entropy

MDPI

Article

Entropy-Weighted Instance Matching Between Different Sourcing Points of Interest

Lin Li [1,2], Xiaoyu Xing [1,*], Hui Xia [1] and Xiaoying Huang [1]

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; lilin@whu.edu.cn (L.L.); xiahui@whu.edu.cn (H.X.); xy.huang@whu.edu.cn (X.H.)

[2] Geo-Spatial Information Science Collaborative Innovation Center of Wuhan University, Luoyu Road 129, Wuhan 430079, China

* Correspondence: xiaoyu_xing@whu.edu.cn; Tel.: +86-27-6877-8879; Fax: +86-27-6877-8381

Academic Editors: Benjamin L. Ruddell and Kevin H. Knuth

Received: 13 September 2015; Accepted: 21 January 2016; Published: 28 January 2016

Abstract: The crucial problem for integrating geospatial data is finding the corresponding objects (the counterpart) from different sources. Most current studies focus on object matching with individual attributes such as spatial, name, or other attributes, which avoids the difficulty of integrating those attributes, but at the cost of an ineffective matching. In this study, we propose an approach for matching instances by integrating heterogeneous attributes with the allocation of suitable attribute weights via information entropy. First, a normalized similarity formula is developed, which can simplify the calculation of spatial attribute similarity. Second, sound-based and word segmentation-based methods are adopted to eliminate the semantic ambiguity when there is a lack of a normative coding standard in geospatial data to express the name attribute. Third, category mapping is established to address the heterogeneity among different classifications. Finally, to address the non-linear characteristic of attribute similarity, the weights of the attributes are calculated by the entropy of the attributes. Experiments demonstrate that the Entropy-Weighted Approach (EWA) has good performance both in terms of precision and recall for instance matching from different data sets.

Keywords: geospatial data; instance matching (POI matching); entropy; word segmentation; category mapping

1. Introduction

The traditional type of geospatial data, Points of Interest (POI, an important instance to convey geographical entity and location information), is attracting increasing attention in the geographical information science (GIS) domain and Location-Based Services (LBS) applications. By conflating the POI from distinct sources, we can exploit complementary attributes to improve data quality, overcome inconsistency, and promote data update [1].

Generally, data conflation (equally with fusion or integration) in GIS should be divided into three phases [2,3]: matching, integrating, and evaluation. In the matching stage, the identification of possible correspondences between multiple spatial data sources should be performed, namely to determine whether two objects from different datasets represent the same place in the physical world. This step is the foundation of data conflation, which receives the most attention in research papers. Integrating is the process of merging attribute values or setting the matched objects as anchors to perform the integration of images [3]. Finally, the evaluation refers to the process of checking the result on the basis of some given parameters, to validate the correctness of the integrated data.

However, POIs from different sources (without a global identifier) [4] with diversity not only in data structures, content, emphases and coverage area but also in the properties (e.g., category, address)

that contain semantic features exhibit discrepancy to some degree. In addition to the problem of inconsistencies caused by the shortage of global identifiers, even the matched attributes have significant discrepancies. Figure 1 shows a pair of corresponding POIs' differences in the spatial (in general, thus refers to longitude and latitude) and name attributes, which were obtained from Google map and Baidu map (both are well-known LBS applications). In Section 4.2, we will discuss this interesting phenomenon in detail (the topological space deviation is non-linear and not a fixed value). In addition, place name is a spontaneous cognitive and socially situated linguistic concept [5], which inevitably results in word or semantic ambiguity. For example, both "ka fei ting" and "ka fei dian" (in the Chinese pronunciation, a pair of synonyms) refers to a place to drink coffee. The heterogeneity occurs in the category attribute as well [6]. For example, a geographical entity in the real-world is categorized by the word "entertainment" in one dataset, whereas "leisure place" is probably used to identify the category value in another. As a consequence, although many approaches have been proposed to matching discrete sourcing geospatial data, the problem of instance matching between different sources in the geography discipline remains a challenge.

Figure 1. A pair of corresponding POIs from different sources (Google and Baidu).

In this paper, we mainly centered on the first step of geospatial data conflation-matching. The remainder of this paper is organized as follows: Section 2 presents related work regarding geospatial data integration and the corresponding objects matching in GIS; Section 3 presents the character of each attribute separately, describes the methodology to convey the similarity metric, and then describes the Entropy-Weighted Approach (EWA) we used to calculate the weights by the entropy of attributes; Section 4 presents experimental data, different models proposed based on EWA and an evaluation of our work; and Section 5 presents the conclusions of the study and discusses further work.

2. Related Work

Integration of geospatial data has gained considerable attention from researchers involved in GIS. Early in 2000, some scholars [7] proposed an approach by applying wrappers to extract geographic data from different heterogeneous sources and convert them into a unified format, and then integrated the formative data through mediators. By using a semantic value and disregarding its representation,

Fonseca [8] constructed geographic ontology for data integration. Du *et al.* [9] converted geo-data sets to ontologies and merged these ontologies into a coherent ontology to integrate disparate geospatial road vector data. Zhu, *et al.* [10] indicated that merging multi-source ontologies based on a concept lattice could reduce the redundancy among different concepts. To simplify the concept lattice, Li, *et al.* [11] described an entropy-based weighted approach to build a concept lattice by using information entropy to merge geo-ontologies.

In addition to the ontology-based method, match instances through finding corresponding objects could be another strategy to achieve geographic information integration [12], which involves identification of matched objects from different datasets that represent the same geographic entity. In the comparison of the ontology approach, it does not need to construct a complicated domain ontology library or repository of knowledge, which appears to be more practical for commercial POI providers. Beeri [4] and Safra, *et al.* [7] assumed that associated objects are closer to each other in the spatial attribute. Based on this assumption, they argued for an algorithm that has higher accuracy than the unilateral nearest neighbor algorithm and described the algorithm in parallel and/or series forms.

Associated object matching through non-spatial attributes (e.g., name, category, description) with semantic features is widely applied in various fields [6,13]. For example, Li, *et al.* [14] extracted the corresponding objects from fuzzy names by using a global clustering algorithm and a global generative model. Their research showed that a non-spatial attribute could also be adopted to match the associated entities. Therefore, studies [15,16] in GIS attempted to match associated geospatial objects by single name (or location description) attribute alone.

The method based on a single attribute (either a spatial or non-spatial attribute) is a relatively simple task; however, considering both the imprecision of the Volunteered Geographic Information (VGI) data attribute value [17,18] and the irregularities in the coding format resulting from linguistic ambiguity is more reasonable. Safra, *et al.* [19] combined the spatial and non-spatial attributes of geospatial data and improved the existing location-based matching algorithms by using Pre-D, Post-R and Pre-F technologies. Scheffler, *et al.* [20] used the spatial property as a fundamental filter and then combined the name metrics to match POIs from different social networking sites. To reflect the importance of property and set threshold flexibility, McKenzie proposed another heuristic approach that applies binomial logic regression [21] to assign weights and used the weighted multi-attributes model to find the corresponding objects.

