International Journal of
Geo-Information

MDPI

Article

Different Sourcing Point of Interest Matching Method Considering Multiple Constraints

Chengming Li, Li Liu *, Zhaoxin Dai and Xiaoli Liu

Chinese Academy of Surveying and mapping, Beijing 100830, China; cmli@casm.ac.cn (C.L.); daizx@lreis.ac.cn (Z.D.); liuxl@casm.ac.cn (X.L.)
* Correspondence: liuli@casm.ac.cn

Received: 5 March 2020; Accepted: 27 March 2020; Published: 31 March 2020

Abstract: Point of interest (POI) matching is critical but is the most technically difficult part of multi-source POI fusion. The accurate matching of POIs from different sources is important for the effective reuse of POI data. However, the existing research on POI matching usually adopts weak constraints, which leads to a low POI matching accuracy. To address the shortcomings of previous studies, this paper proposes a POI matching method with multiple determination constraints. First, according to various attributes (name, class, and spatial location), a new calculation model considering spatial topology, name role labeling, and bottom-up class constraints is established. In addition, the optimal threshold values corresponding to the different attribute constraints are determined. Second, according to the multiattribute constraint values and optimal thresholds, a constraint model with multiple strict determination constraints is proposed. Finally, actual POI data from Baidu Map and Gaode Map in Dongying city is used to validate the method. Comparing to the existing method, the accuracy and recall of the proposed method increase 0.3% and 7.1%, respectively. The experimental results demonstrate that the proposed POI matching method attains a high matching accuracy and high feasibility.

Keywords: POI matching; strong constraints; role labeling; spatial topology; matching accuracy

1. Introduction

With the rapid development of electronic maps and mobile communication technologies, the demands for location-based services have progressively increased [1]. Geographic spatial data represented by points of interest (POIs) has received increasing attention [2–5]. At present, improving data richness and quality with complementary attributes through the fusion of POIs from different sources has become an effective way to rapidly update POI data [6–8]. However, because POI data from different sources usually exhibit issues such as inconsistency, redundancy, ambiguity, and contradiction [9,10], the appropriate method is important for accurately matching POI data from different sources [11]. POI matching from different sources usually refers to the process of discarding POIs representing the same objects but considering POIs representing different objects by comparing the POIs in reference and auxiliary maps with certain constraints. POI matching is a prerequisite and the key part for updating POIs. The rapid and accurate matching of POIs from different sources is critical to enrich and standardize POI databases and realize the effective reuse of data [12,13].

The methods to match multisource POI objects mainly include three categories: methods based on spatial attributes [13,14], methods based on nonspatial attributes, and methods combining both spatial and nonspatial attributes [15,16]. Because there are usually numerous uncertainties during POI matching, methods based on only one type of attribute will lead to poor matching results. A method combining both spatial and nonspatial attributes has the advantage of integrating multi-attributes such as name and spatial distance, which is more commonly being implemented in POI matching.

McKenzie et al., (2014) proposed a weighted multi-attributes strategy for matching POIs, which integrated attributes such as spatial location and distance, name attributes, and thematic similarity, and pointed out that methods combining multiple attributes can effectively solve the issue of a low matching accuracy caused by the use of a single attribute [17]. Based on spatial distance attributes, Huang et al., (2018) applied a nonspatial attribute, i.e., name similarity, to enhance the fusion accuracy of POI data from different sources [15]. Li et al., (2016) and Deng et al., (2019) proposed POI matching methods that combined the similarities of multiple attributes and their corresponding appropriate weights and demonstrated that among the existing methods [18,19], the method integrating the spatial distance, name and class attained the best performance.

However, the existing methods combining spatial and nonspatial attributes usually rely on weak constraints, which may lead to a low POI matching accuracy. For instance, previous research (1) often adopted weak name semantic constraints, where the similarity of names was directly calculated based on character strings, often resulting in different POI objects incorrectly being distinguished as the same objects due to their highly similar names. (2) Previous studies commonly adopted weak class distance constraints, which led to the inaccurate discrimination of POI objects of different classes with a small class distance. (3) Previous research applied spatial attributes that only considered location distance constraints but neglected other factors, such as the spatial topology between objects. All the above mentioned issues resulted in a low POI matching accuracy and poor matching results. Therefore, based on the characteristics of POI objects, this paper proposes a POI matching method integrating multiple determination constraints, which consider spatial topology, name role labeling, and bottom-up class constraints. The proposed method can effectively enhance the matching and fusion accuracy of POIs from different sources.

The paper includes five sections. Section 2 introduces the existing POI matching methods integrating spatial and nonspatial attributes and their shortcomings. The proposed POI matching method with multiple determination constraints is explained in detail in Section 3. Section 4 describes the experiments and presents the results, followed by discussions and conclusions in Section 5.

2. Related Work

2.1. Existing POI Matching Methods Integrating Dpatial and Nonspatial Attributes

At present, the latest POI matching methods combining spatial and nonspatial attributes usually implement weighted matching strategies integrating the spatial location and name, address and class attributes, which can substantially enhance the POI matching accuracy and recall rate. The core calculation algorithms are as follows:

(1) Spatial similarity calculation

The spatial location similarity refers to the geographical proximity of two objects in geographical space. The basic calculation method is the Euclidean distance method. In addition, to eliminate the effects of dimensionality, location parameters are normalized. The spatial similarity $S_{spatial}$ is calculated as follows:

$$S_{spatial} = e^{-\frac{S_{O_iO_j}}{cons}} \tag{1}$$

$$S_{O_iO_j} = \frac{1}{\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}} \tag{2}$$

where $S_{O_iO_j}$ denotes the coordinate similarity, $O_i(x_i, y_i)$ and $O_j\left(x_j, y_j\right)$ are the two sets of coordinates of POI objects from two different sources, and *cons* is a statistic constant that is determined by the training dataset. When $S_{spatial}$ equals 0, the two objects do not match completely, while a value of 1 indicates a complete match.

(2) Name similarity calculation

The common methods for POI name similarity calculation are the Jaro similarity algorithm, the Jaro–Winkler similarity algorithm, and the Levenshtein edit distance algorithm. Their commonality is based on quantitative string similarity representation. For instance, the Levenshtein edit distance algorithm is as defined follows:

$$S_{Lesh} = 1 - \frac{ED\left(NA_i, NA_j\right)}{Max\left\{L_{NA_i}, L_{NA_j}\right\}} \tag{3}$$

where NA_i *and* NA_j are two POI names; ED is the edit distance from NA_i to NA_j; L_{NA_i} is the length of POI name NA_i; and L_{NA_j} is the length of POI name NA_j.

(3) Class similarity calculation

In class similarity calculations, corresponding root node mapping relationships are first established. Then, the class distance between two nodes can be computed according to the determined root node mapping relationships and the depths from nodes to root nodes.

(4) Multi-attribute weighting

Weights are assigned according to the performance of attributes during matching. The overall similarity can be obtained by adopting weight and attribute similarities. When the overall similarity exceeds a certain threshold, the POIs are then considered to be the same. Otherwise, they are considered different POIs.

$$s = \sum_{i=1}^{n} s_i * y_i \tag{4}$$

where s is the overall similarity, s_i denotes the similarity of a single attribute, and γ_i is the weight of the attribute.

