



Article

# POI Data Fusion Method Based on Multi-Feature Matching and Optimization

Yue Wang <sup>1</sup> , Cailin Li <sup>1,2,\*</sup> , Hongjun Zhang <sup>3</sup>, Baoyun Guo <sup>1</sup>, Xianlong Wei <sup>1</sup> and Zhao Hai <sup>1</sup>

<sup>1</sup> School of Civil Engineering and Geomatics, Shandong University of Technology, Zibo 255000, China; 22407010003@stumail.sdut.edu.cn (Y.W.); guobaoyun@sdut.edu.cn (B.G.); 23407010846@sdut.edu.cn (X.W.); 23507020864@sdut.edu.cn (Z.H.)

<sup>2</sup> Hubei LuoJia Laboratory, Wuhan 430079, China

<sup>3</sup> Geographic Information Engineering, Shandong Provincial Institute of Land Surveying and Mapping, Jinan 250102, China; zhanghjgtchy@shandong.cn

\* Correspondence: licailin@sdut.edu.cn

**Abstract:** The key to geospatial data integration lies in identifying corresponding objects from different sources. Aiming at the problem of the low matching accuracy of geospatial entities under a single feature attribute, a geospatial entity matching method based on multi-feature value calculation is proposed. Firstly, when dealing with POI (point of interest) data, the similarity of POI data in terms of name, address, and distance is calculated by combining the improved hybrid similarity method, the Jaccard method, and the Euclidean metric method. Secondly, the random forest algorithm is utilized to dynamically determine the information weights of each attribute and calculate the comprehensive similarity. Finally, taking the area within the Second Ring Road in Beijing as the experimental area, the POI data of Tencent Maps and Amap are collected to verify the method proposed in this paper. The experimental results show that, compared with the existing POI matching methods, the accuracy and recall rate of the results obtained by the POI matching and fusion method proposed in this paper are significantly improved, which verifies the accuracy and feasibility of the matching.



Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 2 November 2024

Revised: 8 January 2025

Accepted: 10 January 2025

Published: 12 January 2025

**Citation:** Wang, Y.; Li, C.; Zhang, H.; Guo, B.; Wei, X.; Hai, Z. POI Data Fusion Method Based on Multi-Feature Matching and Optimization. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 26. <https://doi.org/10.3390/ijgi14010026>

**Copyright:** © 2025 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** POI; entity matching; spatial data integration; matching accuracy; machine learning

## 1. Introduction

In the field of geographic information science (GIS) and location-based services (LBS), point of interest (POI) has attracted much attention as an important data type for conveying geographic entity and location information [1]. POI data from different sources often have problems such as consistency, repeatability, fuzziness and incompatibility, and the accurate matching of these data is crucial to enrich and standardize the POI database and effectively reuse data [2].

GIS data integration is generally divided into two stages: matching and merging [3,4]. The matching stage determines whether different source POI represent the same geographical location by analyzing similar indicators [5], and the merging stage integrates different source data attributes to provide a unified and richer dataset. Among them, matching is the key and most challenging part of data integration. Due to significant differences in structure, content, coverage, and attributes (such as name and address) of POI data from different sources, accurate matching is extremely difficult [6].

Multi-source POI data fusion is the key to smart city research. POI matching methods based on feature similarity can be roughly divided into three categories: those based

on spatial attributes [7,8], those based on non-spatial attributes, and those based on the combination of spatial and non-spatial attributes [9,10]. Relying on a single attribute often leads to poor matching results, and methods combining spatial and non-spatial attributes are more commonly used.

The method based on spatial attributes relies on longitude and latitude information to identify the corresponding object, but the longitude and latitude of different source POI have problems such as error and coordinate system mismatch. Wu et al. used FME Server spatial location and gate address attribute information to fuse data [11]. Luo et al. used the Euclidean metric to determine the matching object [12]. Xu et al. adopted the mutual nearest neighbor algorithm [13]. Safra and Beeri obtained matching entities from three or more data sources through location-based link analysis algorithms [14,15]. Although the method is intuitive, there are substantial uncertainties due to the inconsistency of the coordinate systems of different data sources, the position error introduced by measurement technology, and the nonlinearity of data processing. The method based on non-spatial attributes does not need to consider the difference in latitude and longitude, but requires the POI data storage form of different sources to be relatively uniform, and the non-spatial feature attributes may be labeled incorrectly or information lost due to human factors. The calculation is often reflected by text similarity. For example, Zhang et al.'s experiments show that the edit distance method is the best among the first methods to filter name similarity [8]. Wang et al. segmented and extracted POI address information and updated the data by matching the address tree layer by layer [16]. Li et al. used a global clustering algorithm and global generation model to extract corresponding objects from fuzzy names [17]. Junchul proposed a graph-based method combining the language similarity of string and object name to match text similarity [18]. These methods, based on non-spatial attributes, are easily disturbed by human factors, resulting in defects in non-spatial features and insufficient accuracy.

Methods based on the combination of spatial and non-spatial attributes that integrate multiple attributes (such as name and spatial location) are more common in the field of POI matching and can be subdivided into rule-based and machine learning-based methods [19]. The rule-based approach combines spatial location with non-spatial properties by setting a set of rules. For example, Zhao et al. used the Dempster–Shafer evidence theory combined with an analytic hierarchy process to calculate attribute similarity score weights. Li et al. proposed an instance matching method that combines information entropy with heterogeneous attributes and allocates attribute weights reasonably, which can flexibly set thresholds to obtain corresponding objects with different confidence levels [20]. Zeng et al. used the relationship between human mobility and points of interest to test different weight values within the range of 0.1–0.9 to determine the optimal weight for POI name and spatial similarity [21]. This method does not need to annotate the data to train the model and only needs part of the data to adjust the model, but it has low flexibility and is difficult to apply to different datasets, and cannot dynamically adapt to the diversity and changes in datasets, which limits its application in complex scenarios. The machine learning-based approach takes the similarity scores of multiple attributes as input features and learns how to assign weights and determine classification boundary values to accurately predict whether POI data matches. For example, Xing et al. used TF-IDF and BERT to re-encode original attributes to improve attribute characteristics, and then built a binary classification model based on LightGBM to improve matching performance [22]. Cousseau and others detected duplicate locations with a deep learning model called PlacERN [23]; Piech et al. verified and compared the key components of POI matching, showing that the POI matching classifier with the best matching effect is the combination of random forest algorithm and missing data markers, as well as the mixing of different similarity measures

of different POI attributes [24]. This method does not need to define rules manually, and can automatically learn the weight and filtering threshold of POI attribute information. However, it usually needs to label a large amount of training data, increasing the cost of data preparation and limiting the ability of model generalization.

The above research on POI data matching has the following problems:

- (1) Existing methods mainly focus on the inherent properties of POI, while ignoring the spatial context of POI and human–environment interaction information, which is crucial to improve the accuracy and robustness of POI matching. Traditional methods for calculating text similarity (for example, the Levenshtein distance algorithm) can calculate structural similarity between strings, but do not take into account semantic relationships or contextual information between words. This limitation may affect the accuracy of string similarity matching.
- (2) Due to the limitations of the training data, the existing model may perform well on a specific dataset, but may experience performance degradation on a new dataset.
- (3) There are differences in latitude and longitude between POI data from different sources, mainly due to different coordinate systems and shift algorithms used by various data providers. These differences lead to spatial position errors and mismatches between coordinate systems.

