POI POINT ENTITY MATCHING AND FUSION BASED ON MULTI SIMILARITY CALCULATION

Zhao. Jianghong [1,2,3], Niu. Xinyu [1,3*], Cui. Yuanyuan [1,3], Zhao. Yingxue [4], Guo, Ming [1,3], Zhang, Ruiju [1,3]

[1]School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing , China-zhaojiangh@bucea.edu.cn, niuxinyu0225@163.com, 1095750135@qq.com, guoming@bucea.edu.cn, zhangruiju@bucea.edu.cn
[2]Key Laboratory for Urban Spatial Information of the Ministry of Natural Resources, Beijing 102616, China
[3]Beijing Key Laboratory for Architectural Heritage Fine Reconstruction & Health Monitoring, Beijing 102616, China
[4]Guangzhou Panyu Polytechnic, Guangzhou , China-821862137@qq.com
* Correspondence: niuxinyu0225@163.com

KEY WORDS: POI , Multi-source fusion, Text Similarity, Distance Similarity, Big Data Update, Analytic Hierarchy Process.

ABSTRACT:

This paper presents a multi-source POI matching method with multi feature similarity, which can effectively solve the problem of low matching accuracy of POI data from different sources. The spherical distance method, editing distance method and Jaro Winkler method are combined to calculate the distance, name, address distance and other main attributes of POI data. Then the importance of each feature index is analyzed by using analytic hierarchy process, and the feature weight of each similarity is obtained. The candidate matching objects are screened according to the total similarity to determine the final matching object. Finally, POI points are fused by selecting spherical center coordinates, name aliasing and address normalization methods. Experiments show that the recall and accuracy of this method for POI matching point recognition are significantly higher than those based on name similarity and distance similarity. The recall rate increased by 17.43% and 5.17% respectively, and the accuracy rate increased by 4.37% and 1.22%.It provides more comprehensive and accurate data support for urban function analysis and smart city construction.

1. INTRODUCTION

Point of interest (POI) is a kind of point data abstracted from real natural features or human markers, which usually contains a variety of information, such as name, address, category, longitude and latitude, etc. It is the link between spatial data and non spatial data, an important part of basic geographic information database, and the foundation and premise of building a "digital city" and "smart city".

However, the coverage of POI from a single source is inevitably incomplete. And because most of the current Internet maps obtain data through crowd-sourcing, the data quality is difficult to guarantee. POI information from different channels often has inconsistencies in location information, address and classification attributes. Therefore, it is necessary to identify the matching points from the map POI information of different sources and establish the fusion between the data, so as to facilitate the row statistics, analysis and processing of POI data in subsequent applications more effectively (Quan, Gao, Ma, Sun, & Magnenat-Thalmann, 2013).

There are three schemes for POI fusion methods at home and abroad, the method based on spatial location (Beeri, Doytsher, Kanza, Safra, & Sagiv, 2005; Wang, 2017), the method based on non spatial attributes (Chen, 2014; R. S. Li, 2013), and the method based on Ontology (GUO & Chen, 2015). The position based method can find the corresponding object only according to the longitude and latitude position information, but the longitude and latitude of POI from different sources generally have the problem of errors and inconsistent coordinate systems. The most common method is to use the spherical distance method to calculate the spatial distance similarity and screen the points that meet the fusion conditions (P. F. Li, Zhang, & Sun, 2021). There is also a preliminary distance screening of the two fusion sets through the mutual nearest neighbor method, which is insensitive to the coincidence degree of the two POI data sets (Xu, Zhang, Li, & Liu, 2018). Using the feature that the matching point entity data set has a high correspondence with the corresponding Tyson polygon, a point entity recognition and matching algorithm based on Tyson polygon is proposed (Wu & Wan, 2015). The Euclidean distance method is used to determine the candidate matching objects and provide candidate data sets for the subsequent POI matching (Luo, Ye, & Wang, 2022). The method based on non spatial attributes uses non spatial feature attributes without considering the differences in longitude and latitude. However, it requires that POIs from different sources must have a relatively unified storage mode, and non-spatial feature attributes may have problems with information loss and labeling errors. For the calculation of non spatial attributes, it is often reflected by text similarity. Some scholars use the method of filtering name similarity for the first time through experimental demonstration to select the best method of editing distance method (Zhang, Gao, & Li, 2014). The combination of cosine similarity and editing distance method improves the calculation accuracy of text similarity through three rounds of screening (Sun, 2021). Although the ontology based method is accurate and convenient, there is no mature ontology library for China's POI at present. The validated method is based on the construction of geographical ontology and the encapsulation of ontology attributes, and proposes a method of entity recognition with the same name through attribute similarity. This method uses the attribute similarity of POI, but it needs to carry out the construction and encapsulation of geographical ontology, and the workload of early data processing is large (Fonseca, Agouris, Egenhofer, & Mara, 2013).