However, the previous works mainly have three drawbacks. (1) The names of POIs in VGI have no canonical and authoritative coding standard due to conceptual and semantic ambiguity; traditional text similarity assessment (e.g., *Edit Distance*) would hamper the quantification of the semantic similarity between names because semantic disambiguation is ignored naturally; (2) Because POIs from multi-sources have diverse taxonomies, current research studies in geospatial data match or fusion rarely resolve the heterogeneity between different classifications due to the lack of an ontology library and domain experts; (3) Because the feature of the property similarity metric always exhibits a non-linear distribution, a new technique to confirm that the appropriate weights must be used to exploit the distribution and reflect the deviations of the attributes simultaneously [22,23].

Following this premise, our work is motivated by the need to address these problems mentioned above. The main contributions of this manuscript can be summarized as follows: First, word segmentation-based (along with sound-based) methods are adopted to eliminate the semantic ambiguity to express the name attribute in VGI data. Besides, for category attribute, mapping was established between the two taxonomies to address the heterogeneity and semantic relatedness. During the construction process of conception vectors, new method with the consideration of node depth and descendent node density was adopted to deal with the uneven phenomenon that caused by different classifications. Third, the information entropy technique was utilized to confirm the appropriate weights for integrating these attributes, which could deal with the non-linear characteristic of attribute similarity.

3. The Entropy-Weighted Approach for Finding Matched POIs

3.1. The Strategy of Attribute Selection

Let $P = \{p_1, p_2, ..., p_{n1}\}$ and $Q = \{q_1, q_2, ..., q_{n2}\}$ be two sets of attribute from independent data sources. To select the property used in the weighted multi-attributes model, we define the criteria of attribute selection as follows:

(1) For an attribute category m. If ($m \in P$ and $m \notin Q$) or ($m \notin P$ and $m \in Q$), then define the similarity of this attribute $s_m = 0$, and exclude this property in the weighted multi-attributes model.

(2) If ($m \in (P \cap Q)$), then confirm the calculation of according to the feature of attribute value and include this property in the weighted multi-attributes model.

The rules of attribute selection, spatial, name and category attribute categories are selected from the experimental dataset and are used in the weighted multi-attributes model. For each of the above-described properties, we propose the calculation of property similarity, and then present a series of Entropy-Weighted multi-attributes models to combine these properties together.

3.2. Spatial Similarity

The latitude and longitude of location are always essential components for spatial objects in many GIS applications. Similarly, for the purpose of further development based on their application programming interface (API), most of the commercial application operators offer the spatial attribute of POI. To measure the POI's similarity through the spatial attribute referring to the spatial length on the basis of the coordinate, the most elementary method is computation of the Euclidean metric [24] between two locations. Let $o_i = (x_i, y_i)$ show an object o_i with the projected latitude x_i and the projected longitude y_i. Coordinate similarity $s_{o_i o_j}$ is defined as the inverse of the Euclidean distance denoted as:

$$s_{o_i o_j} = \frac{1}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}} \tag{1}$$

Obviously, $s_{o_i o_j}$ is from 0 to $+\infty$. To eliminate anomalies between different measure scales and obtain reasonable weights through the attribute's information entropy, we should normalize the parameter [7]. We revised the exponential function slightly to ensure the dependent variable is restricted to the range from 0 (no match) to 1 (perfect match) as the independent variable is varied. The similarity displays an inverse correlation with distance $s_{o_i o_j}$, as given below:

$$s_{spatial} = e^{-s_{o_i o_j}/cons} \tag{2}$$

where $s_{spatial}$ is denoted as spatial similarity and *cons* is a constant. Figure 2 shows the similarity performance under a series of values (100, 200, 300, 400, 500 and 600). In fact, *cons* is a statistical characteristic constant depending on the training dataset [12]. Through the analysis of the experimental data, we find that 65% of the matching POIs have Euclidean metric drifts within 200 meters, and with increasing distance drift, the count of matching POIs decreases sharply (*i.e.*, requires a higher slope); as a result, we define *cons* = 500.

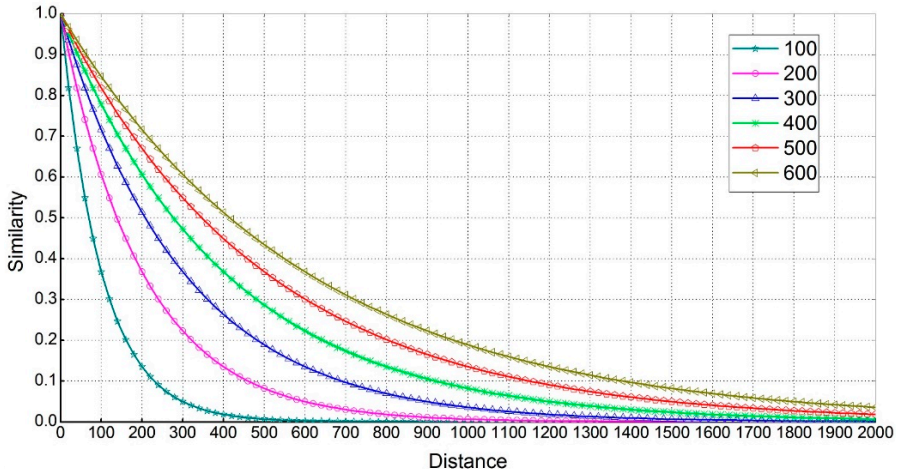

Figure 2. The performance of similarity under different values of the parameter *cons*.

3.3. Name Similarity

The name attribute is generally considered as a distinctive feature to distinguish POI intuitionally. Currently, a classical method for measuring the similarity of the name is the *Edit Distance* [25]. The distance is defined as the minimum number of edit operations (e.g., addition, deletion and change of character) required to convert one string (in text form) to another. Regardless of the location of the character that comprised the name string, one operation is assigned equal weight. *Edit Distance* was considered as an ideal method to measure the similarity metric under the circumstances of the string formally coded. We refer to the similarity under this calculation as s_{Text},

$$s_{Text} = \frac{Edit \text{ Distance } (N_i, N_j)}{Max \{\text{the length of } N_i, \text{ the length of } N_j\}}$$

where N_i and N_j are names of two POIs.