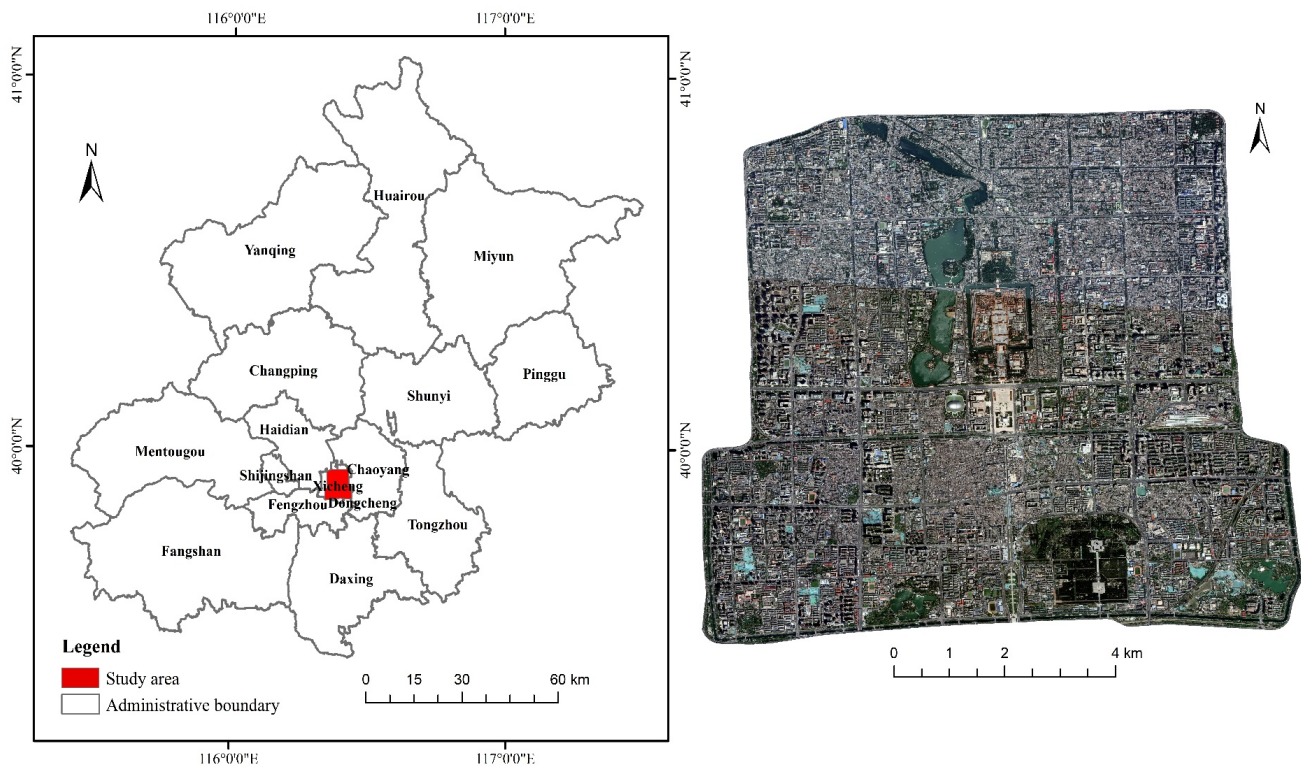
In summary, the focus of this study is to solve the limitations of existing POI matching methods in terms of their dependence on a single feature attribute and the integration of heterogeneous data from multiple sources. By proposing a geospatial entity matching method based on multi-feature-value computation and dynamically determining the information weight of each attribute through the random forest algorithm, the accuracy and recall of multi-source POI matching are effectively improved.

## 2. Materials and Methods

### 2.1. Study Area

The area within the Second Ring Road in Beijing, as one of the core areas of the capital city, surrounds iconic locations such as Tiananmen Square and the Forbidden City, with a total area of approximately 62 km<sup>2</sup>, as shown in Figure 1. The area within the Second Ring Road is not only of great geographical importance but is also a hub for political, cultural, and commercial activities, possessing unique geographical advantages and profound historical and cultural heritages. In this area, rich historical sites, modern commercial centers, and residential areas are integrated, making it not only the urban heart of Beijing but also an ideal case for studying China's rapid urbanization process and the evolution of urban functions.

Currently, most studies on the fusion of multi-source POI data focus on the central urban areas with relatively strong economic strength. Thanks to its unique historical background, complex functional zoning, and highly developed service industry, the area within the Second Ring Road in Beijing provides an excellent sample for data fusion and analysis. Therefore, this study selects the area within the Second Ring Road in Beijing as the research area.



**Figure 1.** Study area.

## 2.2. Data Source and Processing

This study selects the area within the Second Ring Road in Beijing as the research area, aiming to verify the accuracy and effectiveness of the multi-feature POI fusion method. In this study, rich multi-source POI data were obtained from Amap (<https://www.amap.com/>) on 1 November 2022 and Tencent Maps (<https://map.qq.com/>) at 10 May 2023, and the building outline data of OpenStreetMap (OSM) was also introduced. In order to compare the impacts of different data sources on the identification of urban functional areas and provide stable data support for experimental verification, within this area, the POI data of Amap was used as the reference dataset and compared and analyzed with the POI data of Tencent Maps. Among them, Amap provided 73,585 POI data points, and Tencent Maps provided 81,994 POI data points. These POI data points cover multiple categories such as buildings, catering services, companies, schools, hospitals, shopping services, and government institutions, comprehensively reflecting the distribution of urban functional areas in the Dongcheng District. In addition, OSM provided 31,491 building outline data points, which further enhanced the accuracy of functional area identification and provided detailed geographical structure support for subsequent spatial analysis.

Through the open API interfaces provided by map service providers, a total of 155,579 POI data points were obtained in this study. Considering the problems existing in the original data, such as data duplication, missing attribute information, and inconsistent coordinate systems, the following detailed data preprocessing steps were carried out in this paper:

### (1) Data cleaning

Firstly, duplicate removal was performed on the original data to ensure that each POI had only a unique record in the dataset. By comparing the names, addresses, and coordinates of POIs, redundant duplicate data were removed to improve the quality and precision of the data.

Missing value handling: Data with missing key attributes (such as addresses, coordinates, etc.) were deleted to ensure the integrity of the data. Especially in the case of lacking location information, these data were regarded as invalid records to avoid causing interference in subsequent analyses.

## (2) Coordinate system standardization

Since both Amap and Tencent Maps use the GCJ-02 coordinate system, while the WGS-84 coordinate system is more widely used in global positioning and analysis, all POI data were uniformly converted to the WGS-84 coordinate system through their respective API interfaces in this paper. During the coordinate conversion process, precise conversion algorithms were adopted to ensure the spatial positions of the converted data were accurate. The standardized coordinate data provided a reliable basis for subsequent spatial calculations and analyses.

## (3) Attribute information improvement

During the data cleaning process, apart from removing missing or incorrect records, certain supplementary work was also carried out through other data sources in this paper. For some POIs, especially those that only contained names but lacked detailed address information, their complete attribute information was restored as much as possible by querying their actual positions and attributes in relevant map services.

### 2.3. POI Matching Method Considering Multi-Feature Similarity

#### 2.3.1. Name Similarity Calculation

Name similarity calculation plays an important role in accurately identifying the corresponding POI across multiple data sources. Because POI names are usually expressed as string data, subtle differences in names, such as abbreviations, alternative spellings, or typographical errors, can easily lead to mismatches or missing matches. In order to solve this problem, this paper proposes an improved mixed similarity method, which combines character similarity and semantic understanding to calculate name similarity.

#### (1) Data Preprocessing

Before performing similarity calculation, it is necessary to preprocess the name field, including normalizing all names and standardizing the data; remove unnecessary spaces and redundant characters, and eliminate inconsistency and noise; standardize the format of the name to ensure the consistency between all items, thus reducing the calculation errors caused by format differences. In order to effectively reduce the computational complexity and avoid unnecessary recalculation or mismatch, a Levenshtein edit distance algorithm is introduced in this paper. This algorithm can quickly identify and eliminate obviously different POI entities. Through the preprocessing step, the computational complexity and time consumption can be significantly reduced while retaining key information.

#### (2) Embedding Vector Acquisition

To capture the semantic relationships between names, we use the pre-trained multi-lingual BERT model to convert names into fixed-dimension embedding vectors. This method ensures that context and meaning are considered in addition to string-based comparison. For example, POIs named "Café" and "Coffee Shop" may have low string similarity but high semantic similarity, which can be captured using BERT embeddings.