There are few researches on POI matching fusion based on multi similarity. In this study, a POI point entity matching method based on multi similarity calculation is proposed, and the matching results are fused with multiple fields. For POI points from different sources, the spherical distance method is used to calculate the distance similarity, the editing distance method is used to calculate the name similarity, and the Jaro Winkler method is used to calculate the address similarity. The

scientific and objective analytic hierarchy process is used for weighted fusion, and the appropriate comprehensive similarity is selected to screen the matching points. The experimental results are compared with the accuracy of the traditional single attribute POI matching method, and the matching set is fused.

2. METHOD

2.1 Distance similarity calculation

For calculating the longitude and latitude similarity of two POI points, the most simple and effective method is to count the spherical distance between the two points, that is, the shortest distance between the two points on the sphere. The geographic coordinate similarity is defined as (1):

$$\text{distance}(A,B) = R\cos^{-1}(\cos(Lat1)\cos(Lon2))\cos(Lon2 - Lon1) + \sin(Lat1)\sin(Lat2), \quad (1)$$

where Lat1, Lat2 = latitude of A, B
Lon1, Lon2 = longitude of A, B
R = radius of the earth

When the value of distance (A,B) is greater than the set threshold 0.3, the distance similarity is recorded as 1, and it can be judged that A and B match.

2.2 Name similarity calculation

Editing distance, also known as Levenshtein Distance, refers to the minimum number of editing operations required to convert S_1 to S_2 between two groups of strings S_1 and S_2. In this study, Levenshtein.ratio, the distance ratio, is used to calculate the semantic similarity between S_1 and S_2. The calculation method is shown in formula (2):

$$\text{Levenshtein.ratio} = 1 - \frac{\text{Dist}}{\text{sum}}, \quad (2)$$

where Dist = the class editing distance
Sum = the min number of operations from S_1 to S_2

The ratio value indicates the similarity between S_1 and S_2, and the value is between 0 and 1. The minimum value of 0 indicates that they are completely different, and the maximum value of 1 indicates that they are completely matched.

2.3 Address similarity calculation

The algorithm highlights the importance of the same prefix, that is, if two strings are the same in the first few characters, they will obtain higher similarity. Therefore, it is more suitable for the similarity calculation of address fields with similar forms. See publicity (3) for the calculation method.

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

$$d_{jaro-winkler}(S_1,S_2) = d_{jaro}(S_1,S_2) + (lp(1 - d_{jaro}(S_1,S_2))), \quad (3)$$

where $d_{jaro}(S_1,S_2)$ = the Jaro distance between str S_1 , S_2
m = the number of matched characters
t = the number of transpositions
Sum = the min number of operations from S_1 to S_2
l = the length of prefix matching
p = the weight of prefix matching

The higher the final score of Jaro-Winkler distance, the greater the similarity. 0 means there is no similarity, and 1 means an exact match.

2.4 Build a comprehensive similarity model

The research takes Baidu POI and Gaode POI as the research objects, and first preprocesses the multi-source heterogeneous POI data, including data format conversion, unified coordinate system, etc. The similarity discrimination of a single attribute has certain limitations. In the process of recognizing entities with the same name in POI data, some POI data will have problems such as "different address with the same name", "same address with the same name" or "different address with the same name but similar longitude and latitude", which will affect the accuracy of entity recognition with the same name.