2.2. Shortcomings of the Existing Methods

Currently, the commonly used POI matching methods are mostly weighted matching methods that comprehensively consider the spatial location and name, address and class attributes. As mentioned above, these existing methods have certain limitations in terms of multi-attribute calculations and constraint setting. Therefore, a low POI matching accuracy and even incorrect matching results will occur in some scenarios. These scenarios are illustrated below.

Scenario 1: Because name attribute constraints are not rigorous enough, adjacent POI objects with a high name similarity are wrongly matched. As shown in Figure 1, there are two different POIs, and their names only differ by one number. When an existing similarity calculation method is used, the name similarity of these two POIs approaches 1. Moreover, because the two objects are close to each other, it is very likely that they will be considered the same object during matching.

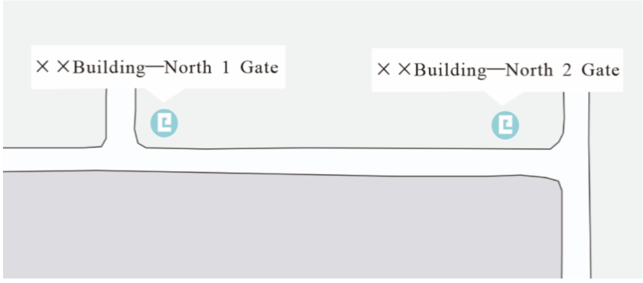


Figure 1. Different POIs with a high name similarity.

Scenario 2: Because class semantic constraints are often weak, if the primary classes of the POI data from different sources are different, the same POI objects are not correctly matched. In Figure 2, the classification employed in Baidu Map is presented in red, whereas that used in Gaode Map is shown in black. At the primary level, bars belong to 'Food' POIs in Baidu Map, but belong to the 'sports and leisure' POIs in Gaode Map. Based on existing methods, the class similarity of the same bar POI is ∞ to POI matching. This leads to a low class similarity, and therefore, the same bar will be considered to be two different objects in the two map systems.

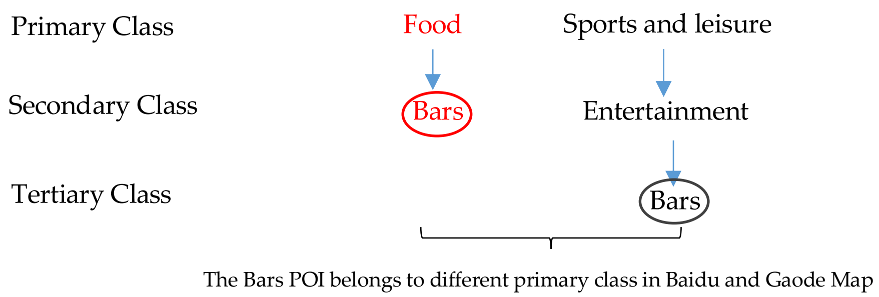
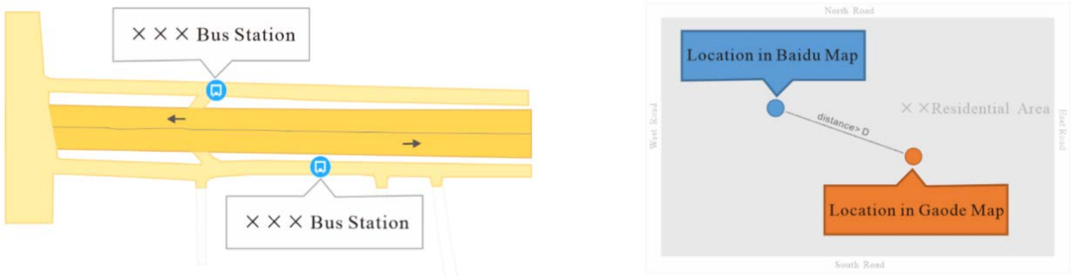


Figure 2. The same POIs with a low class similarity.

Scenario 3: Because only the spatial distance is considered and topological relationships are neglected, the POI objects on opposite sides or on the same plane are not correctly matched. Figure 3a shows two adjacent bus stops with the same name but located on opposite sides of the road. It is difficult to identify the correct POI for matching. In addition, Figure 3b reveals that for one POI object, due to the large residential area, the location of this POI object in the residential area is marked differently in Baidu Map and Gaode Map. The distance is 200 m, which usually exceeds the spatial distance threshold, leading to a low spatial distance similarity and incorrect and missed POI matching results.



(**a**) Two adjacent bus stops with the same name (**b**) The same POI with different location

Figure 3. Incorrect and missed matching only considering the spatial distance.

3. POI Matching Method Considering Multiple Constraints

When matching POIs from different sources, similarity calculations have to be improved for all attributes so that the matching accuracy for each attribute is enhanced and the overall POI matching accuracy can be increased. Therefore, this paper proposes a POI matching method considering multiple determination constraints. A more comprehensive and accurate POI matching is realized by improving the attribute similarity calculation and integrating various determination constraints. More specifically, for the attribute constraints, in addition to the POI name, address and class attributes, spatial constraints such as the topological relationships and distances between matching targets and their adjacent features are also captured. The POIs are adopted from Baidu Map and Gaode Map, and Gaode Map is used as the reference. The proposed method consists of three core parts: multi-attribute constraint calculation, determination of constraint thresholds, and definition of multiple determination constraints. A flow chart is shown in Figure 4.