#### (3) Cosine Similarity and Hybrid Metric

Once the embedding vectors are generated, the cosine similarity metric is used to compare the names. However, to address the specific challenge of handling directional and numerical components (e.g., "Block A" and "Block 2"), lexical annotation is introduced.

Names are broken down into meaningful components, and a weighted similarity score is applied to these components to refine the final similarity result.

#### (4) Inverse Density Weighting

In order to balance the influence of frequent names, the inverse density weight is applied. For example, common names like “Park” may appear frequently, thus reducing their distinguishing ability. Inverse density weighting ensures that rare names have a greater impact on the overall similarity score, thus improving the accuracy of identifying unique entities.

Define a set  $N$  of place names, where each place name occurs with a frequency of  $f(n)$  (where  $n \in N$  represents a place name) and the density weight inverse  $W(n)$  is shown in Formula (1).

$$W(n) = \frac{1}{f(n) + a} \quad (1)$$

where  $W(n)$  is the inverse density weight of the name  $n$ , and  $f(n)$  is the frequency of the name  $n$ , which is a constant to prevent the denominator from appearing to be zero ( $a = 0.01$ ).

#### (5) Final Similarity Calculation

The final name similarity score is calculated by recombining character-based, semantic, and inverse density weights into a mixed similarity function. The calculation formula of comprehensive similarity is shown in Formula (2).

$$S_{name}(i, j) = \alpha \times S_{char} + \beta \times S_{sem} + \gamma \times W(n) \quad (2)$$

Among them,  $S_{char}$  is a score based on character similarity,  $S_{sem}$  is a score based on BERT embedding, and  $W(n)$  is an inverse density weight.  $\alpha$ ,  $\beta$ , and  $\gamma$  are the weight coefficients based on the similarity, semantic similarity, and inverse density of characters, respectively. According to their importance, the values of  $\alpha$ ,  $\beta$ , and  $\gamma$  calculated by AHP are 0.577, 0.192, and 0.115, respectively. This hybrid method allows the system to better handle small spelling changes and meaningful semantic differences, ensuring a more accurate matching process.

### 2.3.2. Address Similarity Calculation

In the process of multi-source POI data fusion, compared with the name attribute, the matching ability of the address attribute is relatively low. In order to effectively identify addresses with similar or identical meanings and conduct data fusion, this study adopts the Jaccard similarity calculation method to evaluate the similarity between addresses. Specifically, the steps for calculating address similarity include: Firstly, read the POI data from different data sources and extract the address information. Then, for each pair of addresses, use the Jaccard similarity calculation function to calculate the similarity score between them. The Jaccard similarity measures the similarity between two addresses based on the proportion of items (for example, words, characters) that appear simultaneously in the two addresses. Specifically, this method calculates the ratio of the intersection and the union of the two address sets. The larger the ratio is, the more similar the addresses are. This method can effectively identify and handle similar addresses in multi-source POI data, thus achieving the purposes of data fusion and duplicate data removal.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

where  $A$  and  $B$  represent the address information in the Amap and Tencent POI datasets, respectively.  $J(A, B)$  is computed based on the ratio of the intersection and concatenation

of the address texts, which is used to evaluate the similarity of the address information between the two datasets.

### 2.3.3. Spatial Similarity Calculation

In the process of multi-source POI data fusion, fully exploiting the spatial attributes of POI data is crucial for accurate identification. Given that POI data are mainly presented as spatial point objects, this paper selects the Euclidean metric method to calculate the similarity of spatial point pairs. Firstly, for two points,  $(x_i, y_i)$  and  $(x_j, y_j)$ , calculate the Euclidean distance between them according to Equation (4) in the projected coordinate system.

$$d(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (4)$$

The closer the calculated distance is, the more similar the distribution of the POIs represented by the two points in space will be, and the higher their similarity will be. Meanwhile, the longitude and latitude are used to conduct unified distance calculation, and the calculation results are normalized to ensure that the spatial distances under different scales are comparable, thus laying a solid foundation for the subsequent fusion and analysis of multi-source POI data and enabling spatial information to assist in the accurate identification of POIs more effectively.

### 2.3.4. Dynamic Feature Weighting with Random Forest

Random forest is a machine learning algorithm based on the integration of decision trees and has a relatively strong feature selection ability. In this paper, the random forest model will automatically generate the importance weights of each feature through the analysis of multiple features. This method no longer relies on a single fixed threshold but dynamically adjusts the weights according to the characteristics of the actual data, thereby improving the matching accuracy. The feature weighting method is divided into the following steps:

- (1) Training Set Construction: Firstly, the POI dataset is divided into a training set and a testing set for model training and verification, respectively. In the training set, the model learns the importance of each feature through the actual matching situations. We use 70% of the POI data for training and 30% for testing to ensure the stability and accuracy of the model.
- (2) Feature Selection: The model takes multiple features into account, such as name, address, coordinates, etc. Each feature will be assigned a weight according to its contribution to the matching result, and the calculation of the weight is based on the impact of the feature on the final matching result.
- (3) Weight Calculation Formula: For each feature, its weight is obtained through feature importance assessment, and specifically, the "Mean Decrease Accuracy (MDA)" is used to measure the feature's contribution to the classifier's performance. The weight calculation formula is as follows:

$$MDA(v) = \frac{1}{nTrees} \sum_{t=1}^{nTrees} (errOOB_t - errOOB'_t) \quad (5)$$

where  $nTrees$  represents the number of decision trees,  $errOOB_t$  is the out-of-bag error of decision tree, and tree  $T$  and  $errOOB'_t$  are the out-of-bag errors of the out-of-bag sample data after random perturbation. The larger the MDA value is, the higher the importance of the feature in the matching process.

- (4) Model Parameter Selection: The key parameters of the random forest model include the number of decision trees ( $nTrees$ ) and the number of features available when each

tree splits nodes ( $mtry$ ). To ensure the stability and accuracy of the model, this paper selects the settings of  $nTrees = 600$  and  $mtry = 3$ . These parameter values are adjusted through the out-of-bag estimate error (OOB Error) and can achieve good performance in practical applications.

- (5) **Dynamic Feature Weighting:** By calculating the importance of features, the model will assign corresponding weights to each feature according to the actual data situation. These weights are applied in the similarity calculation to automatically adjust the influence of each feature on POI matching. The final POI similarity calculation formula is as follows:

$$Similarity\ Score = \sum (Weight_i \times Feature\ Similarity_i) \quad (6)$$

where  $Weight_i$  represents the weight of a specific feature, and  $Feature\ Similarity_i$  represents the similarity score of that feature (such as name similarity, address similarity, etc.), and this value is usually normalized to a range between 0 and 1. The final POI similarity score is obtained by a weighted summation of the similarities of various features.

### 3. Experiment and Analysis

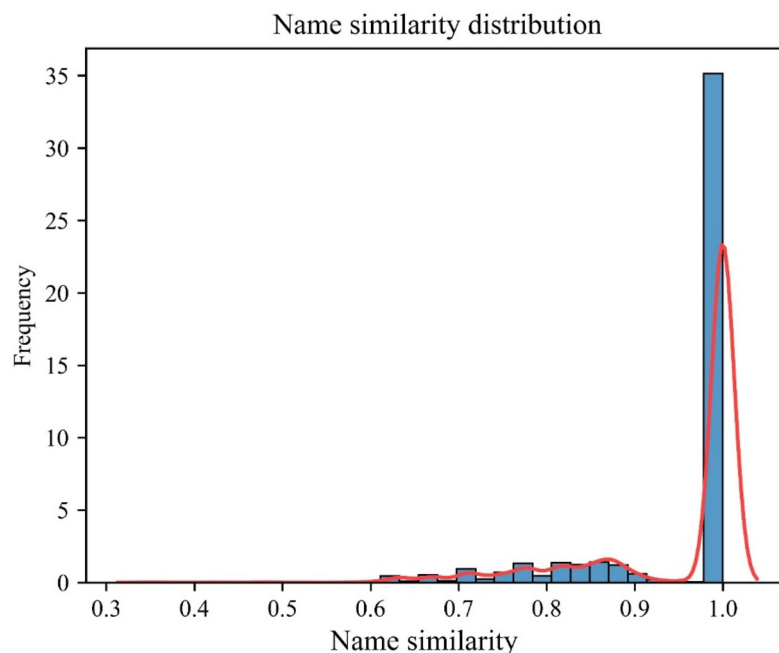
#### 3.1. Feature Similarity Calculation

