需新增的点数，D为真正新增点数。



a 参考数据集一



b 参考数据集二

图3 两个参考数据集的精度分布

3 结 语

实验结果表明，融合语义和位置信息的兴趣点实体数据匹配方法是可行的，能从海量的参考数据集中快速筛选出相似的兴趣点实体数据，并把确定为新增的兴趣点实体数据添加到目标数据集中，实现了兴趣点实体的快速更新，降低了生产成本，提高了生产效率，提升了多源数据的应用价值。

参考文献

[1] 张玲. POI的分类标准研究[J]. 测绘通报,2012(10):82-84

[2] 朱家彪. 公众地理信息公共服务平台系统建设研究:以湖南省为例[J]. 测绘通报,2010(9):60-62

[3] 汤加欣,陈阳,周孟莹,等. 深度学习方法在兴趣点推荐中的应用研究综述[J]. 计算机工程,2022,48(1):12-23

[4] 朱琳,占媛. 省级地理信息公共服务平台数据融合方案研究[J]. 地理空间信息,2021,19(8):61-64

[5] 李鹏飞. 基于互联网地理信息的公共服务平台POI数据增量更新研究[D]. 兰州:兰州交通大学,2020

[6] 刘禄艳. 基于语义的中文文本相似度的算法研究[D]. 西安:西安电子科技大学,2019

[7] 徐鑫鑫. 基于WMD距离的文本相似度算法研究[D]. 太原:太原理工大学,2019

[8] 田红鹏,马博,冯健. 多模型加权融合的文本相似度计算[J]. 计算机工程与设计,2021,42(11):3 239-3 245

[9] 林学民,王炜. 集合和字符串的相似度查询[J]. 计算机学报,2011,34(10):1 853-1 862

[10] 于凯,王立宏. 一种基于加权Jaccard距离的决策树集成选择方法[J]. 烟台大学学报(自然科学与工程版),2020,33(2):204-211

[11] 赵志靖. 编辑距离与语言分类[J]. 宁波大学学报(人文科学版),2018,31(5):63-68

[12] 黄博华,武芳,许俊奎,等. 一种面向同名线要素的距离度量方法[J]. 武汉大学学报(信息科学版),2017,42(3):395-401

[13] 曹闻,刘浩,李润生. 一种改进的Hausdorff距离地图匹配方法[J]. 计算机工程与应用,2013,49(6):159-162

[14] 郑少波,周国祥,张本宏,等. 基于离散Fréchet距离的地图匹配方法[J]. 合肥工业大学学报(自然科学版),2017,40(1):42-46

[15] 滕志军,李昊天,张宇,等. 动态距离权重因子的隐马尔可夫模型地图匹配算法[J]. 河南科技大学学报(自然科学版),2020,41(4):40-45

[16] XI K, GUO R Z, ZHANG Z J, et al. Spatial Context-based Local Toponym Extraction and Chinese Textual Address Segmentation from Urban POI Data[J]. ISPRS International Journal of Geo-information,2020(3):147