However, the entropy of s_{Text} shows that name property has the maximum entropy (compared with the spatial and category attributes). This result indicates that the name attribute has the weakest ability to distinguish POI, which is non-intuitive. We checked the experimental dataset and identified two linguistic and cognitive reasons that may result in the semantic ambiguity: (1) a Chinese character may have many sounds (polyphony) and may be presented in different forms (simplified Chinese character form and traditional Chinese character form) [26]. The reason for this discrepancy is that the POI producer (often via mobile device "on-the-go") may be dispelled: the producer is primarily focused on how the Chinese character sounds [21] and may neglect the character's form; (2) The name in one set uses a fixed phrase for short, whereas another may use the full name (e.g., Yellowstone National Park and Yellowstone Park both refers to a same entity); additional text (landmark or street name) may be attached ahead or behind the name to identify chain stores. For example, KFC is a famous fast food chain, and one may use "KFC-Luoyu Road" as a POI name, whereas "Luoyu Road KFC" may be used in another dataset.

To address the voice and form problem caused by language expression, we use another method to quantify the name similarity, which focuses on the pronunciation of Chinese characters. First, we conduct phonetic transcription via Microsoft voice packet. Second, the similarity of phonetic $s_{phonetic}$ is calculated based on the *Edit Distance* algorithm. In fact, calculation of the similarity via the pronunciation of the character is widely used in various fields [27], including search engines (e.g., Safari, Baidu), artificial intelligence, and so on.

The entropy of $s_{phonetic}$ decreases slightly compared with s_{Text}, which shows that the ability of name to identify a POI increases. To disambiguate the second semantic problem, we proposed the word

segmentation technique [28]. We divide the name on the basis of a small dictionary into sequences of words to enable each separate Chinese character to become meaningful words. For example, three single Chinese characters "Luo"," Yu" and" Road" are always put together to convey the road "Luoyu Road", but in the circumstance of a sparse training dataset, the word segmentation tool performs inefficient word sense discrimination training for undefined word detection. Hence, we construct a small dictionary for a word segmentation tool to extract the road name or landmark efficiently. Next, we establish the word's vector and compute cosine similarity [29] as the name similarity metric $s_{wordseg}$. Through this approach, the semantic feature in the name property could have a better and more actual expression.

3.4. Category Similarity

For the experimental data in Section 4.1, both Baidu map and Sina provide a three-tier hierarchical set of category tags, from which users can be selected to label a POI. Baidu map [30] has 22 broad categories, 218 intermediate categories and 340 minor categories; however Sina [31] offers 20 broad categories, 193 intermediate categories and 500 minor categories. The data producer or volunteer can select an intermediate and/or minor class to label the POI. The type of POI can also aid us in finding matches to some degree; at least, a POI tagged with different types may have more possibilities to represent different entities [32]. Strictly speaking, the category of geospatial data is an ontology question, and using ontologies is a privileged method to achieve interoperability among heterogeneous multi-sources system that with semantic feature [33]. As mentioned above, because of the shortage of ontology library and domain experts, the ontology alignment method was inappropriate for calculating category similarity for a commercial LBS provider.

To address the heterogeneity and semantic relatedness between different classifications in a practical way, this paper divided the process of category similarity calculation into two steps. First, mapping was established between the two taxonomies [32]. For a minor category, we trace up the hierarchy tree, assigning its parent category. The reason for this step is to increase the probability of matching two objects based on category [21], so we only need to build the connection between intermediate categories, as shown in Figure 3.

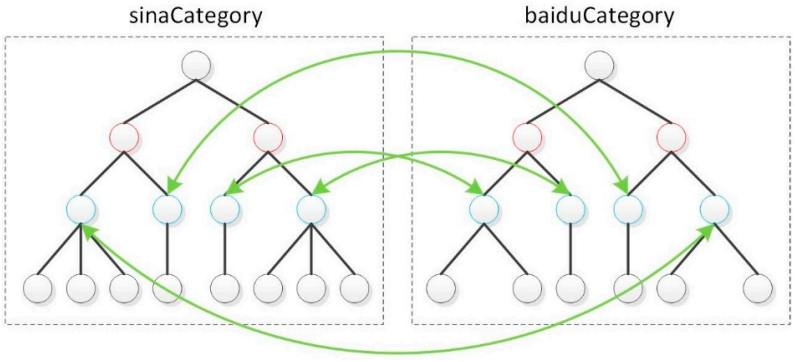

Figure 3. The map of middle categories for different taxonomies.

In the second step, the conception vectors of associated nodes in hierarchical concept graph are constructed, and then the similarity $s_{category}$ is obtained by computing the cosine similarity (via Equation (3)) of the conception vectors. Given the uneven situation caused by different classification systems, concept nodes may not have equivalent descendent concept nodes, so the components of the vectors we defined have different weights (traditionally, the components of a vector are all equals to 1) and the weights depend on the node's position and the descendent nodes' density

in the hierarchical graph. Refer to [34] for a more detailed introduction regarding construction of conception vectors with the consideration of the node depth and descendent node density:

$$s_{category} = \frac{v_p \times v_{p'}}{||v_p|| \, ||v_{p'}||} = \frac{\sum_{i=1}^{n} v_{p_i} \times v_{p_i'}}{\sqrt{\sum_{i=1}^{n} (v_{p_i})^2} \times \sqrt{\sum_{i=1}^{n} (v_{p_i'})^2}} \tag{3}$$

where $s_{category}$ is denoted as category similarity, v_p and $v_{p'}$ are conception vectors, and n is the length of the vectors. For example, we consider two situations as follows: (i) the red circle (in Figure 4a) is a minor category from sinaCategory and the green circle (in Figure 4a) is a minor category from baiduCategory, the two category have same/mapped intermediate categories and two conception vectors could constructed as v_p = [1, 0, 2, 0, 2, 0, 0, 0, 0, 5] and $v_{p'}$ = [1, 0, 2, 0, 2, 0, 0, 0, 5, 0]; (ii) the red circle (in Figure 4b) is a minor category from sinaCategory and the green circle (in Figure 4b) is an intermediate category from baiduCategory, the two conception vectors could constructed as v_p = [1, 0, 2, 2, 0] and $v_{p'}$ = [1, 0, 2, 0, 2, 0, 0, 0, 0, 5], to calculate dot product of the two vectors, we should complete v_p as [1, 0, 2, 2, 0, 0, 0, 0, 0, 0].