##### 3.1.1. Name Similarity

This paper takes the calculation of name similarity as the primary task of feature similarity. Firstly, the Levenshtein edit distance algorithm is used to conduct a preliminary screening on matching pairs to quickly eliminate obviously different names, thus reducing the computational burden. After the preliminary screening, the remaining 43,242 pieces of data will be further processed. To improve the accuracy of similarity calculation, this paper introduces lexical tagging, which effectively reduces the errors caused by specific words (such as numbers and directional words) and significantly enhances the calculation effect of name similarity. The processed names are transformed into semantic vectors through the pre-trained multilingual BERT model and combined with the cosine similarity function to quantify the similarity values between names. This embedding method can capture the context and semantic information of names, enabling names with similar semantics but obvious character differences (such as “café” and “coffee shop”) to obtain relatively high similarity scores as well.

The experimental results show that when the lexical tagging is the same, the cosine similarity values of names are mostly close to 1, indicating a high semantic consistency among place names. For example, the similarity between “Library A” and “Library B” is close to 1, demonstrating their high relevance at the semantic level, while when the lexical tagging is different, the cosine similarity values are relatively low. For instance, the similarity between “North Road of the Park” and “South Road of the School” is obviously low, reflecting the significant semantic differences among place names. The distribution of name similarity shown in Figure 2 further verifies this point. The name similarities of most points of interest are concentrated between 0.6 and 1, showing a relatively high consistency. This indicates that after a series of processing, the calculation results of name similarity are more in line with the actual situation and can effectively capture the similarity relationships between names.

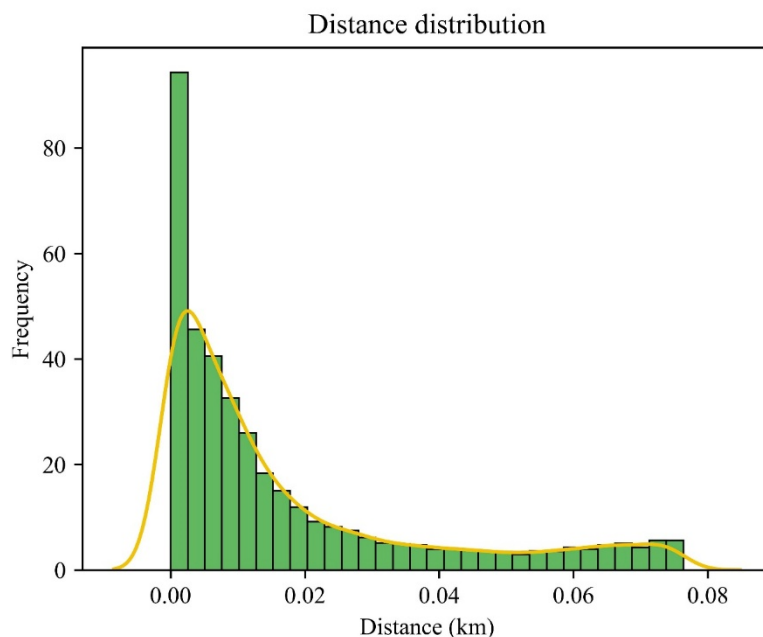




**Figure 2.** Distribution of name similarity calculation results.

### 3.1.2. Spatial Similarity

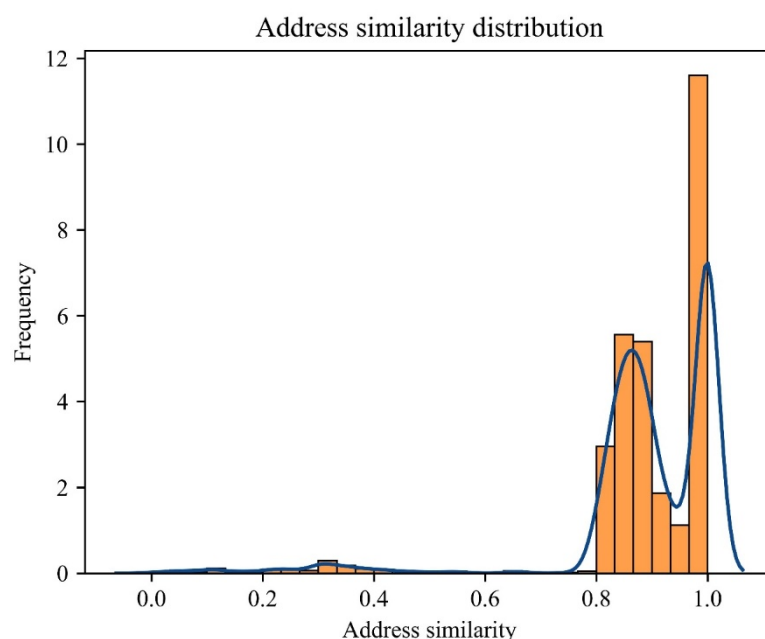
Since both the Amap Map and the Tencent Map use the GCJ-02 coordinate system, which is obtained through encryption from the WGS-84 coordinate system, this encryption algorithm can cause nonlinear shifts in data coordinates. To enhance the usability of spatial attributes, this paper employs the APIs provided by both map service providers for coordinate transformation (Tencent API: <https://cloud.tencent.com/document/product/1301/68448> (accessed on 1 November 2024); Amap API: <https://lbs.amap.com/api/android-sdk/guide/computing-equipment/coordinate-transformation> (accessed on 1 November 2024)) to uniformly convert all data into WGS-84 coordinates. The spatial distances are then calculated using Equation (4), and the results of these calculations are shown in Figure 3.



**Figure 3.** Distribution of spatial distance density.

### 3.1.3. Address Similarity

The results of quantifying the address similarity by using the Jaccard similarity function show that when the similarity of address words is high, the Jaccard value is close to 1, which reflects the high overlap of content. In contrast, Jaccard values with low similarity reveal significant differences (see Figure 4). The experiments show that the Jaccard similarity of most addresses is between 0.5 and 1, indicating that this method can distinguish semantic differences effectively. Compared with traditional methods, Jaccard similarity performs better in addressing specific lexical differences, significantly reducing matching errors. It improves the accuracy and reliability of address matching by quantifying the overlap degree of the word collection, and provides support for the high-quality data analysis of GIS and location services.



**Figure 4.** Distribution of address similarity calculation results.

## 3.2. Model Accuracy Analysis

### 3.2.1. Overall Matching Performance Evaluation

The method proposed in this study performs excellently in terms of overall matching performance. Through a rigorous data processing flow, including data cleaning (removing duplicates and supplementing missing key attributes), coordinate system conversion, and the improvement of attribute information, the enhancement of data quality is ensured, laying a solid foundation for subsequent precise matching.

