


Article

A Semantic Expansion Model for VGI Retrieval

Tao Sun ¹, Hui Xia ², Lin Li ^{1,3,4,*} , Hang Shen ¹ and Yu Liu ¹

¹ School of Resource and Environment Sciences, Wuhan University, Wuhan 430079, China; rbplan@whu.edu.cn (T.S.); shenhang@whu.edu.cn (H.S.); liuyu0201@whu.edu.cn (Y.L.)

² Changjiang Spatial Information Technology Engineering CO, LTD, Wuhan 430079, China; xiahui3@cjwsjy.com.cn

³ The Key Laboratory for Geographical Information Systems, Ministry of Education, Wuhan 430079, China

⁴ Geo Spatial Information Science Collaborative Innovation Center of Wuhan University, 129 Luoyu Rd., Wuhan 430079, China

* Correspondence: lilin@whu.edu.cn; Tel.: +86-138-7150-4963

Received: 19 October 2019; Accepted: 16 December 2019; Published: 17 December 2019



Abstract: OpenStreetMap (OSM) is a representative volunteered geographic information (VGI) project. However, there have been difficulties in retrieving spatial information from OSM. Ontology is an effective knowledge organization and representation method that is often used to enrich the search capabilities of search systems. This paper constructed an OSM ontology model with semantic property items. A query expansion method is also proposed based on the similarity of properties of the ontology model. Moreover, a relevant experiment is conducted using OSM data related to China. The experimental results demonstrate that the recall and precision of the proposed method reach 80% and 87% for geographic information retrieval, respectively. This study provides a method that can be used as a reference for subsequent research on spatial information retrieval.

Keywords: geographical information retrieval; OSM ontology; semantic similarity; query expansion

1. Introduction

The importance of spatial information retrieval is increasing, and web map services now attract a wide range of users. For example, the Baidu map service in 2015 had 302 million monthly active users in China, and queries using the Baidu map service to retrieve living services accounted for 40% of the total search volume [1]. Furthermore, the proportion of people engaging in spatial information retrieval activities who are untrained has increased substantially [2]. These nonprofessional users generally use simple keywords to search for information. However, the actual query results often differ significantly from the results expected by the user [3]. For example, using abbreviations or aliases as query words, such as using “wuda” to search for Wuhan University and its surrounding spatial entities, generates many results with names that contain these two words but are not relevant. Providing users with an intelligent, efficient way to allow them to complete their queries with a few relevant keywords is a key issue that researchers in the information retrieval field around the globe are highly focused on addressing [4].

An obvious approach to solve this problem is query expansion. Expanding a query using words with meanings similar to those in the query increases the chances of matching query targets. Previous research predominantly approached query expansion on the basis of local analysis. The local analysis approach expands a user query by selecting relevant terms from the top-ranked terms initially retrieved in response to a user-provided query. In general, this approach adjusts the weights of expansion terms based on similarity algorithms [5]. However, the result is very sensitive to the quality of the original query, and the effectiveness of the approach may be limited when the top-ranked terms span different topics [6]. Furthermore, this approach does not consider the semantic information of the original query.

Therefore, a query expansion that has no spatial relation with a query target may misguide the result in spatial information retrieval. In our method, many kinds of semantic properties, like numerical and descriptive properties, are considered, and different similarity calculation methods are used for the semantic properties and spatial information to measure similarity accurately.

Spatial information retrieval has prominent spatial features and is a special application in the field of information retrieval. The unique nature of spatial information retrieval is manifested by similarity retrieval, spatiotemporal association retrieval, and uncertainty of knowledge problems [7]. Conventional keyword string matching-based information retrieval techniques cannot meet the aforementioned needs; therefore, higher-level semantic-based information retrieval and matching methods are needed [8]. As an effective knowledge organization and representation method, the ontology technique used in the Semantic Web has shown the merits of structuring information and supporting logical reasoning. This technique has developed rapidly and has been extensively applied in the information retrieval field, particularly for knowledge-based semantic retrieval [9]. Many studies have applied the ontology technique for many topics, including spatial retrieval. Cardoso and Silva used ontology to match geographical features and spatial relationships and readjust the expansion strategy of geographical queries [10]. Derbal et al. proposed a querying approach for geographic information by using domain and user ontologies and developed a web GIS application prototype based on the approach [11]. Fu et al. reported on the use of ontologies developed in the EU Semantic Web project SPIRIT to support the retrieval of documents that are spatially relevant to users' queries [12]. However, ontology model building requires substantial work, and the effect of the information retrieval method is closely related to the quality of the model. Current models are limited to small data sets, and this limitation leads to poor scalability. Moreover, similarity algorithms play a key role in the semantic retrieval field. Most existing methods consider only spatial relations and pay little attention to other implicit relations such as materiality similarity. For example, water is the shared materiality for river and lake.

In recent years, with the emergence of the concept of geographic information crowdsourcing, there has been explosive growth in volunteered geographic information (VGI) [13]. The practice of voluntarily collecting and sharing geographic information has significantly reduced the acquisition cost of geographic information and shortened information update intervals. OpenStreetMap (OSM) is a representative VGI project whose data have been extensively applied in various fields [14]. In the OSM data specification, there is a simple, open semantic description structure called "tag", which provides a convenient means of expanding the semantics of OSM data. By expanding the semantic property items of these tags, we can build a more complete ontology model and measure semantic similarity more accurately with more information.

In this study, an ontology model for OSM tags is constructed to describe geographic properties of features. To ensure that the query results match the expected results and eliminate irrelevant results from a geographic view, a semantic-based query expansion method for geographic information is proposed based on the similarity of the ontology semantic properties and spatial information. In the query process, the user-input keywords are extended by using a series of corresponding similarity algorithms. In addition, the relevant calculation results of keywords are sorted based on their spatial semantic similarity. Experimental results show that the proposed method exhibits better query performance than a non-query expansion method and a conventional information entropy-based query expansion method.

The remainder of this paper is organized as follows. In Section 2, we introduce the data source and our pretreatment process. Section 3 briefly explains our proposed ontology model and expanded method based on semantic ontology similarity. Section 4 presents the experimental results of different query methods. Section 5 discusses the experimental results of those methods. Finally, Section 6 presents conclusions and future research directions.