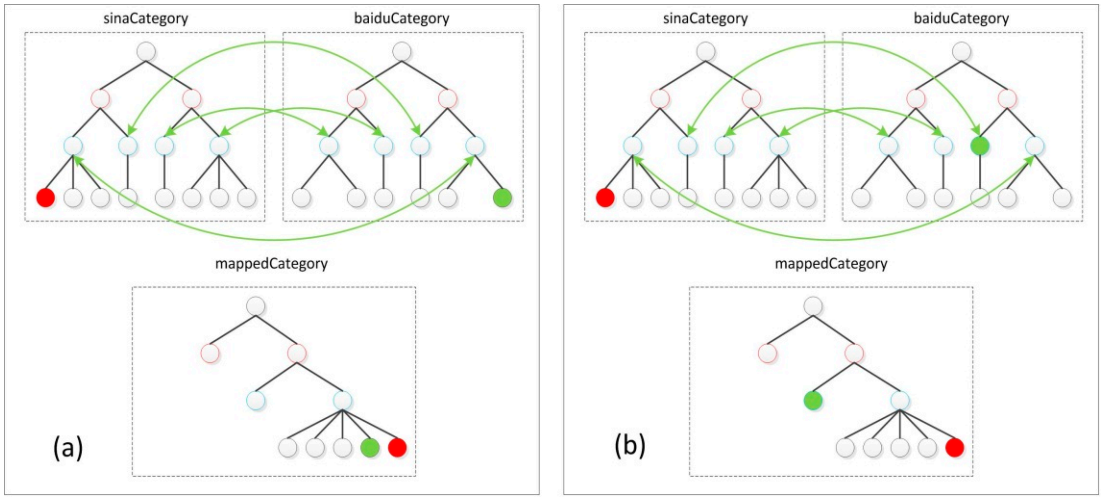

Figure 4. (**a**) Conception vectors constructed from two minor categories; (**b**) Conception vectors constructed from a minor and an intermediate category.

3.5. The Entropy-Weighted Multi-Attributes Method

The work of McKenzie [21] shows that an approach that combines multiple attributes is more appropriate for imprecise attribute values in comparison to the use of a single attribute. We denote the two independent data sets as $D1$ and $D2$, which are obtained from different sources. A pair of POI $\langle o, o' \rangle$ are called corresponding objects if they all represent the same entity, for which $o \in D1$ and $o' \in D2$. The set of attribute items is expressed by $M = \{m_1, m_2, ..., m_n\}$. We demonstrate the weights of the attributes $W = \{w_1, w_2, ..., w_n\}$, where $w_j \in W$ $(0 \leqslant w_j \leqslant 1)$ depict the importance of the attribute m_j. For a unique attribute, there may be several calculation methods. For example, we proposed sound-based and word segmentation-based methods to disambiguate the name of POI, and so the name attribute had three computation ways to express similarity feature. Our intention was not only to allocate suitable attribute weights by entropy for each attribute, but also aims to obtain a better combination of computation ways. So, we refer to w_j as the weight of a calculation method rather than an attribute category. Finally, the similarity of $\langle o, o' \rangle$ is defined as:

$$S(o, o') = \sum_{w_j \in W} (w_j \times s^j_{m_i}(o, o')) \tag{4}$$

$$\sum_{w_j \in W} w_j = 1$$

where w_j is the weight of a calculation method, m_i refers to an attribute belongs to attribute set M, $s^j_{m_i}(o, o')$ is the similarity of attribute m_i with the weight of w_j and $S(o, o')$ indicates a pair of POI's integral similarity.

Information entropy is a method to measure the uncertainty of information that can be used to evaluate its influencing factors and is widely used in various fields (e.g., word alignment, data mining, information theory) related to computer science [35]. Entropy is a well-known method for obtaining the objective weights for a multiple attribute decision making (MADM) problem [23,36], which refers to making preference decisions over the available alternatives that are characterized by multiple attributes. The process of obtaining weights w_j as following steps [23]:

(1) Set the probability distribution of each calculation p_{ij}, $i = 1, 2 \ldots n$, $j = 1, 2 \ldots m$, where n refers to the count of discrete similarity that divided with unique interval, and m equal the amount of w_j (for example, $m = 5$ in the *Five-Methods Model*).

(2) Compute the normalized information entropy E_j, the formula is given as follows [37]:

$$E_j = -h_0 \sum_{i=1}^{n} p_{ij} \cdot \log_2(p_{ij}) \tag{5}$$

where $h_0 = (log_2 n)^{-1}$ is the entropy constant and E_j is the information entropy. In relation to $p_{ij} = 0$, the value of $0 \cdot \log_2(0)$ is defined as 0.

(3) The weights are calculated as follows:

$$w_j = \frac{1 - E_j}{\sum_{k=1}^{m} (1 - E_k)} \tag{6}$$

The above w_j is naturally a normalized form, which indicates the importance of a calculation.

4. Case Study and Discussion

4.1. Experimental Dataset

To test the validity of the method we mentioned above, we collected the POI (Wuchang District, Wuhan City, Hubei Province, China) from Baidu [30] and Sina [31], which are labeled as *Db* (from Baidu) and *Ds* (from Sina), respectively. Baidu map is a well-known LBS application owned by Baidu. Using Baidu map, the developer can apply for a key and use their API (e.g., map service, POI search) conveniently. However, Sina is mainly a social networking site. Users can publish their reviews, check-in data and report a feeling about the place using the application Sina offers. During the process of data upload, with the permission of user, the application obtains the device position of the user through GPS in the mobile device.

Because every record in *Db* and *Ds* represents a unique geographic entity, we preprocess the data to ensure an entity only has one record in the dataset [7]. In the GIS field, the selection of the training dataset is always performed manually [1,4,12,21,32]. Similarly, we select 300 POIs in *Ds* randomly and then find 253 matching POIs in *Db* manually; the pairs of POIs are called dataset *Dm*.

4.2. The Spatial Attribute

For the spatial attribute, we obtained the distance between two corresponding objects using Equation (1). The distance of 80% of the objects achieved a 1000 m shift (as shown by the green bar in Figure 5a), with a few even drifting up to 1700 m. The similarity of the spatial property is obtained using Equation (2), as shown by the green line in Figure 5b. The result far exceeds our imagination of the margin of error for the VGI data. Why did this condition occur? Is there an issue with our method? With these questions, we contacted the two POI providers via e-mail and were told that they both encrypt the longitude and latitude via their own specific algorithm (*i.e.*, they use anonymous coordinate systems, whereas in international norms, the WGS84 coordinate system is used). The encryption algorithm will result in a shift that is not fixed and non-linear; to improve the accuracy matching in the map, one is forced to adopt the map service of the provider (*i.e.*, use the identical encryption algorithm). We used the transfer algorithm offered by the POI provider [30,31] to unite the different coordinate systems and recalculate distances (as the red bar shown in Figure 5a), and the similarity (as the red line shown in Figure 5b) in *Dm* is given by Equations (1) and (2) separately. Clearly, the similarity of most of matched POIs is significantly enhanced, and more factual features in the spatial attribute are presented.