To avoid the above problems, a POI matching point recognition algorithm from different sources based on multi similarity calculation is proposed. The specific method is shown in Figure 1:

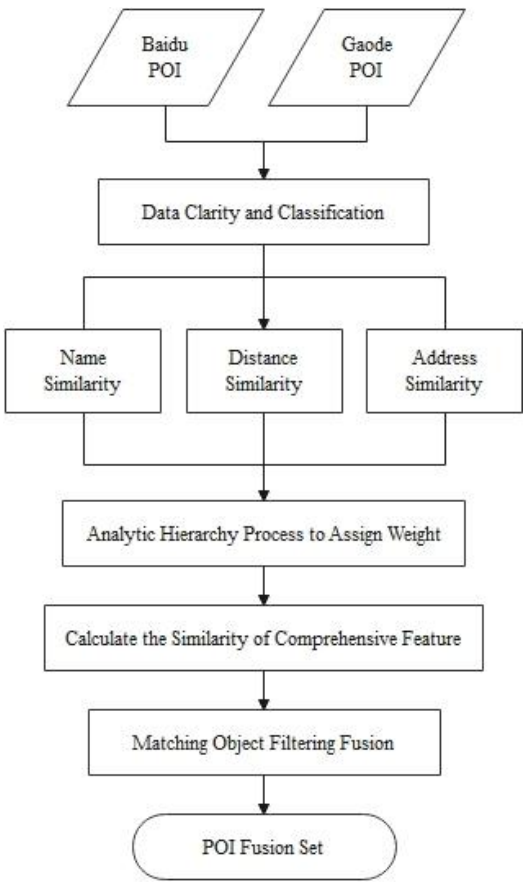

Figure 1. POI matching and fusion technology route.

This method combines the three attributes to form a comprehensive attribute similarity to distinguish the entities with the same name. The problem of single attribute discrimination is effectively solved by using analytic hierarchy process to calculate the weight. The accuracy and efficiency of homonymous entity recognition are greatly improved. The formula for calculating the attribute comprehensive similarity of POI is shown in formula (4).

$$\mathrm{Sim}_{(A,B)} = \frac{\lambda_1 \times \mathrm{Name}_{(A,B)} + \lambda_2 \times \mathrm{Dis}_{(A,B)} + \lambda_3 \times \mathrm{Add}_{(A,B)}}{\lambda_1 + \lambda_2 + \lambda_3} \quad (4)$$

where λ_1 = the weight of name similarity
λ_2 = the weight of distance similarity
λ_3 = the weight of address similarity
$\lambda_1 + \lambda_2 + \lambda_3 = 1$
$\mathrm{Name}_{(A,B)}$ = the value of name similarity
$\mathrm{Dis}_{(A,B)}$ = the value of distance similarity
$\mathrm{Add}_{(A,B)}$ = the value of address similarity

3. EXPERIMENT AND RESULTS

3.1 Study area and data

The study randomly selected a block in Dongcheng District of Beijing as the study area, and captured the POI data of Baidu map and Gaode map in the corresponding area through the web crawler program. After data cleaning such as null value elimination and deduplication, 812 Baidu POI data and 1731 Gaode POI data were finally obtained. Figure 2 shows the research area and experimental data of this paper. The circulars in the figure are Baidu POI data, and the diamonds are Gaode POI data.

