

Review

Progress and Challenges on Entity Alignment of Geographic Knowledge Bases

Kai Sun ^{1,2,3} , Yunqiang Zhu ^{1,2,4} and Jia Song ^{1,2,4,*} 

¹ State Key Laboratory of Resources and Environmental Information System, Beijing 100101, China; sunk@reis.ac.cn (K.S.); zhuyq@reis.ac.cn (Y.Z.)

² Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

³ University of Chinese Academy of Sciences, Beijing 100049, China

⁴ Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

* Correspondence: songj@igsnr.ac.cn; Tel.: +86-010-6488-9906

Received: 23 November 2018; Accepted: 29 January 2019; Published: 6 February 2019



Abstract: Geographic knowledge bases (GKBs) with multiple sources and forms are of obvious heterogeneity, which hinders the integration of geographic knowledge. Entity alignment provides an effective way to find correspondences of entities by measuring the multidimensional similarity between entities from different GKBs, thereby overcoming the semantic gap. Thus, many efforts have been made in this field. This paper initially proposes basic definitions and a general framework for the entity alignment of GKBs. Specifically, the state-of-the-art of algorithms of entity alignment of GKBs is reviewed from the three aspects of similarity metrics, similarity combination, and alignment judgement; the evaluation procedure of alignment results is also summarized. On this basis, eight challenges for future studies are identified. There is a lack of methods to assess the qualities of GKBs. The alignment process should be improved by determining the best composition of heterogeneous features, optimizing alignment algorithms, and incorporating background knowledge. Furthermore, a unified infrastructure, techniques for aligning large-scale GKBs, and deep learning-based alignment techniques should be developed. Meanwhile, the generation of benchmark datasets for the entity alignment of GKBs and the applications of this field need to be investigated. The progress of this field will be accelerated by addressing these challenges.

Keywords: geographic knowledge bases; entity alignment; similarity metrics; similarity combination; knowledge conflation; knowledge integration

1. Introduction

Geographic knowledge bases (GKBs) are the formal and explicit representation of geographic entities and their mutual semantic relationships. The formalized and interconnected geographic knowledge provided by GKBs can help to achieve high-quality and intelligent geographic information services, such as semantic search and personalized recommendations of geographic information, and enhance human–machine interaction (e.g., intelligent question answering). GKBs are also important knowledge sources for constructing geographic knowledge graphs [1–4], which is an important infrastructure for the applications of artificial intelligence (AI) in GIScience.

GKBs have emerged with the continuous development of technologies such as the semantic web and linked data [5,6]. GKBs are mainly represented in three different forms, namely geographic semantic webs, geographic ontologies, and digital gazetteers. Geographic semantic webs (Table 1), which mainly include GeoNames (<http://www.geonames.org/>), LinkedGeoData

(<http://www.linkedgeodata.org/>), the OpenStreetMap (OSM) semantic network (https://wiki.openstreetmap.org/wiki/OSM_Semantic_Network), the Alexandria Digital Library gazetteer (ADL) (<http://legacy.alexandria.ucsb.edu/>), and GeoWordNet, represent and publish geographic knowledge in the format of a resource description framework (RDF). GeoNames is a commonly used database of global toponyms, and all toponyms are categorized into nine feature classes and further subcategorized into 645 feature codes. OpenStreetMap is represented with three basic elements, including nodes, ways, and relations, and all the features are categorized into 28 classes. LinkedGeoData and the OSM semantic network have the same data source as OpenStreetMap (<http://www.openstreetmap.org/>), and the difference between them is that the former performs RDF conversion on all the data of OpenStreetMap and the latter only contains concepts and properties. ADL provides normalized and standardized digital gazetteers and is divided into six feature classes. GeoWordNet is the result of a semantic fusion of GeoNames and WordNet (<https://wordnet.princeton.edu/>), which is currently inaccessible. Some general semantic webs also contain large quantities of toponymic data, such as DBpedia (<http://wiki.dbpedia.org/>).

Table 1. The numbers of concepts, properties, and instances in some geographic semantic webs (obtained on 18 January 2019).