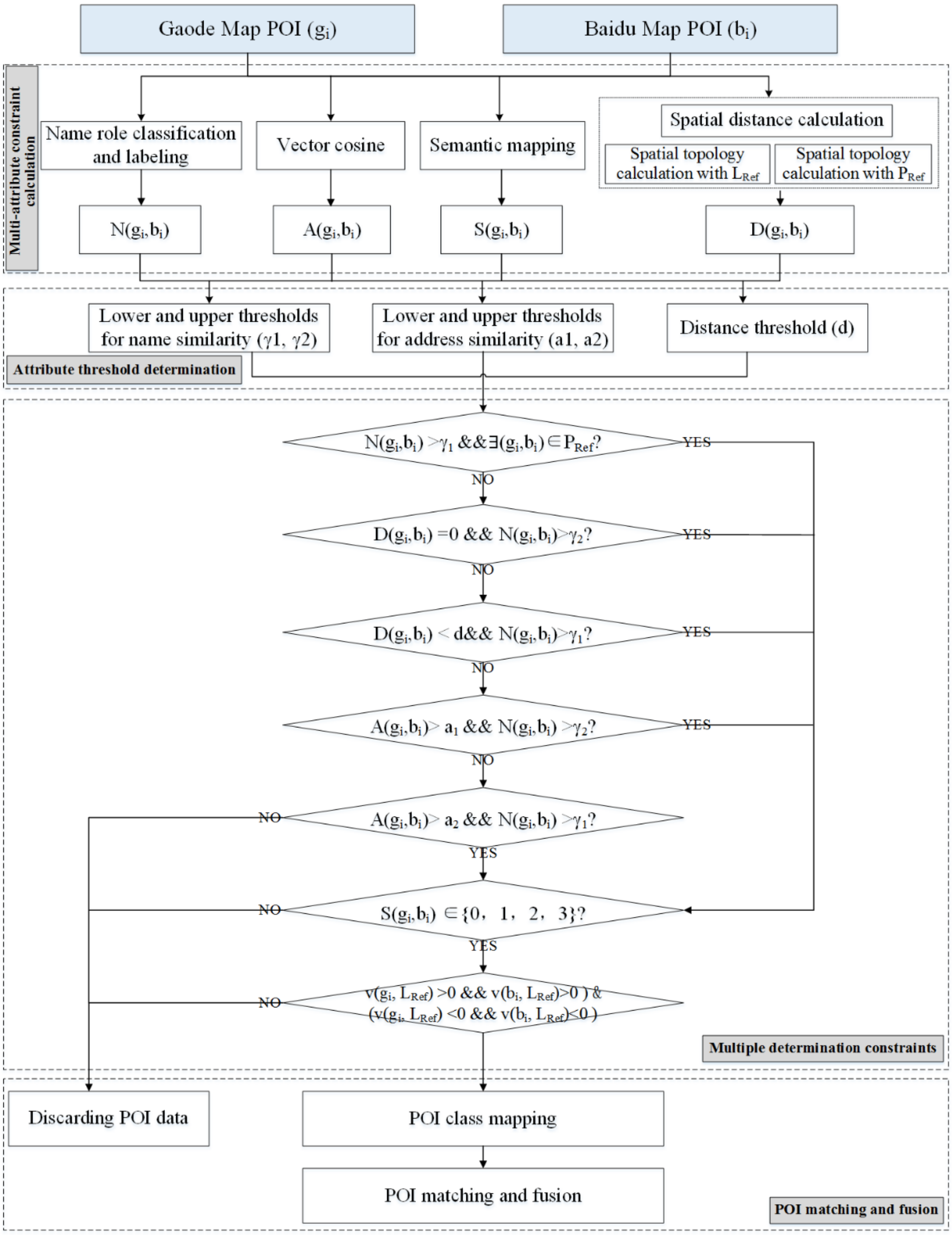

Gaode Map POI (g_i)
Baidu Map POI (b_i)
Multi-attribute constraint calculation
Name role classification and labeling
Vector cosine
Semantic mapping
Spatial distance calculation
Spatial topology calculation with L_{Ref}
Spatial topology calculation with P_{Ref}
$N(g_i,b_i)$
$A(g_i,b_i)$
$S(g_i,b_i)$
$D(g_i,b_i)$
Lower and upper thresholds for name similarity (γ1, γ2)
Lower and upper thresholds for address similarity (a1, a2)
Distance threshold (d)
Attribute threshold determination
$N(g_i,b_i) > \gamma_1$ &&∃$(g_i,b_i) \in P_{Ref}$?
$D(g_i,b_i) = 0$ && $N(g_i,b_i) > \gamma_2$?
$D(g_i,b_i) <$ d&& $N(g_i,b_i) > \gamma_1$?
$A(g_i,b_i) > a_1$ && $N(g_i,b_i) > \gamma_2$?
$A(g_i,b_i) > a_2$ && $N(g_i,b_i) > \gamma_1$?
$S(g_i,b_i) \in \{0, 1, 2, 3\}$?
$v(g_i, L_{Ref}) > 0$ && $v(b_i, L_{Ref}) > 0$) & ($v(g_i, L_{Ref}) < 0$ && $v(b_i, L_{Ref}) < 0$)
YES
NO
Multiple determination constraints
Discarding POI data
POI class mapping
POI matching and fusion
POI matching and fusion

Figure 4. Flow chart of the proposed method.

3.1. Multiattribute Constraint Calculation

3.1.1. Name Similarity Calculation

Name attributes are generally considered distinctive features to distinguish POIs. However, incorrect and missed POI matching results are easily attained by the small differences in numbers and direction words in POI names. To address this issue, this paper proposes a calculation method based on role labeling. Refinement and rigorous calculation are conducted for role labeling of proper names and direction and number words to improve the matching accuracy.

(1) Name role composition

For the Chinese names of most POIs, the word composition is relatively regular. According to their functions, the words in a POI name can be divided into place names (D), proper names (Z), adjectives (X), direction words (F), number words (S), common names (T), and special characters (Y). Hence, the role set of a POI name, NB = {D, Z, X, F, S, T, Y}. Figure 5 shows the semantic role composition of names "Dongying Huanayizhan Kuaijie Jiudian (Huanayizhan Express Inn, Dongying)" and "Kangju Xiaoqu Beiqu—25Haolou (Building no. 25—North Court, Kangju Residential Area)".

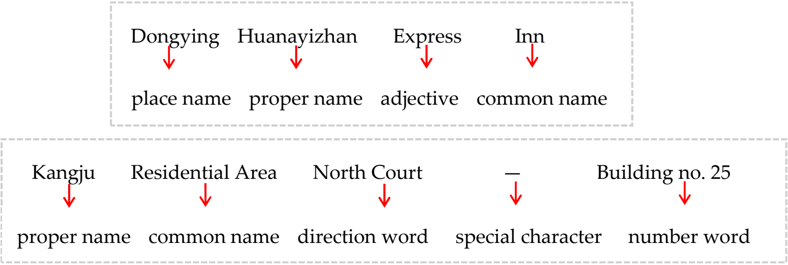


Figure 5. Schematic diagram illustrating the semantic role composition.

(2) Semantic role labeling

First, regular expressions are adopted to identify the direction and number words and special characters in POI names. Subsequently, based on role label dictionaries, including the place name, adjective, and common name, and the hidden Markov model (HMM) model, POI names are tokenized, and corresponding roles are assigned to the words. The dictionary structure is {words, roles, and times}. Finally, all words without any roles are classified as proper names, and their roles are defined as Z.

(3) Similarity calculation

For actual POI names, because roles D, X, T, and Y may be absent and they usually have minor contributions to the overall similarity or even lead to confusion during calculation of the overall similarity, in the proposed method, words with roles D, X, T, and Y are not included in the calculation. Similarity calculations are only based on words with the following roles: proper names (Z), direction words (F), and number words (S). The equation is given below:

$$N = \frac{1}{m}\sum_{i=1}^{m} W_i \tag{5}$$

where N is the name similarity, W_i is the similarity of words with role i, and m is the number of unions for the role labels between the names of two objects. Furthermore, the calculation methods for W_i differ for the different semantic roles. There are two main scenarios:

$$W_i = \left\{ \begin{array}{cc} 1 \text{ or } 0; & \text{if } W_i \in \{\text{direction words, number words}\}.\ W_i = 1 \text{ if they are identical. Otherwise, } W_i = 0. \\ 1 - \frac{ED(N_i,N_j)}{\max\{L_N,L_{Nj}\}}; & \text{if they role of } W_i \text{ is } a \text{ proper name.} \end{array} \right\} \tag{6}$$

where NA_i *and* NA_j are two POI names, ED is the edit distance from NA_i to NA_j, L_{NA_i} is the length of POI name NA_i, and L_{NA_j} is the length of POI name NA_j.