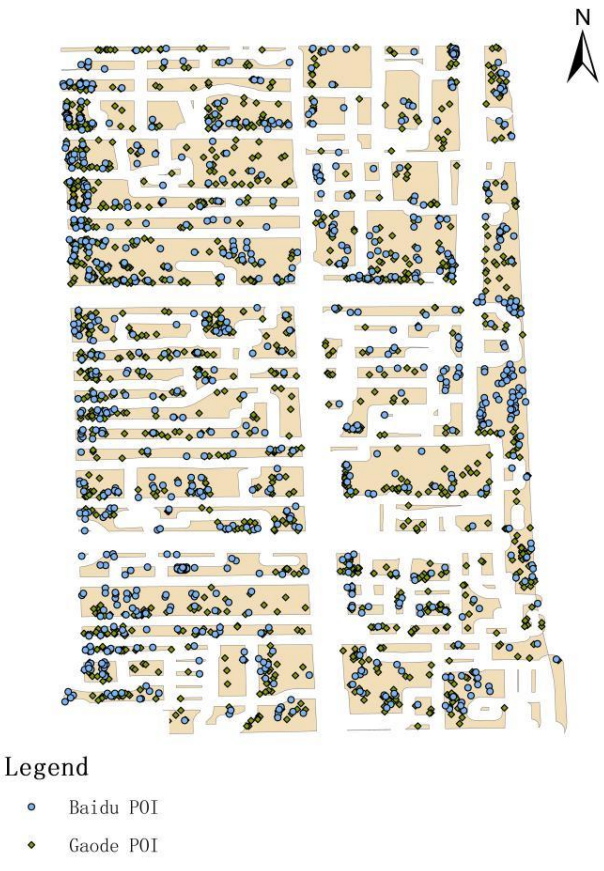

Figure 2. Study area and data.

3.2 Calculate similarity weight

The rationality of feature weight setting is related to the accuracy of matching results. This paper uses analytic hierarchy process to determine the weight of each feature. In order to avoid the situation of "strengthening" or "weakening" some characteristic indexes in the process of manually setting the weight, firstly, the importance of each characteristic index is analyzed by using the analytic hierarchy process (LIU, QIAN, WANG, & HE, 2015), the weight information is obtained, and then the importance is assigned.

Name and spatial location are important criteria for matching, and address information can assist in judgment. The gap caused by the same entity name is the least likely in reality, so the name is the most important. Having the same name and close geographical and spatial location is less likely to cause a gap, so the importance of spatial location is second. Due to the different address formats of data from different sources, address information indicators are used as auxiliary judgment indicators. Build the weight judgment matrix (5).

$$P = \begin{bmatrix} 1 & 3 & 5 \\ 1/3 & 1 & 2 \\ 1/5 & 1/2 & 1 \end{bmatrix}, \quad (5)$$

The maximum eigenvalue of the judgment matrix is obtained λ_{max}=3.004, the eigenvector is normalized to meet the requirements of $\lambda_1+\lambda_2+\lambda_3=1$, and the weights of name, distance and address are 0.648, 0230 and 0.122 respectively. To judge whether the weight setting of matrix P is reasonable, it also needs to pass the consistency test. CR=CI/RI=0.0036/0.58<0.1. It is judged that the matrix meets the requirements of consistency test and the weight distribution is reasonable.

3.3 Select the comprehensive similarity threshold

As the threshold is the judgment condition of matching entity recognition, reasonable setting of its value is the key to accurately identify the entities with the same name. When the attribute comprehensive similarity is greater than the set threshold, the two entities are entities with the same name, otherwise, they are not entities with the same name. This paper selects 200 groups of POI data in the same area of Baidu map and Gaode map to determine the best threshold, and 76 matching entities are obtained after manual verification. The experiment sets different thresholds, counts the number of matching entities identified and the number of matching entities accurately identified under each threshold, and calculates the recall rate, accuracy rate and F1 value. The results are shown in Table 1.

Thres-hold	Numb-er of recalls	Exact numb er	Recall (%)	Accurac-y(%)	F1(%)
0.60	75	65	98.68	86.66	92.28
0.65	74	66	97.36	89.18	93.09
0.70	73	68	96.05	93.15	94.57
0.75	73	70	96.05	95.89	95.96
0.80	72	68	94.73	94.44	94.58
0.85	70	67	92.10	95.71	93.87
0.90	69	67	90.78	97.10	93.83
0.95	67	66	88.15	98.50	93.03
1.00	66	65	86.84	98.48	92.29

Table 1. Recognition results of entities with the same name under different thresholds.