2. Data Acquisition and Preprocessing

2.1. OSM Dataset

Because of the diversity of information provided by a vast number of volunteers, OSM data are more detailed than the data provided by map providers in several regions of the world [15]. In OSM data, geographic information is described using a data model containing nodes, ways and relations. As shown in Figure 1, a node is a simple object that has key-value tags and a couple of coordinates. A node can represent any point information. The way to define a line or area by containing a sorted sequence of nodes and a collection of key-value tags. The relations represent more complex objects, like polygons with holes. Any relationship among nodes, ways and relations can be modeled. Using the tags of OSM data, the property data corresponding to a map element in OSM are represented. Both the “key” and “value” property items carry data in freely formatted text-type values and are shown in the form of “key=value”. This form allows users to customize extensions, which render OSM data highly extensible and significantly enrich the content of the provided geographic information.

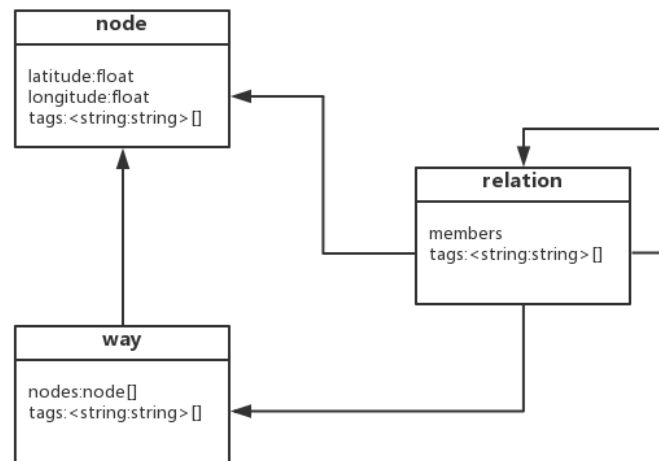


Figure 1. The data model of OSM.

2.2. Property Extraction

In OSM, only the nodes contain location information (i.e., longitude and latitude). The way data and relation data reflect the spatial location information by association with the identification of the nodes. To facilitate the subsequent retrieval of spatial data and the judgment of spatial relationships, we need to establish spatial location information and indices for the way and relation data. Therefore, we preprocess the dataset by extracting the values in the Key and Value fields. These values are used to build an index that will be very important for the subsequent query expansion of place names. The generation and maintenance of quadtrees is relatively simple, so we used a quadtree index. By constructing minimum bounding rectangles, we constructed a quadtree spatial index for each element to improve the subsequent search and retrieval efficiency. In this study, property data are extracted from the offline files of OSM data in XML format and stored in a local dataset.

3. Semantic Model for Information Retrieval

3.1. OSM Ontology Model

In recent years, research on ontologies has gradually matured. However, the definition of ontology and the use of relevant terms are not completely consistent in the literature [16–18]. In addition, OSM has yet to officially release an OSM ontology model with detailed ontology properties. The classification system provided by OSM officials focuses on improving the expression of symbols. In this study, spatial information retrieval pays more attention to spatial information, concept classification, and relations.

Therefore, the classification of OSM elements needs to be extended and modified so that the elements can be applied to the field of spatial information retrieval services. In this study, based on fundamental geographic information features and the hierarchical conceptual model included in OSM, an ontology model applicable to the query and retrieval of OSM data is proposed. The ontology model is built in Protégé and stored in OWL or RDF file format. This model contains several classes: A set of geographic information concepts, a spatial data model, and instance properties. Each class contains several subclasses and entries. Each concept or entity has semantic description properties. Taking the “canal” category in Chinese as an example, the semantic primitives of this category are extracted by applying NLP (natural language processing) tools to a set of phrases, including “aqueduct”, “manual”, “water transfer”, “shipping”, and “across river basin”. Then, by transforming these semantic primitives into the proposed formal ontology model, the semantics of the “canal” category can be represented as a set of several semantic statements as follows:

$$\begin{aligned}
 C = \{ & T_C = \text{“(Canal)”} \cap H_C = \text{“Hypernym : (Waterway, Aqueduct)”} \cap P_C = \text{“(Purpose} \\
 & : \text{(Water transfer, Diversion))} \cup \text{(Purpose : Shipping)} \cap P_C \\
 & = \text{“Nature : (Manual, Artificial)”} \cap R_S = \text{“Topology} \\
 & : \text{(Inter – basin, Across river basin)”}
 \end{aligned}$$

This domain ontology model can facilitate the formal semantic description of the current classification and hierarchy concepts in OSM and further strengthen and quantify the distinctions and associations between concepts. Figure 2 shows a portion of the OSM ontology model structure.

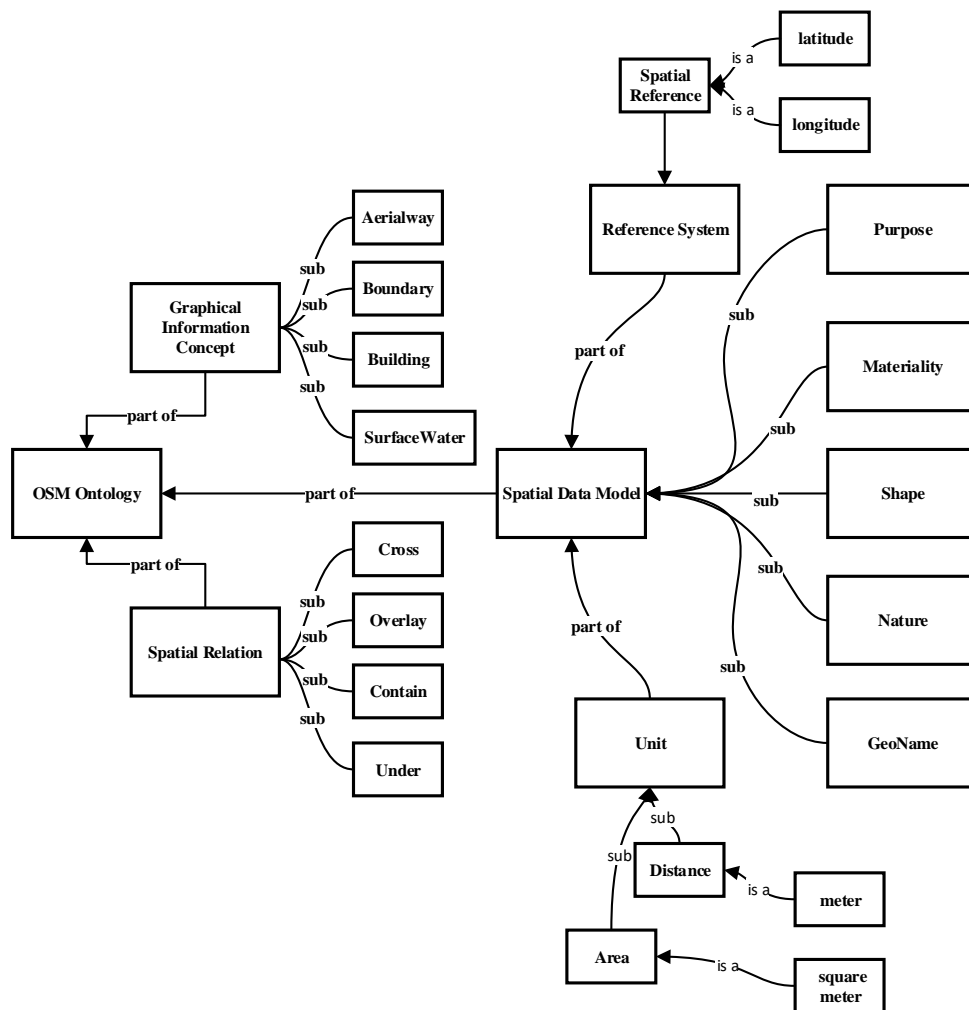


Figure 2. OSM ontology model (partial).