3.1.2. Address Similarity Calculation

Compared to name attributes, address attributes have lower POI matching capabilities. This occurs because, in reality, address descriptions are not standardized for many POIs (for example, certain descriptions include the administrative division while others do not), which leads to large uncertainties. In this paper, address similarity based on cosine similarity is employed for address similarity calculations [20], which includes two steps. First, a description of the administrative division has to be included in the address description, and meaningless special characters are removed. Second, two address pieces are distinguished with the open-source natural language processing framework HanLP, and two vectors are constructed according to the obtained piece words. Then, the similarity is computed based on the cosine values of the two vectors (Equation 5).

$$A(g_i, b_i) = \cos(\theta) = \frac{G.B}{||G||||B||} = \frac{\sum_{i=1}^{n} g_i * b_i}{\sqrt{\sum_{i=1}^{n}(g_i)^2} * \sqrt{\sum_{i=1}^{n}(b_i)^2}} \tag{7}$$

where g_i and b_i are the POIs in Gaode Map and Baidu Map, respectively, G and B are the vectors after address text encoding, and $A(g_i, b_i) \in (0,1)$ is the similarity of the two addresses. The addresses are more similar when the value approaches 1.

3.1.3. Class Similarity Calculation

POIs of the same class may be more similar to each other than POIs of different classes. The data of each map have their own classification system, and the levels and even the class names are different. Gaode Map consists of 23 primary levels, 264 secondary levels and 869 tertiary levels. Baidu Map has 19 primary levels and 138 secondary levels (Figure 6). To more accurately match POI classes from different sources, this paper establishes bottom-up class mapping to strengthen the class constraints and realize the highly accurate matching of class attributes.

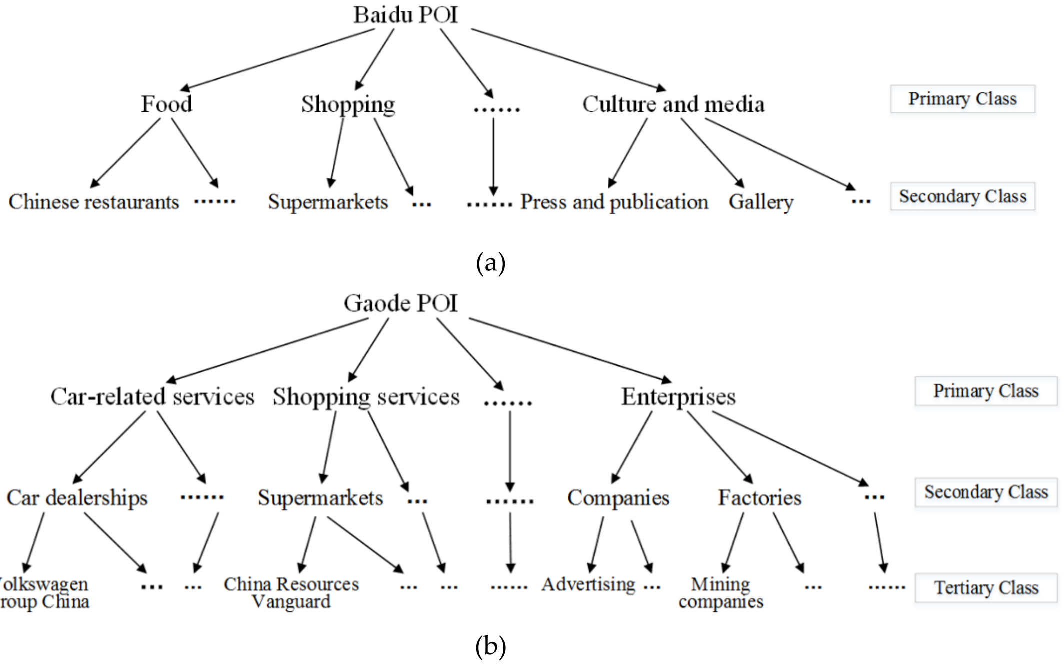


Figure 6. POI class categories of Baidu Map (**a**) and Gaode Map (**b**).

The semantic mapping relationships between the class nodes of the two classification systems are established. In particular, the mapping relationships between the parent classes are assigned to the subclasses by following the hierarchical trees. In this paper, there are three types of mapping relationships. The first type is the complete semantic mapping of class nodes, and the semantic distance is 0. For the second type, the mapping relationships of the class nodes are determined through their parent nodes, and the semantic distance can be 1, 2 or 3 (Deng et al., 2019). For the third type, there are no complete mapping relationships, and the semantic distance is +∞. Figure 7 shows the three main conditions under which the semantic distance is 0. They include (1) Chinese restaurants at the secondary level in Baidu Map and those at the secondary level in Gaode Map (Figure 7a); (2) bars at the secondary level in Baidu Map and those at the tertiary level in Gaode Map (Figure 7b); and (3) car dealerships at the secondary level in Baidu Map and those at the primary level in Gaode Map (Figure 7c). In these figures, the classifications in Baidu Map and Gaode Map are marked in red and black, respectively.

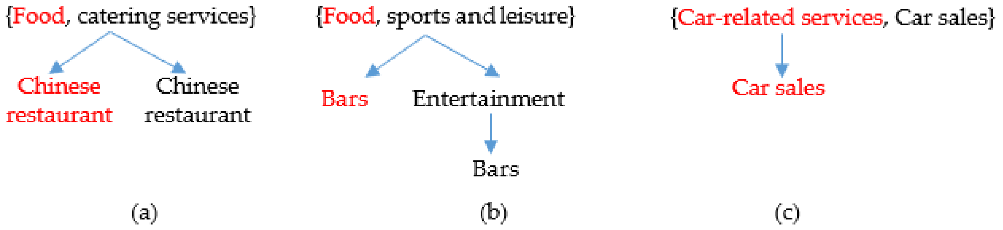

{Food, catering services}
Chinese restaurant
Chinese restaurant
{Food, sports and leisure}
Bars
Entertainment
Bars
{Car-related services, Car sales}
Car sales
(a)
(b)
(c)

Figure 7. Cases with a semantic distance equal to 0.

When the mapping relationships between the class nodes are determined based on their parent classes, the distances are calculated using the following equation:

$$A(g_i, b_i) = \text{Pstep}(gi) + \text{Pstepl}(bi) \tag{8}$$

where $A(g_i,b_i)$ is the semantic distance between a POI in Gaode Map and that in Baidu Map, and $\text{P}_{\text{step}}(g_i)$ and $\text{P}_{\text{stepl}}(b_i)$ are the numbers of steps from the nodes to their parent nodes to determine the mapping relationships based on the parent nodes in Gaode Map and Baidu Map, respectively.

3.1.4. Spatial Constraint Calculation Considering Spatial Topological Relationships

Generally, the closer two POIs are, the higher their matching probability becomes. However, there are large uncertainties if only distance constraints are considered. On the one hand, it is relatively challenging to set distance constraint thresholds. If the thresholds are set too high, some POI objects may not be matched or others will be omitted during matching, while excessively low thresholds will lead to more mismatches. On the other hand, even when two POIs are close to each other, matching errors may result due to inconsistent spatial topological relationships. Hence, this paper proposes a spatial constraint calculation method considering the spatial topological relationships between POIs and other features. The proposed method consists mainly of three components: spatial distance calculation, calculation of the spatial topological constraints between the POIs and line features and calculation of those between the POIs and polygon features.