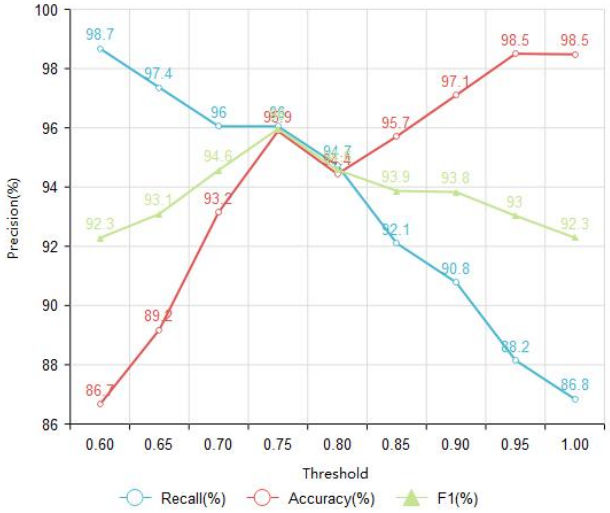

Figure 3. Recall rate, accuracy rate and F1 value of POI points with the same name under different thresholds.

It can be seen from Figure 3 that with the increase of the threshold, the recall rate of entity recognition with the same name is on the downward trend as a whole, and the accuracy rate is on the rise as a whole. This is because when the threshold is gradually increased, if it is a matching entity, its comprehensive similarity value must be greater than or equal to the threshold, so the number of recalled entities with the same name will decrease and the recall rate will decrease. At the same time, due to the increase of the comprehensive similarity, the recognition of the entity with the same name will be more accurate, and the accuracy rate will increase with the increase of the threshold. Therefore, the recall rate and the accuracy rate are negatively correlated, and the two cannot be both high at the same time. In this paper, the F1 value of the precision rate and the recall rate is considered to determine the best threshold. The F1 value is calculated as shown in formula (7).

$$F1 = (2 * \text{Recall} * \text{Accuracy}) / (\text{Recall} + \text{Accuracy}), \quad (7)$$

When the F1 value reaches the peak value, the corresponding threshold value is the best threshold value of entity matching effect. The corresponding threshold value when the F1 value reaches the peak value is 0.75, that is, when the threshold value is 0.75, the POI matching entity recognition effect is the best.

Three groups of POI points in Table 2 are selected as examples. Divide the three points into two groups to calculate name similarity, address similarity, distance similarity and weight them to obtain comprehensive similarity. Among them, the comprehensive similarity of the first group is 0.966105263, and the second group is 0.72056. According to the determined F1 threshold of 0.75, it can be concluded that the first group of comprehensive similarity greater than 0.75 is the same POI point, which is divided into the fusion set for subsequent processing; The second group of comprehensive similarity less than 0.75 is different POI points, which are reserved according to the original attributes.

Number	Name	Name Similarity	Address	Address Similarity	Latitude and Longitude	Distance Similarity	Comprehensive Similarity
1	Beijing Second Hospital Beijing Second Hospital	1	No.36 Youfang Hutong, Xuanwum-ennei Street, Xicheng District, Beijing No.36 Youfang Hutong, Xuannei Street, Xicheng District, Beijing	0.89473684 2	116.37057 13 39.90 0952 35 116.37041 99 39.90 0971 29	1	0.96 6105 263
2	Beijing Second Hospital Build-ing2, Beijing Second Hospital	0.7	No.36 Youfang Hutong, Xuannei Street, Xicheng District, Beijing Beijing Second Hospital, No.36 Youfang Hutong, Xuannei Street, Xicheng District, Beijing	0.68	116.37041 99 39.90 0971 29 116.37040 40 39.90 0738 32	1	0.72 056

Table 2. Example of identifying the same POI points using the comprehensive similarity method.

3.4 POI point matching with the same name

Based on the multi similarity comprehensive calculation method, the matching entity recognition experiment is carried out on the experimental data, and the experimental results are shown in Figure 4. The figure shows the result of superposition of Baidu POI and Gaode POI data and matching point recognition results in a block in Beijing. The red triangle in the figure is the matching points, the diamond is Gaode POI data, and the triangle is Gaode Baidu data.