According to the “Specifications for feature classification and codes of fundamental geographic information” of China and OSM nomenclature, spatial information concepts in this study include eight categories: spatial reference, water system, geomorphology, transportation, habitation, pipeline, boundary and land use types. By further sorting and summarizing the OSM concepts, 15 main ontology properties are selected and extracted to comprehensively formalize the essential features of the spatial information concepts. For example, materiality can distinguish the composition and structure of a geographic information entity. The materiality of lakes and rivers is water in liquid form. Shape is another ontology property. The shape of a road is a line, and the shape of a street block is a polygon. Table 1 lists the 15 ontology properties.

Table 1. List of the semantic properties of the OSM ontology model concepts.

Category	Ontology Properties/Keys	Examples
GeoName	GeoName	Wuhan University
materiality	materiality	solid and sand
cause	cause	natural
spatial morphology	structure	multi-story
	shape	multi-line
	orientation	northern
spatial location	spatial relation	beside
	locality	Hubei province
temporality	position	longitude; latitude
	cyclicity	seasonal
function	lifecycle	scrapped
specificity	function	transportation
measurement	specificity	affiliated
hierarchy	measurement	height
	hierarchy	part-of; subclass; kind-of

3.2. Similarity Query Expansion

An effective information retrieval approach that is generally accepted in the information retrieval field is to first perform a query expansion based on simple or fuzzy user-input search keywords and then use richer search terms to obtain more comprehensive and accurate results [19]. Because geographic information has inherent spatial characteristics, conventional language or lexical meaning-based query expansion methods are not applicable in the geographic information field [20]. Thus, there is an urgent need for a search that uses a query expansion method for the geographic information field that considers the spatial characteristics of geographic information. Since ontology can only fundamentally solve the problem of cognition unity, semantic-based query extension is used to conduct spatial information retrieval.

The level of overlap between the directions of the semantic vectors of the concepts is determined by the similarities between ontology properties. Different similarity calculation methods are used for different types of ontology properties. For the OSM model, the proposed ontology properties are classified into six types: Boolean, synonymous, hierarchical, numerical, descriptive and spatial ontology properties.

(a) The values of the Boolean properties of the concepts can only be true or false (represented by 1 or 0, respectively). For example, whether a road is a one-way road is a Boolean property. The equation for calculating the similarity between Boolean properties is defined as follows:

$$\Phi_{sim}^B(v^c, v^{c'}) = \begin{cases} 1 & v^c = v^{c'} \\ 0 & v^c \neq v^{c'} \end{cases}, \quad (1)$$

where the superscript *B* indicates Boolean properties; v^c represents the value of the Boolean property item of concept *C*; and $v^{c'}$ represents the value of the Boolean property item of concept *C'*.

(b) Synonymous relationships are a type of equivalence relationship that can be evaluated by introducing an external source (e.g., a thesaurus). In this study, the similarity between synonyms is considered to be slightly less important than the similarity between the same properties. For example, a metro station and a subway station can both be used to describe a railway station in a rapid transit system. Referencing the method for calculating text properties by using WordNet to calculate similarity [21,22], the equation for calculating the similarity between synonymous properties is defined as follows:

$$\Phi_{sim}^{syn}(v^c, v^{c'}) = \begin{cases} 1 & v^c \text{ and } v^{c'} \text{ are the same} \\ \text{WordNet}_{sim} & v^c \text{ and } v^{c'} \text{ are synonymous} \end{cases} \quad (2)$$

where the superscript *syn* indicates synonymous properties and WordNet_{sim} represents the calculated similarity between two synonyms in WordNet.

(c) Because there are hierarchical semantic relationships between concepts, there are inevitably hierarchical relationships between the ontology properties that describe the features of the concepts. Semantic relationships in an ontology structure mainly include hypernym–hyponym relationships and whole–part relationships. Hypernym–hyponym relationships are semantic relationships that describe property values that have common characteristics at different logical levels. In comparison, whole–part relationships describe the compositional and structural associations between concepts. Consider a water system class as an example. Lakes and rivers are both subclasses of the water system class. Therefore, there is a certain similarity relationship between these two concepts. Referencing a layer-depth similarity calculation method [23], the equation for calculating the similarity between hierarchical properties is defined as follows:

$$\Phi_{sim}^H(v^c, v^{c'}) = \frac{\sum_{i=1}^n v^c \times v^{c'}}{\sqrt{\sum_{i=1}^n (v^c)^2} \times \sqrt{\sum_{i=1}^n (v^{c'})^2}}, \quad (3)$$

where *H* indicates hierarchical properties and v^c and $v^{c'}$ represent vector matrices of the associated ontology properties of concepts *C* and *C'* in the hierarchical model, respectively.

(d) In numerical ontology property items, specific numerical values are used to describe certain features of geographic information concepts (e.g., water flow rate, road width, and lake area). The semantics of numerical properties can be compared to determine their similarity if they have the same units; otherwise, their semantics cannot be compared. Referencing the data normalization standard method [24], the equation for calculating the similarity between numerical ontology properties is defined as follows:

$$\Phi_{sim}^M(v^c, v^{c'}) = 1 - \frac{|v^c - v^{c'}|}{\text{MAX}(v^c, v^{c'})}, \quad (4)$$

where *M* indicates numerical properties and *MAX* indicates that the maximum value is selected from the set.