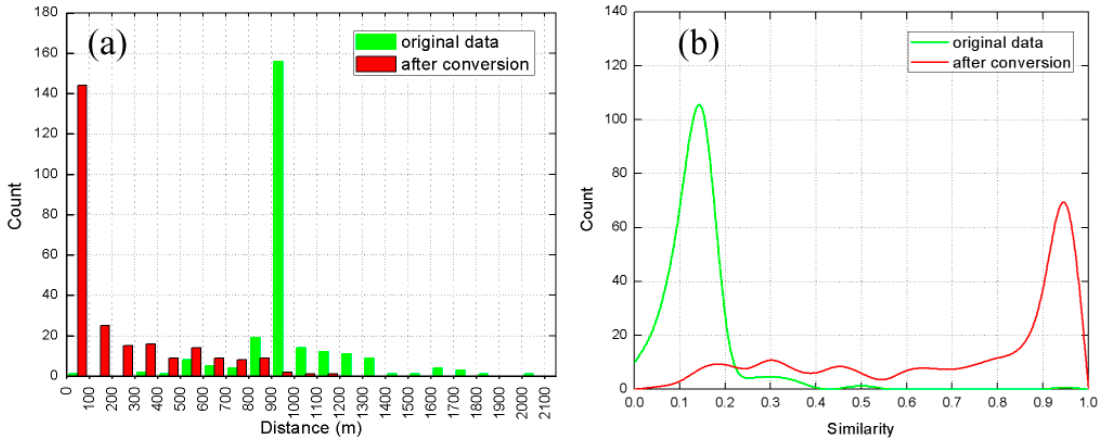

Figure 5. (**a**) The histogram of distance; (**b**) The performance of spatial similarity.

4.3. The Name Attribute

Initially, we calculated the POI name with *Edit Distance* and found that the similarity of the POIs is primarily gathered from 0.35 to 0.85 (as the green line shown in Figure 6 has the maximum entropy). This result illustrates the ambiguity of semantic expression: even the same entity may have a significant discrepancy in name. Using sound-based technology, we found the similarity line to be shifted right overall (as indicated by the blue line in Figure 6), and the entropy decreases slightly. However, by using the word segmentation method, the trend line has a more centralized form (as exhibited by the red line in Figure 6). We can draw two conclusions. First, the word segmentation technique is more appropriate than the sound-based method (at least the strategy we used, which has almost the same morphology with text-based method) for VGI POI to measure the name characteristic. Second, we eliminate the error in name largely by word segmentation and obtain the minimum entropy (compared with the text-based method and the sound-based method), showing that name could distinguish POI largely if an appropriate method is adopted.

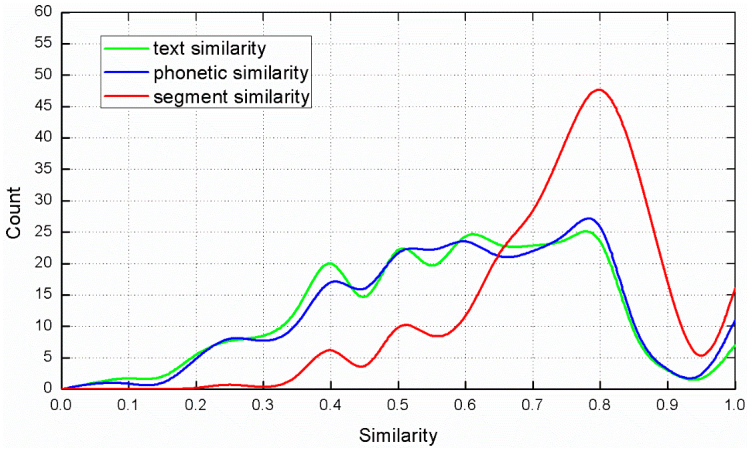

Figure 6. The similarity performance of the name attribute under different calculations.

4.4. The Category Attribute

The category attribute was considered as the weakest feature to identify POI, which was confirmed by our experiment. We map classifications in *Ds* to *Db* and then construct conception vectors with the consideration of the node depth and descendent node density. Finally, we calculate the category attribute similarity feature using Equation (3).

Figure 7 shows the performance of the count with varying values of similarity. We can make the following observations from the figure. First, a fraction of data in *Dm* have a low similarity (< 0.1), because some POI in one dataset are labeled with a correct minor category but may be tagged with the word "else" in another. This situation occurred in this work. Second, the line shows a more dispersive distribution, which is probably related with the informal habit to label a POI. For example, a leisure place labeled with "entertainment" may have a higher similarity metric than a place marked with "catering" or another category, for which the data producer tagged incorrectly and informally.

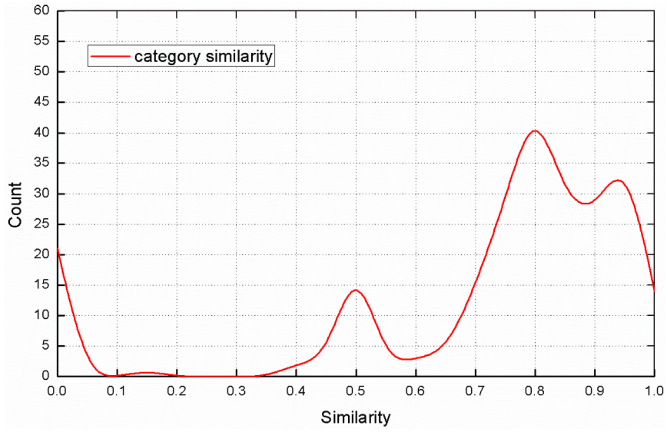

Figure 7. The similarity performance of the category attribute.

4.5. The Entropy-Weighted Multi-Attributes Model Analysis

The pivotal problem in the weighted multi-attributes model is the determination of the optimal weights. To obtain a discrete probability distribution of each calculation method, we calculated the count of different similarity range according to a 0.05 interval, and then divided the count by all pairs of POIs. P(1), P(2), P(3), P(4) and P(5) represent $s_{spatial}$, s_{Text}, $s_{phonetic}$, $s_{wordseg}$ and $s_{category}$ distributions of probability p_{ij}, respectively. Thus, the calculated entropies E_j using Equation (5) are presented in Table 1.

Table 1. The probability of different calculation methodologies and the entropy.