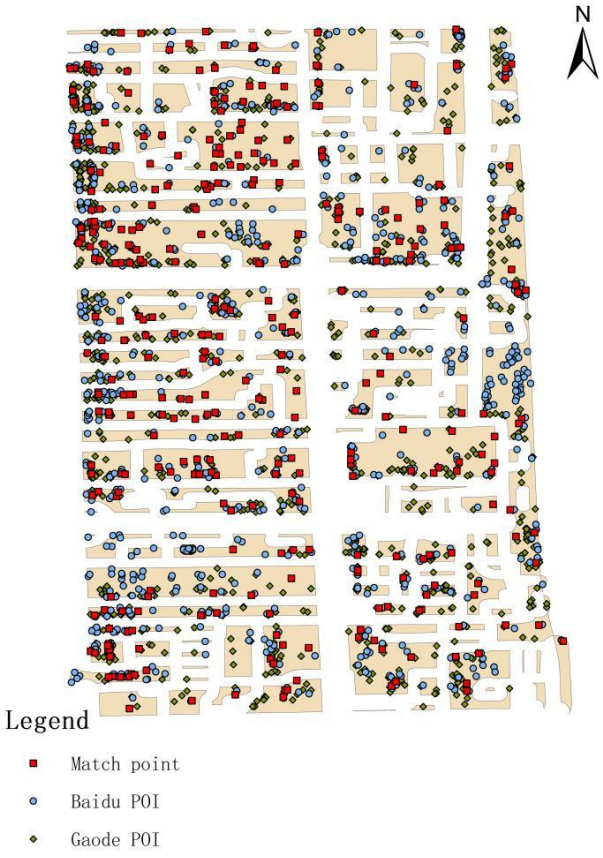

Figure 4. Recognition results of entities with the same name.

A total of 375 entities with the matching points were identified in this experiment. In order to verify the recall and accuracy of the results, the matching points in the two data sets are identified by manual comparison method, and 387 entities with the same name were obtained by manual verification. Through calculation, the recall rate and accuracy rate of the same name entity are 96.90% and 98.40% respectively, and the recognition result is relatively accurate.

3.5 Accuracy comparison

In order to further prove the accuracy of this method, this paper compares Baidu and Gaode POI data in the study area with two methods of matching entities with the same name based on name similarity and distance similarity. Table 3 shows the statistics of the experimental results of three homonymous entities recognition.

Method	Numbe-r of manual checks	Algorit-hm recogni-tion number	Correct identifi-cation number	Recal-l(%)	Accura-cy(%)
Based on name similarit-y	387	318	299	82.17	94.03
Based on distance similarit-y	387	355	345	91.73	97.18
Based on multi similarit-y	387	375	369	96.90	98.40

Table 3. Statistics of experimental results of three POI point matching methods with the same name.

It can be seen from the table that the algorithm based on the name similarity method matches 318 entities with the same name. Among them, the number of matching entities accurately identified was 299, and the recall rate and accuracy rate were 82.17% and 94.03% respectively. The number of matching entities based on distance similarity method is 355, of which the number of correctly identified matching entities is 345, and the recall rate and accuracy rate are 91.73% and 97.18% respectively. The number of entities with the same name matched based on multi similarity comprehensive calculation method is 375. Among them, the number of matching entities accurately identified was 369, and the recall rate and accuracy rate were 96.90% and 98.40% respectively. The method in this study is superior to the other two methods in terms of recall rate and accuracy based on comprehensive similarity, especially the recall rate is increased by 5.17%. This is mainly because the first two recognition methods only use POI single name attribute for recognition, while this method comprehensively considers the three attributes of POI name, address and distance to identify matching entities.

3.6 Fusion processing

In order to make the processed POI points convenient for subsequent research, multi field fusion processing is carried out on the matching result data. The mismatched POI points in the two sources are integrated in a unified format, and the matched POI points are fused according to the attributes of the three fields. For the longitude and latitude of the matching points, the central coordinates of the two fusion points are solved to realize position fusion. The research preserves the naming method of Gaode map for the names of matching points, aliases the names in Baidu map, and synthesizes the two POI points with different addresses for Standardization (according to provincial, urban, street, etc.). Table 4 shows an example of the results of the fusion operation.

Source	Name	Address	Longitude and latitude	(Alias)
Gaode	Dongcheng District Jianguomen Community Health Service Center	No. 9, houzhaojialou Hutong, Dongcheng District, Beijing	116.424188 39.913005	
Baidu	Jianguomen community health service center, Dongcheng District, Beijing	No. 9, zhaojialou Hutong, chaoneinan street, Dongcheng District, Beijing	116.424243 39.912862	
Fusion set	Dongcheng District Jianguomen Community Health Service Center	No. 9, zhaojialou Hutong, chaoneinan street, Dongcheng District, Beijing	116.424216 39.912934	Jianguomen community health service center, Dongcheng District, Beijing

Table 4. Example of matching set fusion results.