(e) For the ontology properties, there is one type of property that describes the features of concepts in a language. For example, the values of the function property items are a set of words describing the functions of the geographic information concept in question. The level of similarity between descriptive properties is determined by the number of common morphemes. Because of the structural features of Chinese words, traditional text similarity assessment (e.g., edit distance) would hamper the quantification of the semantic similarity [25]. Referencing the morpheme center of gravity calculation method [26], the equation for calculating the similarity between descriptive properties is defined as follows:

$$\Phi_{sim}^T(v^c, v^{c'}) = \frac{1}{2}\alpha\left(\frac{k}{m} + \frac{k}{n}\right) + \frac{1}{2}\beta\text{MIN}\left(\frac{m}{n} + \frac{n}{m}\right)\left(\frac{\sum_{i=1}^c L_1(i)}{\sum_{t=1}^m t} + \frac{\sum_{i=1}^c L_2(i)}{\sum_{p=1}^n p}\right), \quad (5)$$

where T indicates descriptive properties; α represents the weight of the number of common morphemes; β represents the weight of the impact of the locations of the common morphemes in the words ($\alpha + \beta = 1$); m and n represent the character lengths of v^c and $v^{c'}$, respectively; k represents the number of characters matched between v^c and $v^{c'}$; and $L_1(i)$ and $L_2(i)$ represent the locations of the matched character i in v^c and $v^{c'}$ in left-to-right order, respectively. For example, for “economic information management” and “commercial information management”, the length of the characters in both sentences is 3, the number of matching characters between the two is 2, and the positive order values of matching characters are 2 and 3. Thus, the similarity is $\Phi_{sim}^T(v^c, v^{c'}) = \frac{2\alpha}{3} + \frac{5\beta}{6}$.

(f) Spatial semantic properties are mainly determined by the spatial association and shape similarity between geo-entities (e.g., interconnected roads and adjacent street blocks). In this study, the similarity between spatial semantics is estimated mainly for four common spatial features, namely distance, shape, size, and topology features [27]. The estimation equation is defined as follows:

$$\Phi_{sim}^G(v^c, v^{c'}) = \frac{1}{p} \sum_{i=1}^p \sigma_i(v^c, v^{c'}), \quad (6)$$

where G indicates spatial properties; p represents the number of common spatial features of spatial entities C and C' ($p \in [0, 4]$ and $p \in N^+$); and $\sigma_i(v^c, v^{c'})$ represents the similarity between the corresponding features. For example, for point targets and linear targets, the common spatial properties are the distance and spatial relationship (Figure 2), and we calculate them separately.

The similarity of the distance feature is inversely proportional to the distance. If the distance exceeds a certain threshold, the similarity is 0. For topological features, the similarity is 0 for disjoint and 1 otherwise.

(g) Because different ontology properties have different levels of importance for element classification, a weighted component is added to the proposed model to describe the importance of the specific ontology properties. The vector structure of the set of ontology properties is defined as follows:

$$Vector = \sum_1^n w_i * \Phi_{sim}^i(v^c, v^{c'}). \quad (7)$$

In the above formula, $Vector$ is the ontology similarity, v is the ontology attribute value, and W_i is the weight value determined by the analytic hierarchy process (AHP).

AHP is a common method for determining weights. First, we structure the decision hierarchy from the top with the similarity and then the ontology attribution weights. We then follow the advice of relevant experts to grade the importance of each attribute from 1 to 9. After constructing a set of pairwise comparison matrices, we calculate the feature vector of the pairwise comparison matrices. We take the feature vector as a weight vector if it passes the consistency test. For convenience, we use a software named “yaahp” to implement the calculation process. Then, we obtain the weight values of different components.

Information entropy is a method to measure the uncertainty of information that can be used to evaluate its influencing factors. Li et al. proposed an approach for matching instances by integrating heterogeneous attributes with the allocation of suitable attribute weights via information entropy [28]. The approach calculates the spatial, text and category attribute similarity between geographic entities and then determines the optimal weights of attribute similarity based on information entropy. The information entropy-based method measures the similarity of geographic entities accurately and is used for comparison.

Based on these equations, a set of relevant keywords can be obtained. This set is then compared with a set of concepts or entities corresponding to the keywords, and the levels of similarity between the words in the two sets are determined and sorted in descending order. Afterwards, based on this

sorted order, the words in the set obtained using the equations in this section are added to the original set of keywords to perform a second query.

3.3. Geo-Data Matching

The basic idea of the ontology similarity-based geographic information query and matching method is described here. First, the relevant descriptive information is extracted from the description of the geographic information retrieval (e.g., keywords, geo-names, and location information). Then, the semantics of the descriptive information are expanded based on the semantic similarity algorithms described in Section 3.2 and the OSM information ontology. Finally, based on the semantically expanded description of the geographic information query, a main set of query results is obtained. Afterwards, based on the order of relevance and the level of match with the user’s request, a certain number of query results are returned.

Based on the user-input query information, a set of query-related concepts, $C = (c_1, c_2, c_3, \dots, c_n)$, is extracted. Then, another set of concepts (C'), which are extensions related to the concepts in set C , is calculated using similarity algorithms. Sets C and C' are combined, and the resulting set is sent to the backend query system to produce a set of matched geo-entities, $E = (e_1, e_2, e_3, \dots, e_m)$. Afterwards, an extended entity set (E') related to set E is calculated using Equation (6). The results are then ranked and returned to the user. Figure 3 shows the flowchart of the ontology concept similarity-based geographic information query-service matching algorithm.

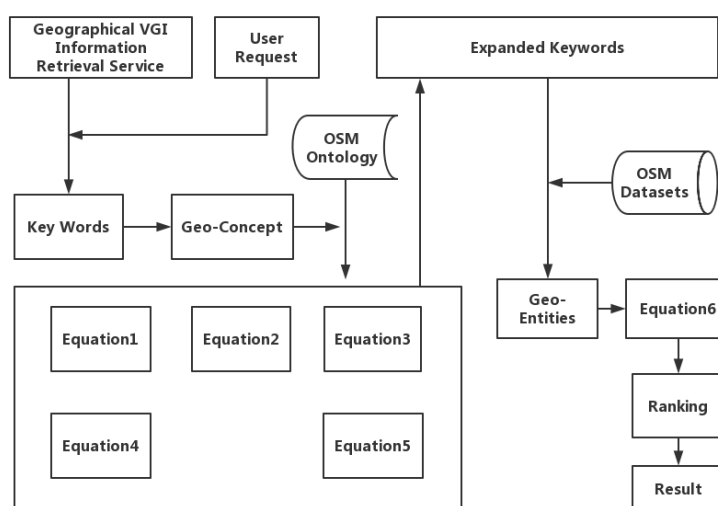


Figure 3. Flowchart of the geographic information query-service matching algorithm.