Similarity	P(1)	P(2)	P(3)	P(4)	P(5)
$0 \leqslant s < 0.05$	0	0	0	0	0.0830
$0.05 \leqslant s < 0.1$	0.0040	0.0040	0.0040	0	0
$0.1 \leqslant s < 0.15$	0.0079	0.0079	0.0040	0	0
$0.15 \leqslant s < 0.2$	0.0356	0.0040	0	0	0.0040
$0.2 \leqslant s < 0.25$	0.0395	0.0237	0.0198	0	0
$0.25 \leqslant s < 0.3$	0.0237	0.0316	0.0356	0.004	0
$0.3 \leqslant s < 0.35$	0.0514	0.0316	0.0277	0	0
$0.35 \leqslant s < 0.4$	0.0277	0.0435	0.0356	0.004	0
$0.4 \leqslant s < 0.45$	0.0198	0.0988	0.0791	0.0356	0.0079
$0.45 \leqslant s < 0.5$	0.0395	0.0356	0.0514	0	0.0119
$0.5 \leqslant s < 0.55$	0.0237	0.1067	0.0949	0.0514	0.0791
$0.55 \leqslant s < 0.6$	0.0079	0.0632	0.0830	0.0277	0.0079
$0.6 \leqslant s < 0.65$	0.0316	0.1067	0.0988	0.0395	0.0119
$0.65 \leqslant s < 0.7$	0.0316	0.0870	0.0791	0.0909	0.0158
$0.7 \leqslant s < 0.75$	0.0277	0.0909	0.0870	0.1067	0.0593
$0.75 \leqslant s < 0.8$	0.0356	0.0909	0.0949	0.1581	0.1146
$0.8 \leqslant s < 0.85$	0.0474	0.1107	0.1225	0.2055	0.1818
$0.85 \leqslant s < 0.9$	0.0514	0.0237	0.0277	0.1502	0.1146
$0.9 \leqslant s < 0.95$	0.1146	0.0119	0.0119	0.0632	0.1067
$0.95 \leqslant s < 1$	0.3794	0	0	0	0.1462
$s = 1$	0	0.0277	0.0435	0.0632	0.0553
E_j	0.765	0.873	0.872	0.739	0.766

Table 1 indicates the following. First, the entropy values of s_{Text}, $s_{phonetic}$ and $s_{wordseg}$ decrease gradually, which indicates that sound-based and word segmentation-based methods could handle the semantic discrepancy problem that results from semantic expression and show a more realistic characteristic of the name attribute. Moreover, s_{Text} gets the maximum entropy compared with $s_{spatial}$, and $s_{category}$, which indicates that the name attribute has the weakest (compared with the spatial and category attributes) ability to distinguish POI. However, the entropy values of $s_{wordseg}$, $s_{spatial}$ and $s_{category}$ increase by various degrees, which shows that name is a strongest distinctive feature to distinguish POI if the proper measurement is chosen.

To obtain a better combination of different calculation methods and the optimal weights, we build the *Five-Methods Model*, the *Four-Methods Model* and the *Three-Methods Model* separately. The *Five-Methods Model* contains $s_{spatial}$, s_{Text}, $s_{phonetic}$, $s_{wordseg}$ and $s_{category}$, which we refer to by the abbreviation *STPWC*. The *Four-Methods Model* includes: $s_{spatial}$, s_{Text}, $s_{phonetic}$ and $s_{category}$ (*STPC* for short); $s_{spatial}$, s_{Text}, $s_{wordseg}$ and $s_{category}$ (*STWC* for short); and $s_{spatial}$, $s_{phonetic}$, $s_{wordseg}$ and $s_{category}$ (*SPWC* for short). The *Three-Methods Model* is comprised of three calculations: $s_{spatial}$, s_{Text} and $s_{category}$ (*STC* for short); $s_{spatial}$, $s_{phonetic}$ and $s_{category}$ (*SPC* for short); and $s_{spatial}$, $s_{wordseg}$ and $s_{category}$ (*SWC* for short). The weights calculated by Equation (6) are provided in Table 2.

Table 2. Different methods involving combinations of models and weights.

	Abbreviation	Spatial	Text	Phonetic	WordSeg	Category
Five-Methods Model	*STPWC*	0.2386	0.1289	0.1299	0.2650	0.2376
Four-Methods Model	*STPC*	0.3246	0.1754	0.1768	–	0.3232
	STWC	0.2742	0.1482	–	0.3046	0.2730
	SPWC	0.2739	–	0.1492	0.3042	0.2727
Three-Methods Model	*STC*	0.3943	0.2131	–	–	0.3926
	SPC	0.3936	–	0.2144	–	0.3920
	SWC	0.3219	–	–	0.3575	0.3205

After obtaining the w_j, we calculate the integral similarity using Equation (4) and then set a series of thresholds T (from 0.1 to 0.9) and compare them with the integral similarity $S(o, o')$ (if $S(o, o') \geqslant T$, then the two POIs are identically matched). As shown in Figure 8a), the precisions of these three methods in the *Four-Methods Model* all improve along with increasing T. The *STPC* model achieves the best performance in precision. However, comparing the three methods in the *Three-Methods Model*, *SPC* (almost equivalent with *STC*) has a higher precision (as depicted in Figure 8b). The *STPC* (the best in *Four-Methods Model*), the *SPC* (the best in *Three-Methods Model*) and the word segmentation-based (has the strongest ability to distinguish POI in a single attribute) are comparable with *STPWC* in precision. From Figure 8c), we can learn the following: (1) all of the weighted multi-attribute models have a better performance in precision than the word segmentation-based method (when $T > 0.45$, the calculation has the strongest ability to distinguish POI among a single attribute); (2) the *SPC* model has the highest precision, which is slightly better than *STPC* and apparently better than *STPWC*; (3) Compared with the Logic regression method [21], which assumed as an effective method for matching POIs, our method has almost equal performance in precision.

Conversely, Recall is a parameter defined as the proportion of positive matched POIs which are correctly identified *Recall = Positive Instances Predicted / Total Positive Instances*, and the parameter represents the ability of finding the corresponding POIs in the dataset. The higher of the Recall value, the stronger of the model's ability to find the corresponding POIs. Figure 8d) shows that the *STPC* model achieves the poorest performance in Recall (compared with *STWC* and *SPWC*), indicating that the word segmentation method, although decreasing the accuracy slightly, achieves a higher ability for matching more POIs. For the same feature shown in Figure 8e), the *SWC* achieves a higher Recall among these *Three-Methods Models*.