(1) Spatial distance calculation

This paper adopts the planar distance between two POIs to measure their similarity. The equation is given below:

$$d(gi, bi) = R * arcos[\cos(y1i) * \cos(y2i) * \cos(x1i - x2i) + \sin(y1i) * \sin(y2i)] \tag{9}$$

where d is the planar distance between two points, R is the approximate radius of the Earth, and x_i and y_i are the latitude and longitude coordinates, respectively, of the two points.

(2) Calculation of the spatial topological constraints between POIs and line features

A line feature (L_{Ref}) is selected as a reference feature, such as a road, and when a POI is located along L_{Ref} or within the threshold distance from L_{Ref}, a spatial constraint relationship exists between the POI and L_{Ref}, that is, the POI belongs to L_{Ref}. The calculation equation is provided below:

$$p \in Line = \{\, \exists Line[i] \ |distance\,(p, Line[i] > M)\, \} \tag{10}$$

where $p \in Line$ denotes that point p belongs to linear object Line, ∃ indicates existence, $Line[i]$ is a certain segment of $Line$, distance $(p,Line[i])$ is the Euclidean distance between point p and *line* segment $Line[i]$, and M is the threshold. The value of M depends on the road classes in the city.

After determining whether a POI belongs to a line object, it is necessary to assess whether the POI is on the left or right side of the linear object. The equation is presented as follows:

$$v = (x1 - x) * (y2 - y) - (y1 - y) * (x2 - x) \tag{11}$$

where x and y are the latitude and longitude coordinates, respectively, of the POI and (x_1, y_1) and (x_2, y_2) are the coordinates of the endpoints of line segment Line[i]. When $v > 0$, the POI is located on the left side of the linear object. Conversely, when $v < 0$, it is on the right side of the linear object.

(3) Calculation of the spatial topological constraints between POIs and polygon features

A polygon feature ($\mathrm{P_{Ref}}$) is selected as a reference feature, such as a residential areas, and when a POI is located in or along the boundary of $\mathrm{P_{Ref}}$, a spatial constraint relationship exists between the POI and $\mathrm{P_{Ref}}$, i.e., the POI belongs to $\mathrm{P_{Ref}}$. The sum of all angles between the edges of $\mathrm{P_{Ref}}$ and the POI is calculated to determine whether the POI belongs to $\mathrm{P_{Ref}}$. The equation is as follows:

$$\mathrm{p} \in \mathrm{Area} = \left\{ \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{angle}(p, Area(i)) = 360 \right\} \tag{12}$$

where p ∈ Area denotes that point p belongs to the planar object Area, and angle(p,$Area$(i)) is the angle between point p and the i-th edge area ($Area$(i)) of the planar object. The equation of angle(p,$Area$(i)) is:

$$\mathrm{angle}(p, Area(i)) = \mathrm{arcos}\left(\frac{(x_1 - x) * (x_2 - x) + (y_1 - y) * (y_2 - y)}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2} + \sqrt{(x_2 - x)^2 + (y_2 - y)^2}} \right) \tag{13}$$

where x and y are the latitude and longitude coordinates, respectively, of the POI and (x_1, y_1) and (x_2, y_2) are the latitude and longitude coordinates of the endpoints of edge Area(i).

3.2. Determination of the Constraint Thresholds

The precision, recall and F1 score for POI matching are computed when name and address attributes and spatial distance are separately used. Thereafter, the optimal thresholds for the name and address similarities and spatial distance are selected. The precision refers to the ratio of the number of expected correct matches to the expected total number of matches (Equation 12). The recall refers to the ratio of the number of expected correct matches to the actual total number of true positive matches (Equation 13). The F1 score evaluates the balance between the precision and recall (Equation 14). F1 is the harmonic mean of the recall and precision. In order to ensure the scientificity and reliability of the thresholds, 3552 POIs from Gaode Map and 1350 POIs from Baidu Map that covers 15 primary classes are used to the performed tests. The classes include car-related services, catering service, shopping services, life services, transport facilities services, enterprises, schools, medical services (hospitals), government agencies, building and block numbers, etc.

$$precision = \frac{TP}{TP + FP} * 100\% \tag{14}$$

$$recall = \frac{TP}{TP + FN} * 100\% \tag{15}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{16}$$

3.2.1. Determination of the Name Similarity Threshold

The upper constraint threshold is selected based on the F1 score. As illustrated in Figure 8, when the name similarity is 0.8, the precision, recall, and F1 score coincide. At this point, the F1 score peaks.

Therefore, a name similarity of 0.8 is chosen as the upper threshold γ1. When the name similarity equals 0.5, the recall starts to decrease considerably. To ensure a relatively high recall, the lower name similarity threshold γ2 is set to 0.5.

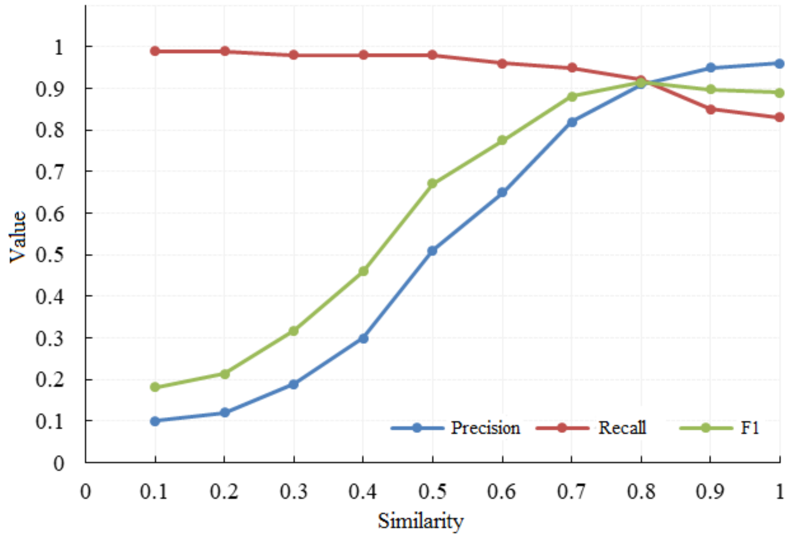


Figure 8. POI matching performance by using the name similarity.

3.2.2. Determination of the Address Similarity Threshold

Similarly, the F1 score is also considered when the upper address threshold is determined. As shown in Figure 9, the F1 score increases initially but rapidly decreases at the later stage and reaches its maximum when the address similarity is 0.8. Hence, the upper address similarity threshold a1 is defined as 0.8. Moreover, when the address similarity equals 0.4, the precision and F1 score simultaneously increase considerably. To ensure a relatively high recall, a lower address threshold a2 of 0.4 is chosen.