For example, if the concept term “permanent lake” is used as a retrieval keyword, the semantic properties of this concept can be obtained through the OSM ontology model built in Section 3.1, as shown in Table 2.

Table 2. List of the semantic properties of “permanent lake”.

Materiality	Formative	Spatiality		Timeliness		Feature
		Space Shape	Spatial Location	Periodicity	Life Cycle	
Water	Natural	Pit shape	Land surface	No	Normal period	Flowing water, storage water, aquaculture

In Table 2, the semantic properties of space shape and life cycle have hierarchy. Using the defined property similarity algorithm, the most similar concepts to “perennial lake” are “lake” and “pond”,

and their semantic similarity to “perennial lake” is 0.55, followed by “inland lake” and “outflow lake”, with a semantic similarity to “perennial lake” of 0.48. According to the results of similarity calculations, “lake”, “pond”, “inland lake”, and “outflow lake” can also be used as matched concept types during space information retrieval for the expansion query.

3.4. Semantic Similarity Calculation Module

Assessment of information retrieval is an activity of an information retrieval system to satisfy the user’s information requirement ability. To better evaluate the space information semantic-based query expansion method, a semantic similarity calculation module is designed for quantitative evaluation.

Relevant operators are determined according to the six semantic similarity calculation formulas proposed in Section 3.2 and then combined into a semantic similarity calculation module. When the data are transmitted to this module, the attributes must be read first to make a judgment according to the type of attribute. Next, the data are put into different types of similarity operators once for calculation to obtain the similarity values among different types of attributes. The similarity is a decimal that ranges from 0 to 1. Then, the overall similarity results can be obtained through comprehensive calculation based on the weight values calculated by the AHP method.

3.5. Evaluation Index

In general, in the information retrieval field, the standard recall, precision and F-measure are used as the main evaluation indices to evaluate search results [29]. Therefore, these three evaluation indices are selected for evaluating the search results in this study. The equations for calculating the recall, precision and F-measure are defined as follows, where RRE represents the number of relevant geo-entities in the query results; ARE represents the number of all relevant geo-entities in the database; RRE is part of ARE ; and RE represents the number of retrieved geo-entities:

$$Recall = \frac{RRE}{ARE} \times 100\%. \quad (8)$$

$$Precision = \frac{RRE}{RE} \times 100\%. \quad (9)$$

The F-measure comprehensively reflects both the recall and precision and; therefore, can satisfactorily reflect the query performance.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}. \quad (10)$$

In the information retrieval field, the order of returned results is very important for the user to obtain accurate information. Generally, users only pay attention to the top query results [30]. Because of the limitation on the number of query results that can be presented on one page, users normally only pay attention to the query results on the first one or two pages. Therefore, an auxiliary evaluation index, $AP@n$ [31], is introduced to evaluate the performance of the proposed query expansion method in ranking the relevant real query results. $AP@n$ represents the average precision based on the relevant documents in the first n number of query results and is calculated using the following equation:

$$AP@n = \frac{1}{RRE@n} \sum_{rank_i}^i (i \leq n), \quad (11)$$

where $RRE@n$ represents the number of relevant geographic entities in the first n number of query results and $rank_i$ represents the rank of the relevant result i in the first n number of query results.

4. Results

4.1. Recall Rate and Precision Rate

In this study, open access data related to China were downloaded in October 2016 from the official OSM website as the data source. Statistical analysis revealed that the experimental data set contained 34,147,539 nodes, 2,502,525 ways, and 30,361 relations. Spatial information and gazetteer semantic dictionaries were constructed to achieve better performance in splitting query sentences and query intention.

In the experiment, the non-query expansion method (i.e., initial search results provided by the official OSM retrieval service), the conventional information entropy-based query expansion method, and the semantic-based query expansion method proposed in this study were compared. All results were sorted using a conventional vector space model.

Twenty queries were performed for different regions of China and query keywords. The queries included commonly searched objects such as tourist attractions, transportation sites, roads, schools, and commercial facilities. We screened the correct answers from the data set manually before the experiment. In order to make the answers more objective, we followed specific rules when screening. In the data preprocessing stage, we constructed minimum bounding rectangles and a quadtree spatial index for each element. We eliminated irrelevant results with low spatial associations with the target in the screening stage. At the same time, we considered a variety of relationships according to the type of query target. For example, for query targets like roads, we also considered ancillary facilities such as gas stations.

The query keywords were divided into two groups, with one group representing geo-names and the other representing geo-entities. The geo-names came from the gazetteer semantic dictionaries, and geo-entities came from the downloaded OSM experimental data set. Tables 3 and 4 list the query keywords and average query results, respectively.

Table 3. Query keywords and their involved regions.

Group	Geo-Names		Geo-Entities	
	Key Words	Location	Key Words	Location
1	Optics Valley Square	Hubei Province	Moon Lake	Hubei Province
2	Wuhan University	Hubei Province	Yangtze Grand Bridge	Hubei Province
3	Three Gorges Dam	Hubei Province	Highway No. S14	Hubei Province
4	Sun Yat-Sen Mausoleum	Jiangsu Province	Tianhe Airport	Hubei Province
5	Old Summer Palace	Beijing	Hongli Road	Shenzhen
6	Canton Tower	Guangzhou	Yalu River	Jiling Province
7	West Lake	Zhejiang Province	East Third Ring Road	Beijing
8	Ultima Thule	Hainan Province	Wanquan River	Hainan Province
9	The Bund	Shanghai	Five Old Men Peaks	Jiangxi Province
10	Potala Palace	Tibet	Desert Road	Xinjiang

Table 4. Comparison of the results obtained using various query methods.

Retrieval Method	Mean Recall	Mean Precision	Mean F-measure
Non-Expansion	43%	49.2%	45.9%
Information Entropy-Based	65.4%	75.8%	70.2%
Semantic-Based (Proposed)	80.5%	89.3%	84.7%

Table 4 shows that the semantic-based query expansion method proposed in this study has higher query performance than the non-query expansion method and the information entropy-based query expansion method. In addition, compared to the query expansion method that is solely based on entropy, the proposed method exhibits significantly higher precision while ensuring recall. This higher precision occurs because the query is expanded for a second time in the query expansion process

based on the extended concepts obtained by analyzing the similarities between semantic properties based on the OSM ontology model. Then, the spatial semantic properties are subjected to relevance analysis based on the characteristics of spatial information. These steps improve the accuracy of the query results.