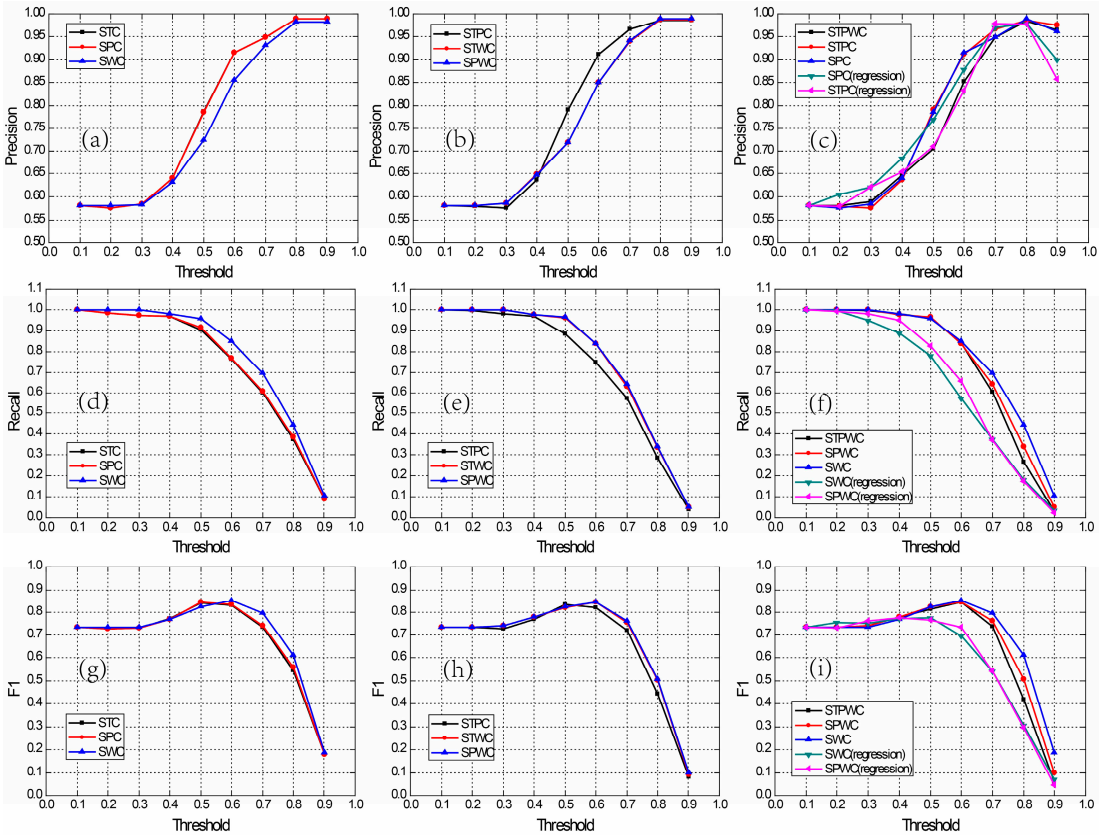

Figure 8. (**a**) *prec.* of *STC*, *SPC* and *SWC*; (**b**) *prec.* of *STPC*, *STWC* and *SPWC*; (**c**) *prec.* of *STPWC*, *STPC* and *SPC*; (**d**) Recall of *STC*, *SPC* and *SWC*; (**e**) Recall of *STPC*, *STWC* and *SPWC*; (**f**) Recall of *STPWC*, *SPWC* and *SWC*; (**g**) F1 of *STC*, *SPC* and *SWC*; (**h**) F1 of *STPC*, *STWC* and *SPWC*; (**i**) F1 of *STPWC*, *SPWC* and *SWC*.

Comparing *STPWC*, *SPWC*, *SWC*, *SWC (regression)* (the weights obtained by the Logic regression method) and *SPWC (regression)* (the weights obtained similarly by the Logic regression method) in Figure 8f, we can draw the conclusion that sound-based technology and word segmentation technology that eliminate the semantic ambiguity problem can substantially improve the Recall. In addition, when we set $T > 0.6$, *SWC* achieves the best performance of all of the weighted multi-attribute models. It also indicates that our method is much better than the Logic regression method in recall, which means it has stronger ability to identify corresponding POIs in different datasets.

We not only attempt to obtain a higher accuracy but also hope to achieve a higher recall in practice, to find as many corresponding instances as possible, so that the objects are truly matched. The parameter $F1$ value was introduced to measure the harmonic mean of precision and recall, defined as $F1 = (2 \times r \times p)/(r + p)$, where r refers to recall and p refers to precision. A higher $F1$ indicated that a model can identify more POIs as matched and the POIs are truly corresponding. Figure 8g) shows that *SPWC* (almost same with *STWC*) achieves the best performance among these *Four-Methods Models*. In addition, *SWC* achieves the best performance (shown as Figure 8h) among the *Three-Methods Models*. Figure 8i) shows that when $T \leqslant 0.6$, *STPWC*, *SPWC*, and *SWC* achieve similar performances, but decrease gradually when $T > 0.6$. Besides, our method apparently outperform the Logic regression method when $T > 0.6$, even though at the peak of the curve ($T = 0.6$), *STPWC*, *SPWC*, and *SWC* have a $0.1 \sim 0.15$ increase, which will has great significance for instance matching among millions of VGI POIs. Overall, the *SWC* model can achieve both a higher accuracy and a higher recall simultaneously, *i.e.*, the method can find more matched POIs and the POIs are truly corresponding, especially for a threshold greater than 0.6.

5. Conclusions and Future Work

On the one hand, with the popularity of LBS and mobile positioning devices, large numbers of VGI POIs are increasingly collected, which raises a crucial problem regarding their integration, especially in geographical data interoperability. On the other hand, for heterogeneous POIs from multiple sources, finding the corresponding objects is difficult because of the lack of global identifiers and imprecise attribute values.

This paper first proposed the rule of attribution selection, analyzed the conundrum in POI matching for spatial, name and category attribution, and determined a strategy for weighted multi-attributes for matching POIs. Second, the name attribute achieves the maximum entropy under *Edit Distance* similarity calculation, which naturally neglects the conceptual disambiguation and linguistic ambiguity. We added sound-based and word segmentation-based methods to express the similarity metric of the name property and found that the entropy reduced obviously; we also demonstrated that an appropriate method could eliminate the semantic ambiguity caused by cognitive and informal linguistic expression in name. Finally, for an imprecise and fuzzy POI property in VGI, we allocated suitable attribute weights by entropy to tackle the problem raised by the non-linear similarity feature of the nondeterministic attribute item. The experiments show that the Entropy-Weighted Approach can match more POIs (Recall) and the matched POIs are truly correct (Precision). The work provides strong evidence that information entropy theory can be used in the field of geospatial instance matching. In practice, one can flexibly set an appropriate threshold value to obtain the corresponding objects at various confidence levels.

In the future, other important aspects also should be taken into account. For example, we will pay more attention to additional attributes(review, phone number and so on) that homogeneously have the potential to identify a POI to improve the model, and we will focus on how to merge attribute values and evaluate the quality of the merged results, namely the second step and third step of geospatial data conflation.

Acknowledgments: This work was supported by the 863 High Technology of China (No. 2013AA12A202) and the Special Fund for Surveying, Mapping and Geographical Information Scientific Research in the Public Interest (No. 201412014). We gratefully acknowledge the editor and anonymous reviewers for their constructive suggestions.