Name	Number of Concepts	Number of Properties	Number of Instances	Formalized Format
GeoNames	654	28	11,809,910	OWL
LinkedGeoData	1222	137	3,000,000,000	NT
OSM Semantic Network	1222	137	Null	RDF
ADL	210	Null	8,000,000	RDF

Geographic ontologies [7,8], such as the semantic web for Earth and environmental terminology (SWEET) ontology, GeoNames ontology (<http://www.geonames.org/ontology/documentation.html>), and some other domain ontologies [9,10], also include substantial geographic knowledge. Digital gazetteers, such as the Getty thesaurus for geographic names (TGN) (<http://www.getty.edu/research/tools/vocabularies/tgn/>), the GEOnet names server (GNS) (<http://geonames.nga.mil/gns/html/>), and the geographic names information system (GNIS) (<https://geonames.usgs.gov/domestic/>), play an important role in the recognition and ambiguity resolution of geographic entities.

These GKBs with different sources and forms, however, are independent, resulting in a problem of locally well-organized but overall dispersed and independent “information islands” [11]. This problem is also termed semantic heterogeneity, which exists in the lexicon, structure, spatial position, category, shape, data-type of property, range of property, and property value of the entities from different GKBs. The semantic heterogeneity is a significant barrier for the integration of different GKBs [12]. GKBs integration is to form the resulting global GKBs by identifying mappings between local GKBs. The integrated global GKBs can solve overlapping and gap of knowledge, therefore providing more comprehensive and clear geographic knowledge to facilitate retrieval, integration, and exchange of geospatial information.

An effective way to deal with semantic heterogeneity is entity alignment (i.e., entity resolution [13,14], duplicate detection [15], record linkage [16], etc.). For two entities of the same type from different GKBs, entity alignment can judge the relationship between them and find correspondences by measuring the semantic correlation over multidimensional information, thereby eliminating the inconsistency, such as conflict and referent ambiguity of entities in the heterogeneous GKBs.

Entity alignment, a fundamental semantic technology, has become a research hotspot as knowledge bases provide an effective way to encode meaning of information [17,18]. In computer science, the last decades have witnessed extensive studies in this field [17], and many well-known alignment systems have been developed, such as the association rule ontology matching approach (AROMA) [19,20], AgreementMaker [21,22], as well as generic ontology matching and mapping

management (GOMMA) [23]. In GIScience, there have also been a great number of studies on this subject. The topics of these previous studies can be roughly separated into three categories: Similarity metrics, alignment techniques, and processing frameworks. The first category focuses on the similarity measures based on the heterogeneous features of entities. The second category includes the methods of similarity combination, alignment judgement, and result evaluation. The third category centers on the effective employment of the similarity metrics and alignment techniques.

While numerous studies exist and have led to many achievements, there is a lack of a unified definition and framework on the general concept, and a systematic summarization. To fill this gap, we present a formal and explicit framework, which could provide readers, especially novices, an introduction to understand the entity alignment as a field of study [24]. Based on this coherent framework, we organize the insights from previous studies and provide a comprehensive review to help readers to understand the existing algorithms and techniques. We also explore the key challenges for future works to facilitate progress in this field. This paper makes the following contributions:

- A formal and explicit coherent framework for the entity alignment of GKBs;
- A systematic classification and summarization of previous studies in terms of the algorithms of similarity metrics, similarity combination, alignment judgement, and result evaluation;
- A set of challenges for future research.

The remainder of this paper is structured as follows. In Section 2, the definitions and framework for the entity alignment of GKBs are formally defined and presented. In Section 3, the algorithms from previous studies are systematically examined from the three aspects of similarity metrics, similarity combination, and alignment judgement. Section 4 shows the methods and benchmarks for evaluating the alignment result. Section 5 articulates the key challenges for future research. Finally, Section 6 summarizes this study.

2. Definitions and Framework for Entity Alignment of GKBs

2.1. Basic Definitions

2.1.1. Problem Statement

The process for the entity alignment of GKBs can be simplified as shown in Figure 1. Given two entities e_1 and e_2 , from two GKBs, entity alignment is performed to determine whether they are matching pairs. Some optional parameters (e.g., the relevant weights or thresholds) and background knowledge (e.g., generally accepted knowledge bases or domain lexicons) can be used to facilitate this process.