For simple queries of geo-entities, because the semantic properties of spatial features are considered and the coexistence relationships between geographic information concepts (i.e., dependence relationship) are sorted when constructing the OSM ontology model, the spatial feature properties of geospatial entities are expanded when extending the tag information of the OSM data. As a result, compared to the non-query expansion method and the information entropy-based query expansion method, the geospatial entity query results obtained using the proposed method are more in line with the expected results, and the relevant geo-entities retrieved by the proposed method are more comprehensive. This improvement in the query results is reflected in the data, as the proposed method significantly outperforms the other two methods in terms of recall, precision, and F-measure. Figure 4 shows a comparison of the query results of the three methods for geo-names and geo-entities for 10 different groups of keywords.

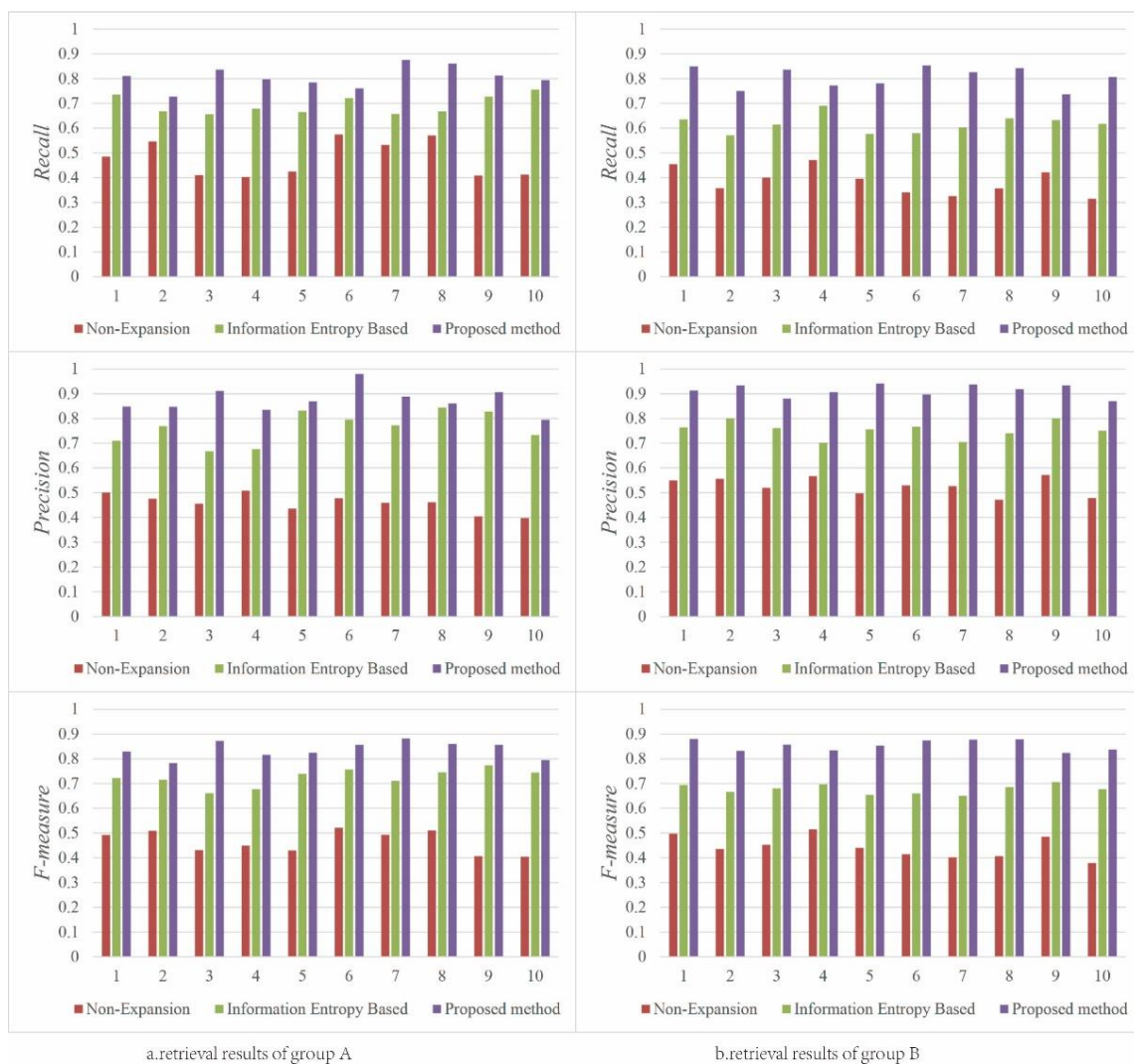


Figure 4. Comparison of the query results for geo-names and geo-entities for various methods: (a) Retrieval result of normal geo-names; (b) retrieval result of normal geo-entities.

Figure 4a clearly shows that the proposed method produces relatively good search results. This method has an average recall of approximately 80% and an average precision of approximately

87% for queries of general geo-entities, and these values are approximately 10% higher than those of the non-query expansion method and the information entropy-based query expansion method. Figure 4b shows the query results for geospatial information entities. Because of the query expansion on spatial feature semantics, the proposed method produces significantly better query results than the other two methods. The proposed method has an average recall of greater than 75% and an average precision of greater than 86%, which are approximately 20% higher than those of the other two methods.

4.2. Top n Results of a Query

The experimental results obtained for the keywords in the 10 groups are used to analyze the performance of the $AP@n$ for the three methods. Based on the users' search habits, n is set to 20. Figure 5 shows a comparison of the performance of the three methods in terms of $AP@20$ (the average precision based on the relevant documents in the first 20 number of query results).

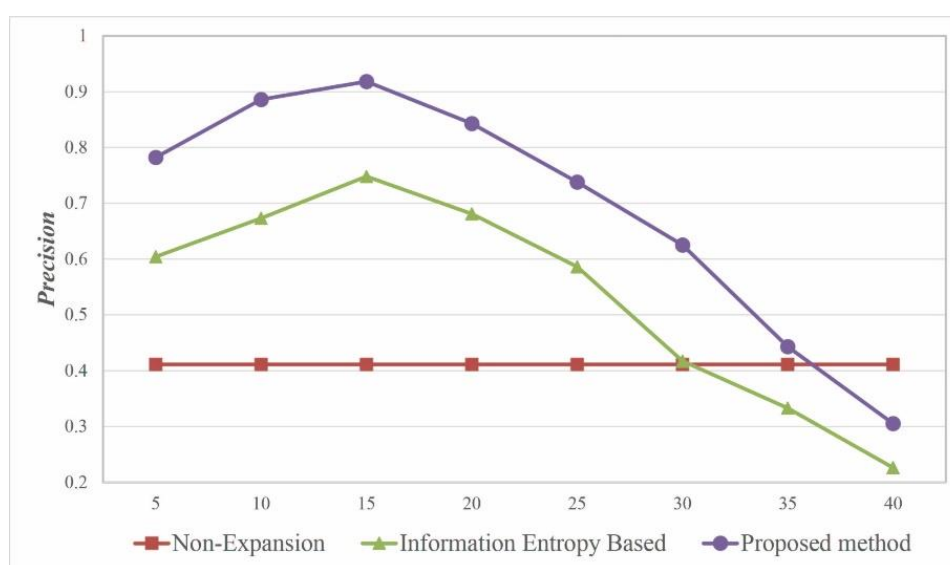


Figure 5. Comparison of the performance of the three methods in terms of $AP@20$.