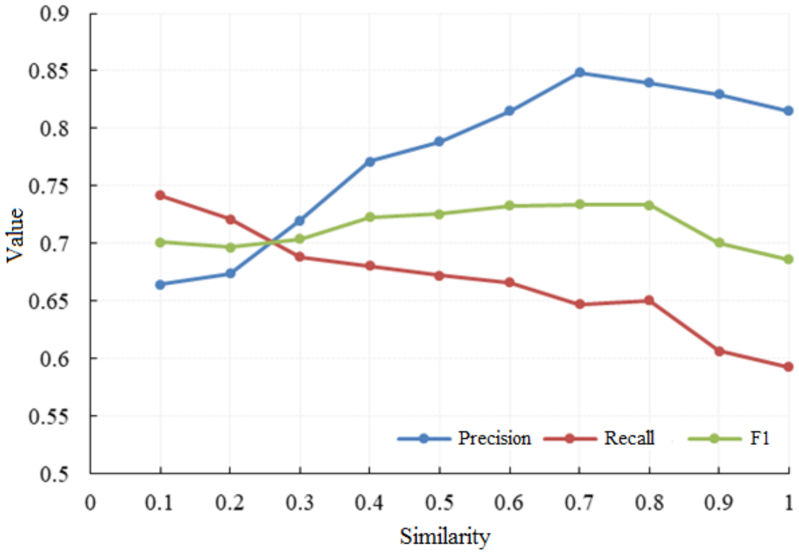


Figure 9. POI matching performance by using the address similarity.

3.2.3. Determination of the Spatial Distance Threshold

Figure 10 demonstrated that when the spatial distance is smaller than 10 m, the precision attains its highest value, but with increasing distance, the precision gradually decreases. In contrast, the opposite trend is observed for the recall. The recall gradually increases with increasing distance. When the distance exceeds 30 m, the F1 score stabilizes. By considering all three indicators, because the F1 score reaches its maximum and the precision and recall are both relatively high when the spatial distance is 50 m, the spatial distance threshold "d" is set to 50 m.

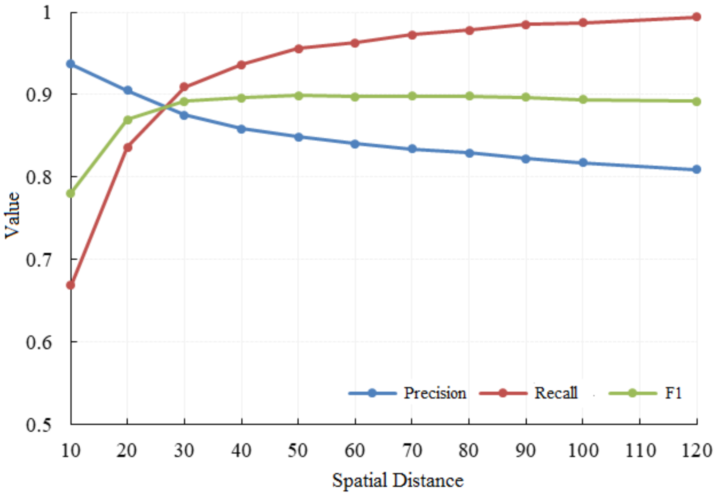


Figure 10. POI matching performance by using the spatial distance.

3.3. Multiple Determination Constraints

To match POIs more accurately, this paper defines multiple determination constraints based on the multi-attribute constraint calculation results. First, as many preliminary matching sets are selected as possible. Second, the preliminary matching sets are subjected to secondary filtering to remove objects that do not satisfy the set conditions. Finally, the POIs that satisfy the multiple determination constraints are matched and fused.

(1) Preliminary matching set selection

Various combinations of attribute constraints are employed to select as many targets as possible for matching, which is aimed at maximizing the recall. The POIs that satisfy any of the following constraints are included for secondary filtering. The constraints are given below:

$$\text{Constraint 1: } N(g_i,b_i) > \gamma_1 \ \&\& \ \exists(g_i,b_i) \in P_{Ref}$$

$$\text{Constraint 2: } D(g_i,b_i) = 0 \ \&\& \ N(g_i,b_i) > \gamma_2$$

$$\text{Constraint 3: } D(g_i,b_i) < d \ \&\& \ N(g_i,b_i) > \gamma_1$$

$$\text{Constraint 4: } A(g_i,b_i) > a_1 \ \&\& \ N(g_i,b_i) > \gamma_2$$

$$\text{Constraint 5: } A(g_i,b_i) > a_2 \ \&\& \ N(g_i,b_i) > \gamma_1$$

where g_i and b_i are the two types of POIs to be matched, $N(g_i,b_i)$ is the name similarity, $D(g_i,b_i)$ is the distance between the POIs, $A(g_i,b_i)$ is the address similarity, γ_1 and γ_2 are the upper and lower name similarity thresholds, respectively, "d" is the distance threshold, and a_1 and a_2 are the upper and lower address similarity thresholds, respectively.

(2) Preliminary matching set filtering

Constraints are imposed on the classes and the spatial topological relationships of the preliminary matching sets to further enhance the matching accuracy. The class constraint is given below:

$$S(g_i,b_i) \in \{0,1,2,3\}$$

where $S(g_i,b_i)$ is the class semantic distance between the two POIs.

The constraint on the spatial topological relationship of the POIs with other line features is provided as follows:

If $\exists(g_i,b_i) \in L_{Ref}$, then $(v(g_i, L_{Ref}) >0 \ \&\& \ v(b_i, L_{Ref})>0)$ or $(v(g_i, L_{Ref}) <0 \ \&\& \ v(b_i, L_{Ref})<0)$ has to be satisfied, where L_{Ref} is the same line object to which the two POIs belong, and $v(g_{i/}b_i, L_{Ref})$ is the direction value of the POI and the line object.

4. Experiments and Analyses

4.1. Experimental Data

POI data in Dongying, Shandong Province, are employed to validate the accuracy and effectiveness of the proposed method. The data are adopted from Gaode Map and Baidu Map, and the data from Gaode Map are considered as the references. The POI data of an arbitrarily selected area covering 1000 (m) × 1200 (m) in Dongying city are selected as the experimental data. There are 2220 and 1350 POIs from Gaode Map and Baidu Map, respectively (Figure 11). The classes of the POI data include building and block numbers, catering services, companies, schools, hospitals, shopping services, and government agencies.

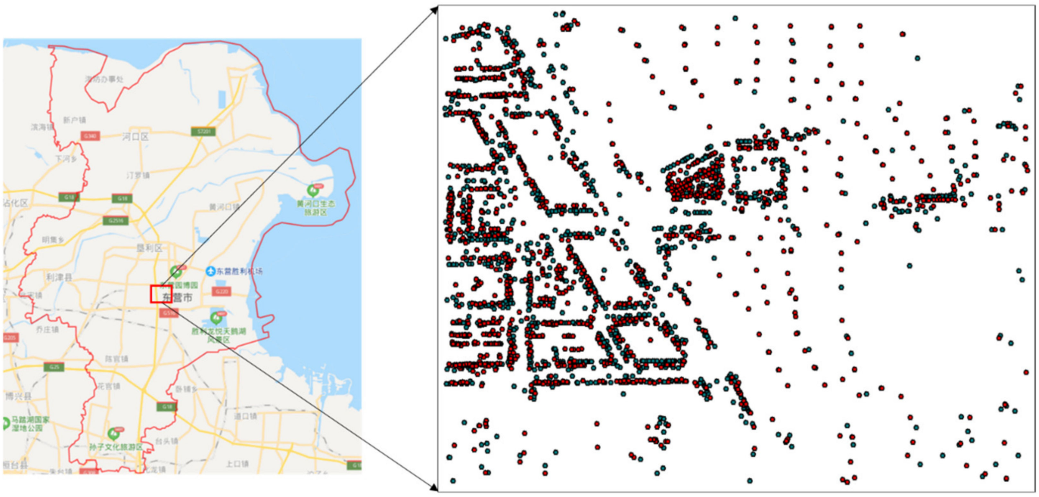


Figure 11. POI matching performance by using the spatial distance.