Author Contributions: This article was mainly conducted by Lin Li and Xiaoyu Xing. Lin Li supervised this research and his comments were considered throughout the paper; Xiaoyu Xing and Hui Xia collected and analyzed the data, designed and performed the experiments; Xiaoyu Xing wrote the paper; Xiaoying Huang contributed to the correlation analysis. All authors have read and approved the final manuscript.

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

References

1. Hastings, J.T. Automated conflation of digital gazetteer data. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1109–1127. [CrossRef]
2. Porter, R.; Collins, L.; Powell, J.; Rivenburgh, R. Information space models for data integration, and entity resolution. *Proc. SPIE* **2012**, *8396*, 263–276.
3. Ruiz, J.J.; Ariza, F.J.; Urena, M.A.; Blazquez, 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]
4. Beeri, C.; Doytsher, Y.; Kanza, Y.; Safra, E.; Sagiv, Y. Finding Corresponding Objects when Integrating Several Geo-Spatial Datasets. In Proceedings of the 13th ACM International Workshop on Geographic Information Systems, Bremen, Germany, 4–5 November 2005; Association for Computing Machinery: New York, NY, USA, 2005; pp. 87–96.
5. Kitchin, R.M. Increasing the integrity of cognitive mapping research: Appraising conceptual schemata of environment behaviour interaction. *Prog. Hum. Geogr.* **1996**, *20*, 56–84. [CrossRef]
6. Michalowski, M.; Ambite, J.L.; Thakkar, S.; Tuchinda, R.; Knoblock, C.A.; Minton, S. Retrieving and semantically integrating heterogeneous data from the web. *IEEE Intell. Syst.* **2004**, *19*, 72–79. [CrossRef]
7. Safra, E.; Kanza, Y.; Sagiv, Y.; Beeri, C.; Doytsher, Y. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 69–106. [CrossRef]
8. Fonseca, F.T.; Egenhofer, M.J.; Agouris, P.; Câmara, G. Using ontologies for integrated geographic information systems. *Trans. GIS* **2002**, *6*, 231–257. [CrossRef]
9. Du, H.; Anand, S.; Alechina, N.; Morley, J.; Hart, G.; Leibovici, D.; Jackson, M.; Ware, M. Geospatial information integration for authoritative and crowd sourced road vector data. *Trans. GIS* **2012**, *16*, 455–476. [CrossRef]
10. Zhu, J.; Wang, J.; Li, B. A formal method for integrating distributed ontologies and reducing the redundant relations. *Kybernetes* **2009**, *38*, 1870–1879.
11. Li, J.; He, Z.; Zhu, Q. An entropy-based weighted concept lattice for merging multi-source geo-ontologies. *Entropy* **2013**, *15*, 2303–2318. [CrossRef]
12. Samal, A.; Seth, S.; Cueto, K. A feature-based approach to conflation of geospatial sources. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 459–489. [CrossRef]
13. Garla, V.N.; Brandt, C. Semantic similarity in the biomedical domain: An evaluation across knowledge sources. *BMC Bioinform.* **2012**, *13*. [CrossRef] [PubMed]
14. Li, X.; Morie, P.; Roth, D. Semantic integration in text: From ambiguous names to identifiable entities. *AI Mag.* **2005**, *26*, 45–58.
15. Vasardani, M.; Winter, S.; Richter, K.F. Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2509–2532. [CrossRef]
16. Wang, W.; Stewart, K. Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Comput. Environ. Urban Syst.* **2015**, *50*, 30–40. [CrossRef]
17. Mulliganni, C.; Janowicz, K.; Ye, M.; Lee, W.-C. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Spatial Information Theory*; Egenhofer, M., Giudice, N., Moratz, R., Worboys, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 350–370.
18. Yang, B.S.; Zhang, Y.F.; Lu, F. Geometric-based approach for integrating vgi pois and road networks. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 126–147. [CrossRef]

19. Safra, E.; Kanza, Y.; Sagiv, Y.; Doytsher, Y. Integrating Data from Maps on the World-Wide Web. In *Web and Wireless Geographical Information Systems*; Carswell, J.D., Tezuka, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 180–191.
20. Scheffler, T.; Schirru, R.; Lehmann, P. Matching Points of Interest from Different Social Networking Sites. In *KI 2012: Advances in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 245–248.
21. 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]
22. Jost, L. Entropy and diversity. *Oikos* **2006**, *113*, 363–375. [CrossRef]
23. Lotfi, F.H.; Fallahnejad, R. Imprecise shannon's entropy and multi attribute decision making. *Entropy* **2010**, *12*, 53–62. [CrossRef]
24. Arsigny, V.; Fillard, P.; Pennec, X.; Ayache, N. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **2006**, *56*, 411–421. [CrossRef] [PubMed]
25. Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **2001**, *33*, 31–88. [CrossRef]
26. Liu, W.; Cai, M.; Yuan, H.; Shi, X.; Zhang, W.; Liu, J. Phonotactic Language Recognition Based on Dnn-HMM Acoustic Model. In Proceedings of the 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, 12–14 September 2014; pp. 153–157.
27. Meltzoff, A.N.; Kuhl, P.K.; Movellan, J.; Sejnowski, T.J. Foundations for a new science of learning. *Science* **2009**, *325*, 284–288. [CrossRef] [PubMed]
28. Mattys, S.L.; Davis, M.H.; Bradlow, A.R.; Scott, S.K. Speech recognition in adverse conditions: A review. *Lang. Cognit. Process.* **2012**, *27*, 953–978. [CrossRef]
29. Nie, X.; Feng, W.; Wan, L.; Xie, L. Measuring Semantic Similarity by Contextual Word Connections in Chinese News Story Segmentation. In Proceddings of the 2013 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 8312–8316.
30. Baidu. Available online: http://developer.baidu.com/map/index.php (accessed on 20 June 2015).
31. Sina. Available online: http://open.weibo.com/ (accessed on 20 June 2015).
32. Sehgal, V.; Getoor, L.; Viechnicki, P.D. Entity Resolution in Geospatial Data Integration. In Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems, ACM-GIS'06, Arlington, VA, USA, 6–11 November 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 83–90.
33. Sanchez, D.; Batet, M. A semantic similarity method based on information content exploiting multiple ontologies. *Expert Syst. Appl.* **2013**, *40*, 1393–1399. [CrossRef]
34. Liu, H.Z.; Bao, H.; Xu, D. Concept vector for semantic similarity and relatedness based on wordnet structure. *J. Syst. Softw.* **2012**, *85*, 370–381. [CrossRef]
35. Dincer, I.; Cengel, Y. Energy, entropy and exergy concepts and their roles in thermal engineering. *Entropy* **2001**, *3*, 116–149. [CrossRef]
36. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
37. Machado, J. Fractional order generalized information. *Entropy* **2014**, *16*, 2350–2361. [CrossRef]