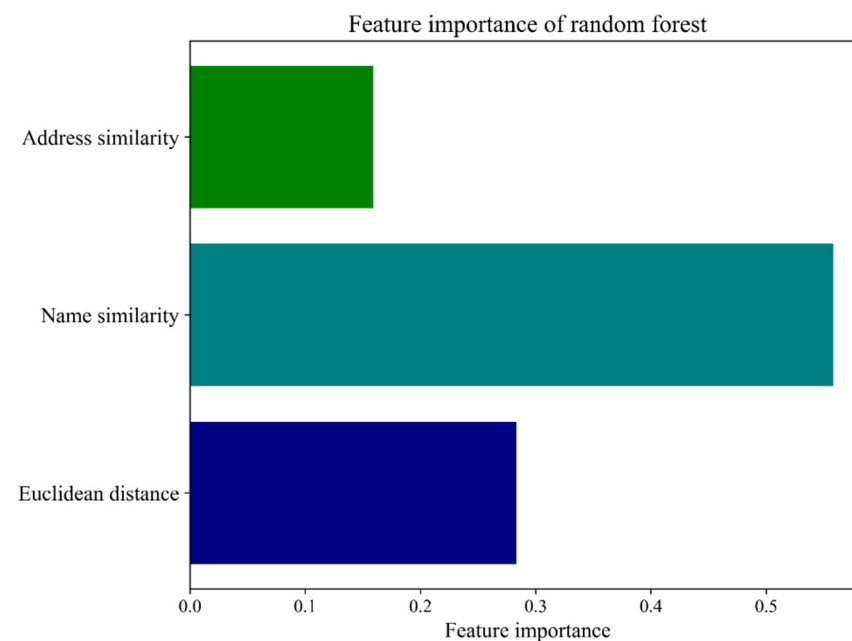
In the actual matching experiment, the performance of the method in this paper on the test set is satisfactory. Specifically, the precision reaches 0.983, which indicates that among all the POI pairs judged to be matched, approximately 98.3% of the POI pairs are truly matched, demonstrating the high accuracy of the model in matching judgment. The recall rate is 0.978, meaning that the model can successfully identify more than 98.0% of the actual matching POI pairs, with an extremely low omission rate, indicating that the model exhibits excellent capabilities when dealing with POI pairs that actually have a matching relationship. The F1 score is 0.980. As a comprehensive measurement indicator of precision and recall, this value further verifies that the model has achieved an excellent balance between the two.

In conclusion, the method in this paper performs outstandingly in key indicators such as precision, recall, and F1 score, fully demonstrating its powerful ability in multi-

source POI data matching. This method can not only provide high-quality matching results for geospatial data integration but also offer reliable data support for the subsequent applications of POI data, ensuring the effective fusion and precise application of the data.

### 3.2.2. The Influence of Feature Similarity on Matching Results

In the process of POI (point of interest) matching, name similarity, address similarity, and spatial distance similarity are three important indicators which affect the accuracy and efficiency of matching to varying degrees. According to the feature importance values in the output results shown in Figure 5, the name similarity is 0.558, the address similarity is 0.159, and the spatial distance similarity is 0.283. These values reflect the degree of dependence and importance of different features in the POI matching task for the random forest model.



**Figure 5.** The importance of different features in random forest models.

Firstly, the importance value of name similarity reaches 0.558, indicating that the name plays a dominant role in POI matching. As a key factor in identifying locations, the name is usually an important clue for judging whether two POIs match or not. When the names of two POIs are highly similar or exactly the same, they are very likely to represent different records of the same geographical location. For example, in commercial POIs, the names of different stores of the same chain brand are often the same or similar. Through the calculation of name similarity, the records of these stores in different data sources can be quickly and accurately identified. However, name similarity is not absolutely reliable. Common names (such as “park”, “street”, etc.) frequently appear in different locations, which may lead to misjudgments. To solve this problem, strategies such as lexical tagging, semantic vectors of the BERT model, and inverse density weighting have been introduced in the study to improve the accuracy of name similarity calculation.

Secondly, the importance value of address similarity is 0.159, which is relatively low. This is mainly due to data quality problems in the address field, where there are a large number of irregular inputs, spelling mistakes, or inconsistent formats. Even if the actual geographical locations are the same, these problems may lead to a low calculated address similarity, weakening its effectiveness in matching. In addition, data from different sources may adopt different ways of describing addresses, which makes the address representations of the same location vary greatly. Although name similarity has already included some

location information to a certain extent, in some specific cases, such as when the names are similar but the spatial distances are far apart, address similarity can serve as an auxiliary judgment basis to help exclude some incorrect matches.

Finally, the importance value of spatial distance similarity is 0.283, which is between the name similarity and the address similarity. As an objective physical indicator, spatial distance plays an important role in POI matching. In an urban environment, the distances between buildings are usually small. When the spatial distances of two POIs are very close, they are very likely to be different records of the same location. For adjacent stores or different merchants in the same building, their spatial distances are close. Through the calculation of spatial distance similarity, the model can be effectively assisted in making matching decisions. A large spatial distance can almost directly rule out the possibility of two POIs matching. Therefore, spatial distance not only helps to confirm the possibility of close matching but also can quickly screen out POI pairs that are impossible to match due to long distances, thus improving the overall efficiency and accuracy of the model.

In conclusion, name similarity, address similarity, and spatial distance similarity each play different roles in POI matching. By comprehensively considering these indicators, the accuracy and efficiency of POI matching can be effectively improved, providing more precise data support for urban management and services.

### 3.2.3. Comparative Analysis of Different Matching Methods

To comprehensively evaluate the effectiveness of the proposed method, this paper conducts a detailed comparison using the same dataset with several other common POI matching methods, including the multi-feature similarity calculation method based on a fixed threshold [25] and the matching method considering multiple constraints [26]. In the comparative experiment, we evaluated different methods from multiple key indicators, and the specific results are shown in Table 1:

**Table 1.** Comparison of POI matching performance.

Method	Number of Successful Matches	Number of False Matches	Number of Missing Matches	Precision	Recall	F1 Score
Based on fixed thresholds	13,591	3527	2691	79.4%	83.4%	0.814
Multiple constraint based	12,895	487	323	96.4%	97.5%	0.969
Methodology of this paper	13,100	240	290	98.1%	97.6%	0.979

The table data reveals significant differences in POI matching performance across methods. The fixed threshold-based approach successfully matched 13,591 POIs but exhibited a high number of false matches and missing matches, leading to lower precision and recall. Its final F1 score of 0.814 indicates that this method is limited when dealing with complex matching scenarios.

In contrast, the multiple constraint-based method significantly reduced false and missing matches, achieving a precision of 96.4% and recall of 97.5%, with an F1 score of 0.969. This demonstrates that introducing multidimensional constraints can significantly enhance the reliability and accuracy of matching. However, this method may still face unoptimized issues in certain cases.

The methodology proposed in this paper further improves matching performance, successfully matching 13,100 POIs with only 240 false matches and 290 missing matches. It achieved a precision of 98.1%, a recall of 97.6%, and an F1 score of 0.979. This highlights the method's outstanding ability to optimize false and missing matches, particularly in complex

scenarios, by integrating semantic similarity, spatial distance, and dynamic feature weight allocation. These factors collectively enhance both the accuracy and stability of matching.

In summary, the proposed method outperforms other methods across multiple performance metrics. Its significant reduction in false and missing matches underscores its innovation and applicability, offering a more reliable solution for complex POI matching tasks. This efficient matching mechanism not only improves the precision of multi-source data integration but also provides a robust data foundation for GIS and related fields.

### 3.3. Matching Quality Evaluation

#### 3.3.1. Matching Effect of Different Geographical Regions

In the central urban areas, especially in areas with frequent commercial activities, points of interest (POIs) are densely distributed. Despite the rich data volume, the actual data fusion faces numerous challenges. Taking the Beijing Department Store area as an example, Tencent Map records 1082 POIs, while Amap has 744. However, the preliminary matching only yields 46 results. Although the accuracy rate is 100%, the overall effect is not satisfactory. This is mainly due to the significant differences in data details and store recording methods between the two data sources.