4.2. Overall Accuracy Analysis

To validate the accuracy of the proposed method, the results based on the proposed and existing POI matching methods considering multiple attributes are compared and analyzed using the same dataset of Dongying. The Baidu Map and Gaode Map use the BD-09 and GCJ-02 coordinate system, respectively, both of them were obtained by the WGS-84 encryption. Before matching, all datasets have undergone WGS84 coordinate transformation and data cleaning. The POI matching results are assessed using precision, recall and F1 score indicators. Table 1 summarizes the overall accuracy of the proposed and existing methods and manual discrimination.

Table 1. Overall Point of Interest (POI) matching performance.

Method	Number of Successful Matches	Number of Incorrect Matches	Number of Missing Matches	Precision	Recall	F1
the existing method	1,245	97	47	96.1%	92.2%	0.941
the proposed method	1,159	28	41	96.4%	97.5%	0.969
manual discrimination	1,162	0	0	100%	100%	1

From Table 1 and Figure 12, the numbers of incorrect and missing matches for the proposed method are smaller than those for the existing method. More specifically, the number of incorrect matches decreases by 76.7%. The POI matching precision for both the proposed and existing methods considering multiple attributes is approximately 96%, indicating that both methods can match POIs reasonably well. The proposed method slightly outperforms the other methods. However, the recall of the proposed POI matching method is 7.1% higher than that of the existing method, suggesting that the proposed method has more accurate POI descriptions and a high probability of classifying the POI to be matched into the correct class. Moreover, the F1 score of the proposed method is substantially higher than that of the existing method. This demonstrates that the proposed method notably outperforms the existing method in terms of the POI matching accuracy.

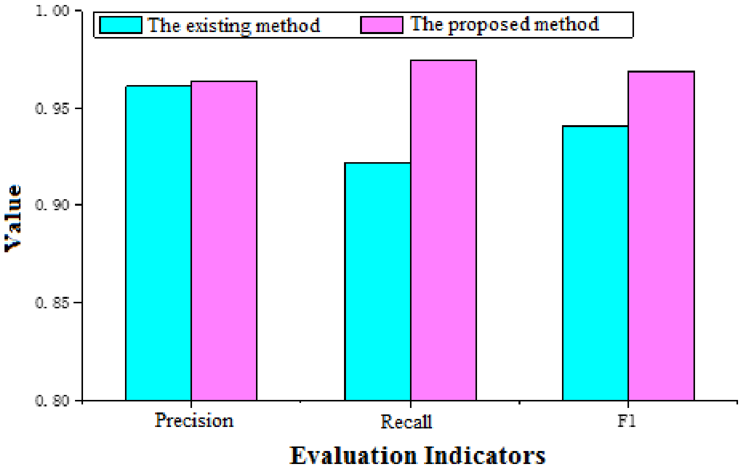


Figure 12. Overall POI matching performance.

Although the proposed method can effectively reduce the incorrect and missing matches, they are still present when the proposed method is employed (as indicated in Table 1). This may be the result of the following reasons. First, the same POI, with the same pronunciation, may be described with different words. Second, for the same POI, its full name is used in one of the sources, but an abbreviation is employed in another source, leading to a low name similarity. Third, the order of the words in the name of the same POI differ among the different sources. For example, "Binzhou Road/Fuqian Street (intersection)" and "intersection of Fuqian Street and Binzhou Road" actually refer to the same POI. Fourth, the attribute information of the POI itself is incorrect.

4.3. Superiority Analysis

To validate the superiority of the proposed method in POI matching, the matching performance in different scenarios is evaluated and analyzed (Table 2 and Figure 13). As indicated in Table 2, the proposed method effectively conducts POI matching in the three scenarios, which have been mentioned and defined in Section 2.2.

More specifically, in Scenario 1, the number of inaccurate matches by the existing method is the highest, while there are no inaccurate matches by the proposed method. This suggests that the proposed method can considerably reduce the incorrect matches of POIs that are close to each other but possess a high name similarity. The POIs in Scenario 1 are mostly located in residential and commercial areas. In Scenario 2, the proposed method effectively avoids missing matches of the same POI objects even if they are of different primary classes. The POIs in Scenario 2 are prone to mismatching when the primary classes of the POI objects are transportation facilities or sports leisure. In Scenario 3, the proposed method has zero incorrect and missing matches. This confirms that the proposed method with spatial constraints can significantly enhance the matching precision and recall because it also considers spatial topological relationships. In Dongying, the POIs in Scenario 3 are mostly located along both sides of roads and in large commercial areas.

Table 2. POI matching performance under the different scenarios.

Method	Scenario 1		Scenario 2		Scenario 3	
	Incorrect Matches	Missing Matches	Incorrect Matches	Missing Matches	Incorrect Matches	Missing Matches
existing method	63	0	4	1	2	5
proposed method	0	0	0	0	0	0

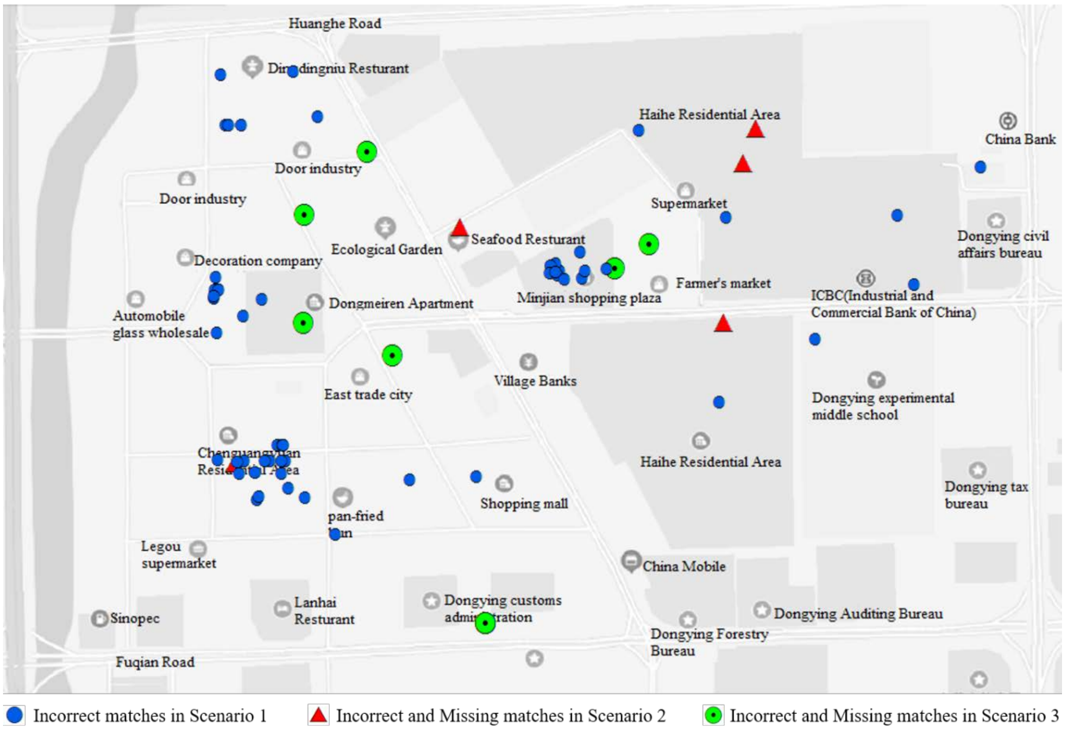


Figure 13. Spatial distribution of the POIs in the three scenarios.