4. CONCLUSION

Based on the spatial and non-spatial attributes of different data sources, matching and carrying out multi-source data fusion is an attempt to comprehensively apply multi-source heterogeneous data. In the era of big data, massive data resources have been created, and data integration can reduce data production costs and improve data efficiency and quality.

Matching point recognition is the key technology to realize POI data fusion from different sources. In view of the deficiencies in the similarity measurement and weight determination of each feature, this paper studies the similarity calculation of the name information, location information and address information of POI points, and uses the analytic hierarchy process to give weights to each feature, so as to scientifically integrate the similarity between POI objects. This method is better than the POI matching method based on name similarity and distance similarity. It is more suitable for the direct matching of multi-source heterogeneous POI data, and can meet the needs of efficient fusion of multi-source POI data. At the same time, it also provides more abundant and accurate data support for other studies.

REFERENCES

Beeri, C., Doytsher, Y., Kanza, Y., Safra, E., & Sagiv, Y., 2005: Finding corresponding objects when integrating several geo-spatial datasets. *ACM*.

Chen, R., 2014: *Study on the Method of Matching and Fusion Based on the Multi-source POI Data.* (Master), Lanzhou Jiaotong University, Available from Cnki

Fonseca, F. T., Agouris, P., Egenhofer, M. J., & Mara, G. C., 2013: Research Article Using Ontologies for Integrated Geographic Information Systems.

GUO, X. J., & Chen, J. J., 2015: Research on Identical Entity Matching Based on GIS Ontology. *Computer Applications and Software, 32*(02), 66-68+112.

Li, P. F., Zhang, Y., & Sun, Q. K., 2021: Point of Interest Synthetic Similarity Caculation Method and its Application. *Science of Surveying and Mapping, 46*(09), 178-183. doi:10.16251/j.cnki.1009-2307.2021.09.023

Li, R. S., 2013: *Multi-source POI Information Fusion Based on Natural Language Processing.* (Master), Ocean University of Chin, Available from Cnki

LIU, H. L., QIAN, H. Z., WANG, X., & HE, H. W., 2015: Road Networks Global Matching Method Using Analytical Hierarchy Process. *Geomatics and Information Science of Wuhan University, 40*(05), 644-651. doi:10.13203/j.whugis20130350

Luo, G. W., Ye, J. Y., & Wang, J. F., 2022: Multi-Source POI Matching Method Based on Multi-Feature Similarity. *Bulletin of Surveying and Mapping*(04), 96-100. doi:10.13474/j.cnki.11-2246.2022.0117

Quan, Y., Gao, C., Ma, Z., Sun, A., & Magnenat-Thalmann, N., 2013: *Time-aware point-of-interest recommendation.* Paper presented at the Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.

Sun, L. C., 2021: *Research on Multi-source POI Information Fusion and Credibility Evaluation Technology.* (Master), Lanzhou Jiaotong University, Available from Cnki

Wang, Y., 2017: Research on Crawling and Consistency Processing of POIs from Deep Web. *Acta Geodaetica et Cartographica Sinica, 46*(03), 399.

Wu, J. H., & Wan, Y. Y., 2015: Point Entity Matching Algorithm Using Tyson Polygon. *Science of Surveying and Mapping, 40*(04), 97-100+120. doi:10.16251/j.cnki.1009-2307.2015.04.022

Xu, s., Zhang, Q., Li, Y., & Liu, J. Y., 2018: Fusion Algorithm of Multi-Source Interest Points Based on Distance Category. *Journal of Computer Applications, 38*(05), 1334-1338.

Zhang, W., Gao, X. Y., & Li, R. S., 2014: Multi-Source POI Data Fusion Based on the Spatial Location Information. *Periodical of Ocean University of China, 44*(07),111-116. doi:10.16441/j.cnki.hdxb.2014.07.018