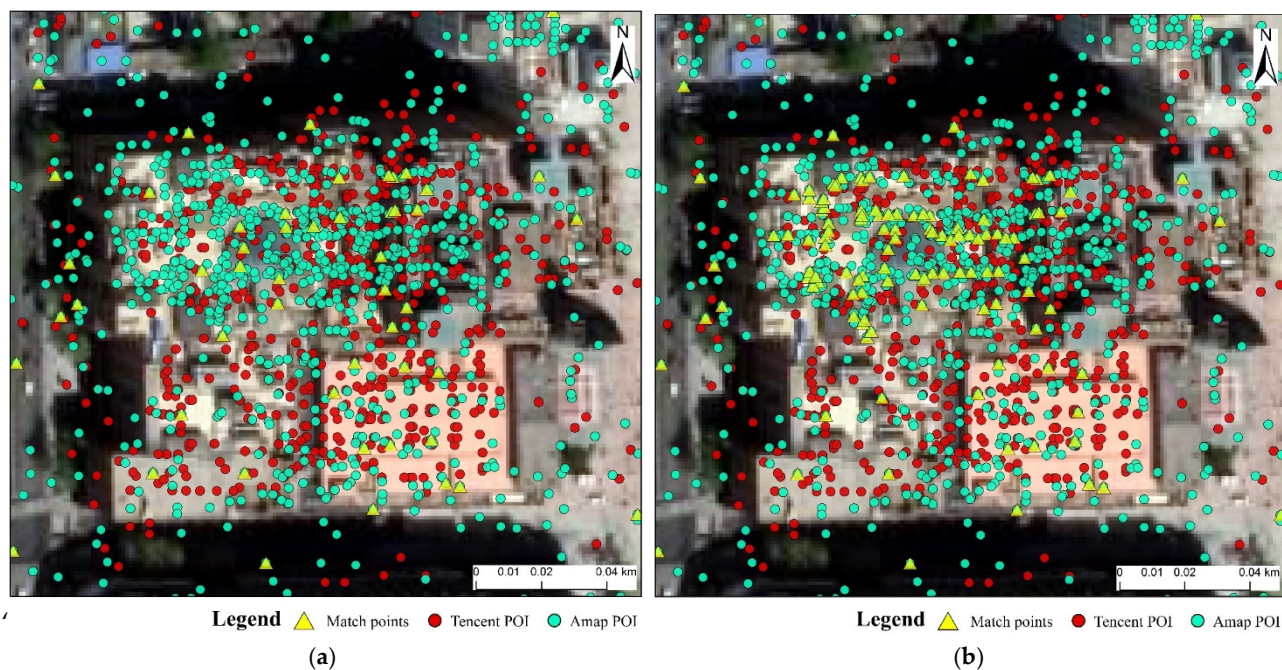
For example, Amap uses mall serial numbers (such as “001 (Beijing Department Store)”) instead of specific store names, while Tencent Map uses specific store names (such as “Fashion Women’s Clothing Store”). The difference in naming methods complicates the calculation of name similarity and hinders the accuracy of preliminary matching. In addition, the presence of non-commercial data points such as escalators and rest areas also increases the difficulty of data processing and interferes with the matching algorithm, reducing the number of effective commercial POI matches.

Single-feature matching methods perform poorly in this scenario. Name similarity cannot effectively distinguish stores due to the existence of serial numbers and irregular names. Relying solely on spatial distance or address similarity often leads to overly broad or repeated results, thus causing misjudgments.

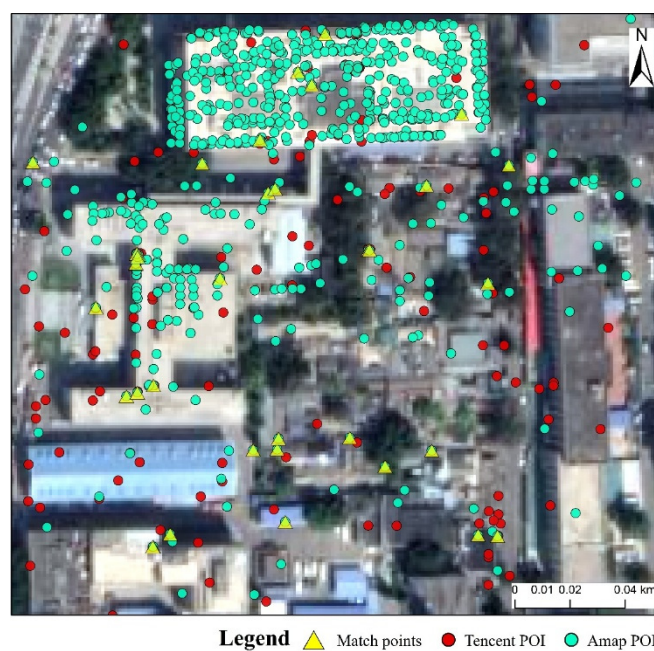
To address the above problems, this study proposes a multi-feature similarity method that comprehensively calculates name, address, and spatial distance. Meanwhile, it is recommended to introduce a mall-store distribution map during the matching process to more accurately identify the positions of stores and their relative relationships within the same mall. Through this method, the number of matching items in this area has increased from 46 to 108, significantly improving the matching accuracy and algorithm adaptability in complex scenarios (see Figure 6).

In the POI matching of public service facilities in the core urban areas (such as schools and hospitals), this method also performs remarkably well. The addresses of these facilities are usually accurate and their spatial positions are fixed. For example, the address of a school is unique and unchanged. By accurately matching the address and combining with spatial distance judgment, it is possible to accurately identify the same school across different data sources.

Taking Peking University First Hospital as an example, Tencent Map provides 221 POIs, and Amap provides 691. Eventually, 40 precise matches are achieved, with an accuracy rate of 100% (see Figure 7). This verifies that this method can utilize address and spatial data to achieve the precise matching of public service facility POIs.



**Figure 6.** Before and after commercial POI optimization: (a) is the matching result before optimisation; (b) is the matching result after optimisation.



**Figure 7.** Map of POI matching results around Peking University First Hospital.

Within the Second Ring Road in Beijing, although the POI data are dense, there are data quality problems in the matching of residential communities and building numbers. Some building numbers in residential communities are only labeled as “Building 1” or “Unit 2”. However, this method can still achieve effective matching by using spatial distance information and a dynamic feature-weighting strategy, especially for spatial distance. In a local residential community matching test, Tencent Map has 428 POIs, and Amap has 302. Eventually, 92 matches are obtained, with an accuracy rate of 100% (see Figure 8). Even in the face of non-standard or incomplete data, this method can compensate by adjusting weights to ensure high-precision and stable overall matching results.



**Figure 8.** The matching results of the Fahua South Lane Subdivision.

### 3.3.2. Matching Effect of Different POI Types

In multi-source data fusion, different types of POIs have different matching performance due to their own characteristics, which affects the accuracy and effectiveness of data fusion. For this reason, for the six types of POIs, namely, shopping, transportation facilities, infrastructure, construction properties, tourist attractions, and companies, 300 samples are randomly selected (to ensure that the number of samples is sufficient and statistically significant), and the following matching accuracy data are obtained (Table 2).

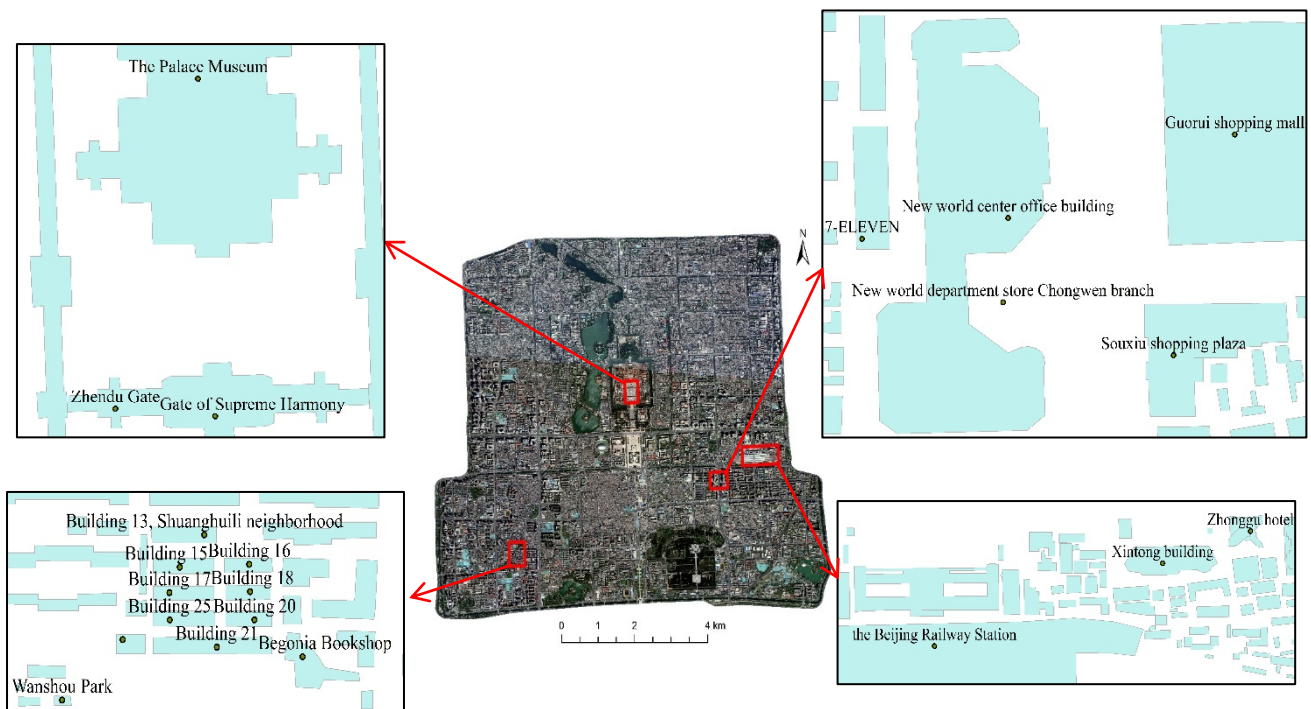
**Table 2.** Detailed table of matching performance of different types of POI samples.