Overall, the proposed method outperforms the other two methods in ranking the first 20 relevant results.

4.3. Optimal Number of Extension Words

Figure 6 shows the effect of the scale of expansion on query performance based on an analysis of the precision. When there are approximately 15 expanded keywords, the information entropy-based query expansion method and the proposed semantic-based query expansion method both exhibit relatively good query performance. When there are more than 35 expanded keywords, the query performance of the proposed method in this study is inferior to that of the non-query expansion method. Therefore, the optimal threshold for the number of expanded keywords is approximately 15.

The experimental results show the following: (1) The proposed semantic-based query expansion method is effective and efficient for VGI retrieval. (2) The proposed method considerably outperforms a non-query expansion method and an information entropy-based query expansion method in terms of recall and precision. (3) For queries on geo-entities, the proposed method exhibits significantly higher query performance than the other two methods because it considers the expansion of the spatial semantic properties. In addition, the method achieves optimal performance when the number of expanded query keywords is set to approximately 15. (4) The query results obtained using the method in this paper are ranked toward the top and are more aligned with the expected results.

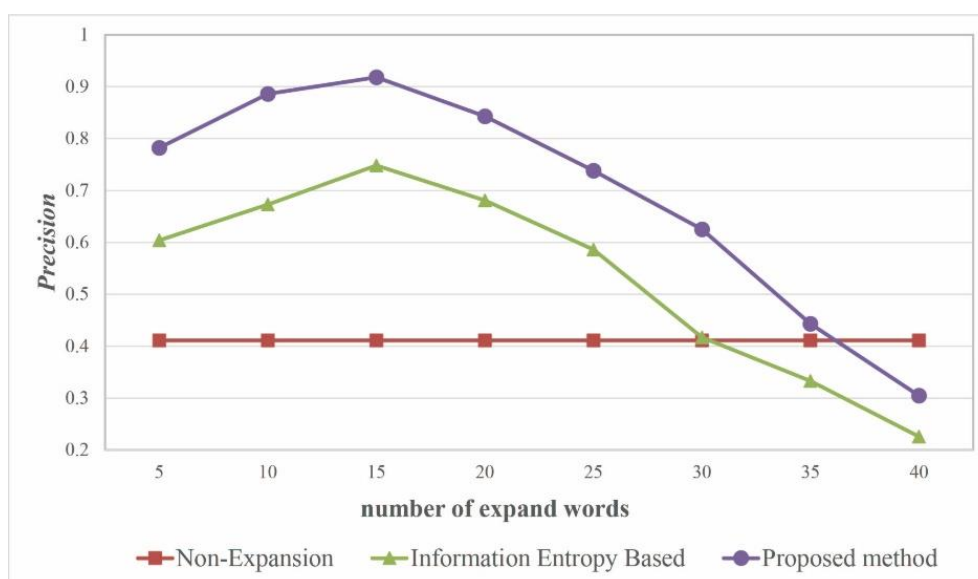


Figure 6. Relationship between the number of expanded keywords and the precision.

5. Discussion

In addition, the proposed method is validated through experimentation based on OSM data related to China. Compared to the non-query expansion method and the information entropy-based query expansion method, our method clearly produces relatively good performance. In addition, the order of returned results is very important in the information retrieval field. The proposed method eliminates irrelevant results with low spatial association in the subsequent spatial semantic similarity estimation process. This result suggests that the keywords expanded in the query expansion process by the proposed method are more closely aligned with the expected results.

The scale of the expansion is another important factor that affects the query performance. If the number of expanded keywords is too small, they cannot sufficiently represent the query requirements of the user; if the number of expanded keywords is too large, they will reduce the query result accuracy.

As shown in Figure 4a, the information entropy-based method and the proposed method have similar performance in groups 6 and 10. These results suggest that there is no significant difference between the numbers of relevant query results. Thus, our semantic model appears to have limited enhancement compared to the information entropy-based method for queries about tourist attractions. However, these problems could be solved if we adjust the weight in formula 6 according to the type of keywords.

Furthermore, limited by time and manpower, the current experiment was based on a small sample. In order to test the performance of the method comprehensively, our keyword sample covered common query categories. From the point of view of type, our sample consisted of point features, line and polygon features. And in the view of category, our samples included transportation facilities, tourist attractions, public facilities, etc. However, there may be some limitations of this study have not been exposed that can also provide directions for further research.

6. Conclusions and Future Work

In this study, using the tags in OSM data combined with spatial information semantic features, an OSM ontology model with semantic property items was constructed, a corresponding semantic similarity algorithm model was designed, and a query expansion method was proposed based on the similarity of the OSM ontology semantic properties. This method aims to address VGI retrieval with extensible tags and defines standard semantic expansion property items. In addition, the proposed method was validated through experimentation based on OSM data related to China.

Further research is necessary, and our future work will focus on several areas, including improving the framework of the semantic model for VGI proposed in this study and optimizing the query result ranking algorithm, to achieve an efficient query method that is more suitable for VGI.

Author Contributions: Conceptualization, Tao Sun and Lin Li; methodology, Tao Sun; software, Hui Xia; Validation, Tao Sun and Hui Xia; formal analysis, Hui Xia; Resources, Hang Shen; data curation, Hang Shen; writing—original draft preparation, Tao Sun; writing—review and editing, Tao Sun and Yu Liu; visualization, Hang Shen; supervision, Lin Li; project administration, Lin Li; funding acquisition, Lin Li.

Funding: This study is funded by the National Key R&D Program of China, grant number 2017YFB0503701.