4.4. *Citywide Validation*

Experiments on the entire city of Dongying (with an area of 7923 km^2) are further conducted to validate the proposed POI matching method. A total of 138.376 and 80.167 POIs are captured from Gaode Map and Baidu Map, respectively. With the proposed method, a total of 149.276 POIs are fused and matched, with an increase of 10.900 POIs. Furthermore, the recall of the matched POIs in the citywide experiment are basically the same as those of the experimental data, and the matching results totally won provincial quality requirements, which again demonstrates the accuracy, robustness, and versatility of the proposed method.

5. Conclusions and Discussions

POI matching is a prerequisite and the key part of POI fusion and updating using different map sources. The accuracy of POI matching is important to the improvement and standardization of POI databases. However, the existing research on POI matching usually adopts weak constraints, which leads to a low POI matching accuracy. Therefore, this paper proposes a POI matching method that considers the spatial topology and bottom-up class and strict name role constraints and multiple determination constraints. This effectively enhances the POI fusion and matching accuracy for data from Gaode Map and Baidu Map. The proposed method is validated using actual POI data in Dongying, Shandong Province. The main conclusions are as follows:

(1) Regarding the overall accuracy, the numbers of incorrect and missing POI matches by the proposed method are both smaller than those by the existing method. In particular, the number of incorrect matches is reduced by 76.7%. The POI matching precision of both the proposed and existing methods is approximately 96%. However, in terms of the recall and F1 score, the proposed method effectively increases their values by 7.1% and 0.3, respectively, highly demonstrating that the proposed method considerably outperforms the existing method in terms of the POI matching accuracy.

(2) In terms of superiority, there are no incorrect and missing matches in any of the three considered scenarios with the proposed method, suggesting its superiority.

(3) In citywide validation, the recall is basically the same as those of the experimental data, and the matching results totally won provincial quality requirements. This strongly supports the robustness and versatility of the proposed method.

The proposed method is primarily experimentally validated using data from Baidu Map and Gaode Map, which is also applicable to matching POIs from other data sources, but these matching results require further verification. In future research, more attention will be given to the mutual complementation and calibration of POI information from multiple data sources.:

Author Contributions: C.L. conceived the original idea for the study, and all co-authors conceived and designed the methodology. L.L., Z.D. and X.L. conducted the processing and analysis of the data. C.L., L.L. and Z.D. drafted the manuscript. All authors have to read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, 2018YFB2100702.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aliannejadi, M.; Crestani, F. Personalized Context-Aware Point of Interest Recommendation. *ACM Trans. Inf.* **2018**, *36*, 45. [CrossRef]
2. Huang, Y.; Xiong, H.; Leach, K.; Zhang, Y.; Chow, P.; Fua, K.; Barnes, L.E. Assessing social anxiety using gps trajectories and point-of-interest data. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016.
3. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J., Jr.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* **2015**, *53*, 36–46. [CrossRef]
4. Liu, X.; Long, Y. Automated identification and characterization of parcels with OpenStreetMap and points of interest. *Environ. Plan. B Urban Anal. City Sci.* **2015**, *43*, 341–360. [CrossRef]
5. Zhang, Z.; Liu, L.; Li, L.; Zhang, X. A point of interest recommendation method using user similarity. *Web Intell.* **2018**, *16*, 105–112. [CrossRef]
6. Kim, J.; Vasardani, M.; Winter, S. Similarity matching for integrating spatial information extracted from place descriptions. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 56–80. [CrossRef]
7. Lamprianidis, G.; Skoutas, D.; Papatheodorou, G.; Pfoser, D. Extraction, integration and analysis of crowdsourced points of interest from multiple web sources. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Dallas, TX, USA, 4–7 November 2014; Volume 11, pp. 16–23.
8. Scheffler, T.; Schirru, R.; Lehmann, P. *Matching Points of Interest from Different Social Networking Sites*; Springer: Berlin, Germany, 2012; pp. 245–248.
9. Hochmair, H.; Juhász, L.; Cvetojevic, S. Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. In Proceedings of the LBS 2018: 14th International Conference on Location Based Services, Zurich, Switzerland, 15–17 January 2018; pp. 293–313.
10. Novack, T.; Peters, R.; Zipf, A. Graph-Based Matching of Points-of-Interest from Collaborative Geo-Datasets. *ISPRS Int. J. Geo Inf.* **2018**, *7*, 117. [CrossRef]

11. Mckenzie, G.; Janowicz, K.; Adams, B. Weighted multi–attribute matching of user–generated points of interest. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, 5–8 November 2013.
12. Xia, Y.; Luo, S.; Zhang, X.; Bae, H.Y. Organization and Retrieval Method of Multimodal Point of Interest Data Based on Geo-ontology. *Adv. Sci. Technol. Lett.* **2014**, *45*, 49–54.
13. Yang, B.; Zhang, Y.; Lu, F. Geometric-based approach for integrating vgi pois and road networks. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 126–147. [CrossRef]
14. Zhang, W.; Gao, X.; Li, R. Multi-Source POI data fusion based on the spatial location information. *Period. Ocean Univ. China* **2014**, *44*, 111–116.
15. Huang, M. Multi-source POI Duplication Detection Method in Map World Fujian Based on Word Segmentation. *Geospat. Inf.* **2018**, *16*, 51–53.
16. Yang, B.; Zhang, Y. Pattern-mining approach for conflating crowdsourcing road networks with POIs. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 786–805. [CrossRef]
17. Mckenzie, G.; Janowicz, K.; Adams, B. A weighted multi–attribute method for matching user–generated Points of Interest. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 125–137. [CrossRef]
18. Li, L.; Xing, X.; Xia, H.; Huang, X. Entropy-Weighted Instance Matching Between Different Sourcing Points of Interest. *Entropy* **2016**, *18*, 45. [CrossRef]
19. Deng, Y.; Luo, A.; Liu, J.; Wang, Y. Point of Interest Matching between Different Geospatial Datasets. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 435. [CrossRef]
20. Zhang, Z.Y.; Wang, J.; Cheng, H.M. An approach for spatial index of text information based on cosine similarity. *Comput. Sci.* **2005**, *32*, 160–163.