POI Category	Number of Samples	Number of Correct Matches	Number of Incorrect Matches	Accuracy Rate
Accuracy rate	300	290	10	96.67%
Transportation facilities	300	294	6	98.0%
Infrastructure	300	297	3	99.0%
Building properties	300	300	0	100%
Tourist attractions	300	300	0	100%
Companies and enterprise	300	300	0	100%

From the data, it can be seen that the accuracy of POIs in the building properties, tourist attractions, and companies and enterprises categories is 100%, while the infrastructure category is 99.0%, transportation facilities is 98.0%, and the shopping category is 96.67%.

In the shopping POI category, large chain brands have a significant matching effect due to high name consistency, but the emerging formats still provide important support for business region data due to changing names and large record differences. The transportation facilities' POI matching relies on significant spatial characteristics. Large hubs (such as airports) are easily matched by spatial distance, while parking lots are matched accurately by combining space and address information. Although reconstruction and data lag may bring about certain impacts, spatial characteristics still ensure good performance. Due to fixed locations and standardized information, infrastructure POI shows a superior matching effect through accurate address and spatial similarity calculation, providing reliable support for urban management and planning. Construction, real estate and tourist attractions' POI show a high degree of information consistency in multi-source data due to

clear geographical locations and high name recognition, and achieve efficient matching. Corporate POI also has good matching characteristics due to the standardization of business activities and clear registered addresses. Overall, the method integrates multi-feature similarity calculation to effectively deal with complex features of different POI types, especially in complex data environments, showing excellent stability and accuracy, and providing innovative support for multi-source data fusion. Overall, the method integrates multi-feature similarity computation and can effectively handle complex features of different POI types, especially in complex data environments, such as the various scenarios covered in Figure 9, demonstrating excellent stability and accuracy and providing innovative support for multi-source data fusion.



**Figure 9.** An example of a matching result.

In short, different types of POI matching have their own advantages and challenges. In commercial POIs, the brand advantage makes some types of matching more effective, but there are still difficulties in matching emerging businesses. Transportation facility POIs, due to the influence of urban construction, are able to ensure higher matching accuracy by virtue of their obvious spatial characteristics. Infrastructure, construction, real estate, tourist attractions, and corporate POIs are usually able to achieve excellent matching results due to their stable and standardized information.

#### 4. Discussion

This study proposes a multi-source POI data fusion method based on multi-feature similarity calculation, which demonstrates significant advantages in improving the matching accuracy, precision, and computational efficiency of POI data. It provides an efficient and accurate solution for geospatial data integration and POI data applications.

Compared with traditional matching methods based on fixed thresholds and multiple constraints, this method performs better in terms of accuracy, recall rate, and F1 score (see Table 2). For example, when matching commercial POIs, by comprehensively considering features such as name, address, and spatial location, it can accurately identify similar commercial venues in different data sources, enhancing the integrity and accuracy of



data. Taking public service facility POIs as an example, by introducing lexical tagging and semantic vectors of the BERT model to improve the calculation of name similarity, and combining these with the calculation of address and spatial similarity, the matching accuracy is significantly improved.

In urban core areas and major functional areas, this method exhibits high matching accuracy and integrity, and can accurately reflect the commercial layout and other aspects. However, in remote areas or regions with lagging data updates, due to the insufficient coverage of POI data or untimely updates, there are still a small number of matching problems.

Compared with traditional methods based on fixed thresholds, this method has obvious advantages. In terms of feature utilization, it comprehensively considers multiple features such as space, text, and semantics, overcoming the problem of inaccurate matching caused by traditional methods relying on a single feature or some features. For instance, traditional methods based on spatial attributes have limitations when dealing with issues like coordinate system mismatch. This method combines non-spatial attributes, enhancing the reliability of matching. Innovatively, the random forest algorithm is introduced for dynamic feature weighting, avoiding the shortcomings of relying on fixed thresholds or manual weight setting. The model can automatically adjust weights according to the characteristics of the data, improving flexibility and adaptability. For example, when the similarity of POI names is low but the address and spatial distance are similar, it can automatically increase the weight of spatial features, optimizing the matching results and reducing missed matches.

In terms of computational efficiency, optimization means such as using cosine similarity for the calculation of embedded vectors and spatial indexing techniques (such as KD-trees) are adopted, significantly reducing the computational burden. Spatial indexing techniques store and query spatial data in a structured way, improving computational efficiency and avoiding the computational bottleneck caused by large amounts of data in traditional methods. The introduction of cosine similarity accelerates the calculation of text and semantic similarity, especially when dealing with long texts or high-dimensional features, where its computational advantages are more prominent. In a big data environment, the computational efficiency of this method is higher than that of traditional methods with high computational complexity (see Table 3).

**Table 3.** Comprehensive comparison table of POI matching method performance.

Method	Precision	Recall	F1 Score	Time Complexity
Levenshtein distance	86.30%	87.10%	0.867	$O(n^2)$
Mutual nearest neighbor	84.50%	85.20%	0.848	$O(n \log n)$
Proposed hybrid similarity	98.8%	99.4%	0.991	$O(n \log n)$

Although this study has achieved good results, there are still some deficiencies. The quality and update frequencies of different data sources vary, which may lead to deviations in the fusion results. In the future, it is necessary to explore more refined data quality management strategies and real-time data update mechanisms. The algorithm also needs to be optimized and expanded to adapt to more complex geographical environments and diverse POI data types. Additionally, property data released by government departments can be introduced to improve residential data.

Against the backdrop of accelerating urbanization, accurate POI data are of crucial importance. This method has significant advantages in matching multiple POI data types and different regions. It surpasses traditional methods in terms of precision, recall rate, computational efficiency, flexibility, and adaptability. Although there are challenges in

handling POI matching in complex commercial areas such as shopping malls, it is expected that by introducing more data sources and optimizing the algorithm, the matching accuracy can be further improved. This provides precise data support for smart city construction and intelligent commercial services, promoting the development of technological innovation and application in the field of geospatial information science, and possessing both theoretical and practical value.

## 5. Conclusions

This study proposes an innovative POI matching and optimization method based on multi-feature value computation, aiming to address the accuracy issues of existing POI matching technologies. By integrating an improved hybrid similarity method, Jaccard method, and Euclidean distance metric, this study achieves the comprehensive calculation of multi-dimensional features such as name, address, and spatial distance, effectively improving the accuracy and efficiency of multi-source POI data fusion. In name similarity calculation, we introduced vocabulary tagging and pre-trained BERT models, combined with cosine similarity and inverse density weighting strategies, to effectively overcome the limitations of traditional methods in handling complex name matching. In address similarity calculation, we utilized the Jaccard method to accurately identify similar addresses, thereby improving matching accuracy. In spatial similarity calculation, we employed the Euclidean distance metric to calculate spatial distances, ensuring the precision of spatial matching.