Acknowledgments: The authors would like to thank editors and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Kobayashi, S.; Fujioka, T.; Tanaka, Y.; Inoue, M.; Niho, Y.; Miyoshi, A. A geographical information system using the google map api for guidance to referral hospitals. *J. Med. Syst.* **2009**, *34*, 1157–1160. [[CrossRef](#)] [[PubMed](#)]
2. Purves, R.; Jones, C. Geographic information retrieval (gir). *Comput. Environ. Urban Syst.* **2006**, *30*, 375–377. [[CrossRef](#)]
3. Jones, C.B.; Alani, H.; Tudhope, D. *Geographical Information Retrieval with Ontologies of Place*; International Conference on Spatial Information Theory; Springer: Berlin, Heidelberg, 2001; pp. 322–335.
4. Bergmann, R.; Gil, Y. Similarity assessment and efficient retrieval of semantic workflows. *Inf. Syst.* **2014**, *40*, 115–127. [[CrossRef](#)]
5. Wei, C.-P.; Hu, P.J.-H.; Tai, C.-H.; Huang, C.-N.; Yang, C.-S. Managing word mismatch problems in information retrieval: A topic-based query expansion approach. *J. Manag. Inf. Syst.* **2014**, *24*, 269–295. [[CrossRef](#)]
6. Singh, J.; Sharan, A. Relevance feedback-based query expansion model using ranks combining and word2vec approach. *IETE J. Res.* **2016**, *62*, 591–604. [[CrossRef](#)]
7. Bai, L.; Yan, L.; Ma, Z.M. Determining topological relationship of fuzzy spatiotemporal data integrated with xml twig pattern. *Appl. Intell.* **2012**, *39*, 75–100. [[CrossRef](#)]
8. Alazzawi, A.N.; Abdelmoty, A.I.; Jones, C.B. An ontology of place and service types to facilitate place-affordance geographic information retrieval. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18–19 February 2010; pp. 1–2.
9. Li, L.; Liu, Y.; Zhu, H.; Ying, S.; Luo, Q.; Luo, H.; Kuai, X.; Xia, H.; Shen, H. A bibliometric and visual analysis of global geo-ontology research. *Comput. Geosci.* **2017**, *99*, 1–8. [[CrossRef](#)]
10. Cardoso, N.; Silva, M.J. Query expansion through geographical feature types. In Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, Lisbon, Portugal, 9 November 2007; pp. 55–60.
11. Derbal, K.; Bordogna, G.; Pasi, G.; Alimazighi, Z. Spatial Querying Supported by Domain and User Ontologies: An Approach for Web GIS Applications. In *Flexible Query Answering Systems 2015*; Springer: Berlin, Heidelberg, 2016; pp. 353–365.
12. Fu, G.; Jones, C.B.; Abdelmoty, A.I. OTM Confederated International Conferences “on the Move to Meaningful Internet Systems”. In *Ontology-Based Spatial Query Expansion in Information Retrieval*; Springer: Berlin, Heidelberg, 2005; pp. 1466–1482.
13. Brown, G.; Kelly, M.; Whittall, D. Which ‘public’? Sampling effects in public participation gis (ppgis) and volunteered geographic information (vgi) systems for public lands management. *J. Environ. Plan. Manag.* **2014**, *57*, 190–214. [[CrossRef](#)]
14. OSM. Osm Mapping Projects. Available online: http://wiki.openstreetmap.org/wiki/Mapping_projects (accessed on 27 August 2017).
15. Codescu, M.; Horsinka, G.; Kutz, O.; Mossakowski, T.; Rau, R. Osmonto—an ontology of openstreetmap tags. In Proceedings of the State of the Map Europe (SOTM-EU) Conference, Vienna, Austria, 15–17 July 2011.
16. Lopez-Pellicer, F.J.; Silva, R.J.; Chaves, M. Linkable geographic ontologies. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18–19 February 2010; pp. 1–8.

17. Janowicz, K.; Schade, S.; Bröring, A.; Keßler, C.; Maué, P.; Stasch, C. Semantic enablement for spatial data infrastructures. *Trans. GIS* **2010**, *14*, 111–129. [[CrossRef](#)]
18. Kauppinen, T.; Henriksson, R.; Sinkkilä, R.; Lindroos, R.; Väätäinen, J.; Hyvönen, E. *Ontology-Based Disambiguation of Spatiotemporal Locations*; IRSW: Tenerife, Spain, 2008.
19. Carpineto, C.; Romano, G. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv. CSUR* **2012**, *44*. [[CrossRef](#)]
20. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Geographic knowledge extraction and semantic similarity in openstreetmap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81. [[CrossRef](#)]
21. Zhu, X. A novel wordnet-based approach for measuring semantic similarity. *J. Inf. Comput. Sci.* **2015**, *12*, 4919–4927. [[CrossRef](#)]
22. Gao, J.-B.; Zhang, B.-W.; Chen, X.-H. A wordnet-based semantic similarity measurement combining edge-counting and information content theory. *Eng. Appl. Artif. Intell.* **2015**, *39*, 80–88. [[CrossRef](#)]
23. Liu, H.; Bao, H.; Xu, D. Concept vector for semantic similarity and relatedness based on wordnet structure. *J. Syst. Softw.* **2012**, *85*, 370–381. [[CrossRef](#)]
24. Ponniah, P. Data Normalization Method. In *Database Design and Development*; IEEE Press: Piscataway, NJ, USA, 2015.
25. Cao, J.; Wu, -X.; Xia, Y.; Zheng, F. Pinyin-indexed method for approximate matching in Chinese. *J. Tsinghua Univ. Sci. Technol.* **2009**, *S1*, 1328–1332. (In Chinese)
26. Wu, Z.-Q. The Development of Post-Control Words during Economical Information Retrieval. Master's Thesis, Nanjing Agricultural University, Nanjing, China, 1999.
27. An, X.-Y.; Sun, Q.; Xiao, Q. A Shape Multilevel Description Method and Application in Measuring Geometry Similarity of Multi-scale Spatial Data. *Acta Geod. Cartogr. Sin.* **2011**, *40*, 495–502.
28. Li, L.; Xing, X.; Xia, H.; Huang, X. Entropy-weighted instance matching between different sourcing points of interest. *Entropy* **2016**, *18*, 45. [[CrossRef](#)]
29. Büttcher, S.; Clarke, C.L.; Cormack, G.V. *Information Retrieval: Implementing and Evaluating Search Engines*; MIT Press: Cambridge, MA, USA, 2016.
30. Spink, A.; Wolfram, D.; Jansen, M.B.J.; Saracevic, T. Searching the web: The public and their queries. *J. Am. Soc. Inf. Sci. Technol.* **2001**, *52*, 226–234. [[CrossRef](#)]
31. Karypis, G. Evaluation of Item-Based Top-N Recommendation Algorithms. In Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM), Atlanta, GA, USA, 5–10 November 2001.



© 2019 by the authors. 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 (<http://creativecommons.org/licenses/by/4.0/>).