Specifically, this study applies the random forest algorithm to dynamically adjust the weights of various features. This innovative method breaks through the traditional POI matching technology based on fixed thresholds, enabling the automatic adjustment of feature weights according to the actual characteristics of the data. It flexibly adapts to changes in different data sources and regions, improving both matching accuracy and efficiency while reducing computational resource waste.

Through experimental verification in the Second Ring Road area of Beijing, our research method has shown excellent performance in several key metrics. The experimental results indicate a precision of 0.983, a recall of 0.978, and an F1 score of 0.980, significantly outperforming existing mainstream POI matching techniques. These results fully demonstrate the effectiveness of the proposed method in multi-source POI data fusion, especially in handling large-scale, heterogeneous, and complex geospatial datasets, where it exhibits significant advantages.

The innovative method in this study provides effective data support for smart city development and offers new insights for geographic information service optimization and resource allocation. With the acceleration of urbanization, the application scenarios for POI data are gradually expanding, and our method has broad prospects in areas such as traffic management, public services, and urban planning. Specifically, in navigation systems, location recommendations, and urban functional zoning, the results of this study can effectively enhance the intelligence level of systems and improve user experience.

However, despite the positive experimental results obtained in this study, there is still some room for improvement. First, the difference in the quality and update frequency of different data sources may affect the fusion accuracy, and the matching method based on the real-time update mechanism will be explored in the future; second, the computational efficiency of the algorithm can be further optimized with the expansion of the dataset size, and it is proposed to combine the advanced technology to improve the accuracy and efficiency to cope with the complex environment and diverse data types in the future. In summary, this study opens up a new direction for POI matching technology and multi-source geospatial data fusion application, and we will continue to explore ways to improve

the accuracy and efficiency in the future, so as to promote the continuous development of related fields.

**Author Contributions:** Conceptualization, Yue Wang; methodology, Cailin Li; software, Yue Wang; validation, Cailin Li and Baoyun Guo; formal analysis, Xianlong Wei; investigation, Zhao Hai; data curation, Yue Wang; writing—original draft preparation, Yue Wang; writing—review and editing, Cailin Li and Hongjun Zhang; visualization, Yue Wang; funding acquisition, Cailin Li and Hongjun Zhang. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Shandong Provincial Natural Science Foundation and the Open Fund of Hubei Luoja Laboratory, under grant numbers No. ZR2022MD039 and No. 230100026.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

## References

- Zhang, J.Q.; Shi, W.B.; Xiu, C.L. Urban Research Using Points of Interest Data in China. *Geogr. Sci.* **2021**, *41*, 140–148.
- Xue, B.; Zhao, B.Y.; Li, J.Z. Methods for evaluating and improving the quality of POI data in geographic big data. *J. Geogr.* **2023**, *78*, 1290–1303.
- Ruiz, J.J.; Ariza, F.J.; Ureña, M.A.; Blázquez, E.B. Digital Map Conflation: A Review of the Process and a Proposal for Classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [[CrossRef](#)]
- Porter, R.; Collins, L.; Powell, J.; Rivenburgh, R. Information Space Models for Data Integration, and Entity Resolution. *Proc. SPIE* **2013**, *8396*, 92–103.
- Peter, C. *Data Matching—Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*; Springer: Berlin/Heidelberg, Germany, 2012.
- 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](#)]
- Yang, B.; Yunfei, Z.; Lu, F. Geometric-Based Approach for Integrating VGI POIs and Road Networks. *Int. J. Geogr. Inf. Sci.* **2013**, *28*, 126–147. [[CrossRef](#)]
- Zhang, W.; Gao, X.Y.; Li, R.S. Multi source POI data fusion of spatial location information. *J. Ocean Univ. China* **2014**, *44*, 111–116.
- Yang, B.; Zhang, Y. Pattern-Mining Approach for Conflating Crowdsourcing Road Networks with POIs. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 786–805. [[CrossRef](#)]
- Huang, H.; Yao, X.A.; Krisp, J.M.; Jiang, B. Analytics of Location-Based Big Data for Smart Cities: Opportunities, Challenges, and Future Directions. *Comput. Environ. Urban Syst.* **2021**, *90*, 101712. [[CrossRef](#)]
- Zhang, W.; Xia, L.F. Multi source heterogeneous POI fusion method and application. *Surv. Map. Rep.* **2018**, *3*, 143–146.
- Luo, G.W.; Ye, J.Y.; Wang, J.F. *A Multi-Source POI Matching Method Based on Multi Feature Similarity*; Bulletin of Surveying and Mapping: Shanghai, China, 2022; pp. 96–100.
- Xu, S.; Zhang, Q.; Li, D.; Liu, J.Y. A multi-source interest point fusion algorithm based on distance categories. *Comput. Appl.* **2018**, *38*, 1334–1338.
- Safra, E.; Kanza, Y.; Sagiv, Y.; Doytsher, Y. Integrating Data from Maps On The World-Wide Web. In *Web and Wireless Geographical Information Systems*; Springer: Berlin/Heidelberg, Germany, 2006.
- Beeri, C.; Doytsher, Y.; Kanza, Y.; Safra, E.; Sagiv, Y. Finding Corresponding Objects When Integrating Several Geo-Spatial Datasets. In Proceedings of the 2005 ACM International Workshop on Geographic Information Systems, Bremen, Germany, 4–5 November 2005.
- Wang, Y.; Liu, J.P.; Guo, Q.S.; Luo, A. A standardized processing method for network POI address information considering location relationships. *J. Surv. Mapp.* **2016**, *45*, 623–630.
- Li, X.; Morie, P.; Roth, D. Semantic Integration in Text: From Ambiguous Names to Identifiable Entities. *AI Mag.* **2005**, *26*, 45–58.
- Jun, Z.; Xin, H.; Yinghua, L.; Junhao, W.; Wei, Z. A Point-of-Interest Recommendation Method Using User Similarity. *Web Intell.* **2018**, *16*, 105–112.
- Sun, K.; Hu, Y.; Ma, Y.; Zhou, R.Z.; Zhu, Y. Conflating Point of Interest (POI) Data: A Systematic Review of Matching Methods. *Computers, Environment and Urban Systems*. *arXiv* **2023**, arXiv:2310.15320.
- Li, L.; Xing, X.; Xia, H.; Huang, X. Entropy-Weighted Instance Matching Between Different Sourcing Points of Interest. *Entropy* **2016**, *18*, 45. [[CrossRef](#)]

21. Zeng, W.; Fu, C.W.; Arisona, S.M.; Schubiger, S.; Burkhard, R.; Ma, K.L. Visualizing the Relationship Between Human Mobility and Points of Interest. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2271–2284. [[CrossRef](#)]
22. Xing, X.; Lin, H.; Zhao, F.; Qiang, S. Local POI Matching Based on KNN and LightGBM Method. In Proceedings of the 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 23–25 September 2022; pp. 455–458.
23. Cousseau, V.; Barbosa, L. Linking Place Records Using Multi-View Encoders. *Neural Comput. Applic.* **2021**, *33*, 12103–12119. [[CrossRef](#)]
24. Piech, M.; Smywinski-Pohl, A.; Marcjan, R.; Siwik, L. Towards Automatic Points of Interest Matching. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 291. [[CrossRef](#)]
25. Zhao, J.; Niu, X.; Cui, Y.; Zhao, Y.; Guo, M.; Zhang, R. POI Point Entity Matching And Fusion Based On Multi Similarity Calculation. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *10*, 87–92. [[CrossRef](#)]
26. Li, C.; Liu, L.; Dai, Z.; Liu, X. Different Sourcing Point of Interest Matching Method Considering Multiple Constraints. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 214. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.