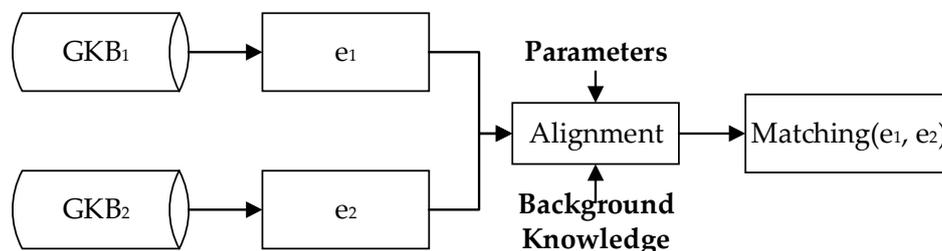


Figure 1. Schematic diagram for entity alignment of geographic knowledge bases (GKBs). (Note: GKB_1 and GKB_2 represent two geographic knowledge bases to be integrated, and e_1 and e_2 are two entities to be aligned from GKB_1 and GKB_2 , respectively.).

The entity sets from GKBs can be defined as in Equation (1):

$$E(GKB) = \{(GC \cup GP \cup GI) | GC, GP, GI \in GKB\}, \quad (1)$$

where $E(GKB)$ represents the entity set of a GKB and GC , GP , and GI represent the sets of geographic concepts, properties, and instances of the GKB, respectively.

The entity alignment of GKBs can be defined as in Equation (2):

$$Align(e_1, e_2) = \{Fun(Sim_1(e_1, e_2), Sim_2(e_1, e_2) \dots) ? Matching(e_1, e_2) : Null | e_1 \in E(GKB_1), e_2 \in E(GKB_2), Sim(e_1, e_2) \in [0, 1]\}, \quad (2)$$

where $Sim(e_1, e_2)$ represents the normalized similarity scores computed over heterogeneous features of two entities: e_1 and e_2 ; $Fun(Sim_1(e_1, e_2), Sim_2(e_1, e_2) \dots)$ represents the judging function for the relationship between them, leveraging their multidimensional similarity scores. $Matching(e_1, e_2)$ indicates that e_1 and e_2 are a matching pair, and $Null$ means that e_1 and e_2 are a nonmatching pair.

2.1.2. Explanation for Heterogeneities in Geographic Entities

The heterogeneities refer to the various differences that may arise between entities. Based on in-depth analysis of the connotations and extensions of the entities in GKBs, we identified eight types of heterogeneities: Heterogeneity in lexicon, structure, spatial position, category, shape, data-type of property, range of property, and property value. To provide a formal representation for them, we used the function $Diff(a, b)$ to represent the heterogeneities between a and b . Given two entities, $e_1 \in E(GKB_1)$ and $e_2 \in E(GKB_2)$, the eight types of heterogeneities can be defined as follows.

Heterogeneity in Lexicon (HL)

The lexical characteristics include the labels and comments of entities. HL can be defined as in Equation (3):

$$HL(e_1, e_2) = ((Diff(e_1.Label, e_2.Label)) \cup (Diff(e_1.Comment, e_2.Comment))), \quad (3)$$

where $e.Label$ and $e.Comment$ are the label and comment of entity e , respectively.

Heterogeneity in Structure (HS)

For a specific entity, structural characteristics refer to its linked entities, including hypernyms (Hypers), hyponyms (Hypos), and siblings (Siblings). HS can be defined as in Equation (4):

$$HS(e_1, e_2) = ((Diff(e_1.Hypers, e_2.Hypers)) \cup (Diff(e_1.Hypos, e_2.Hypos)) \cup (Diff(e_1.Siblings, e_2.Siblings))), \quad (4)$$

where $e.Hypers$, $e.Hypos$, and $e.Siblings$ are the hypernymic, hyponymic, and sibling nodes of entity e , respectively.

Heterogeneity in Spatial Position (HSp)

HSp refers to the difference in the spatial positions of entities and is defined as in Equation (5):

$$HSp(e_1, e_2) = Diff(e_1.Pos, e_2.Pos), \quad (5)$$

where $e.Pos$ is the spatial position of entity e .

Heterogeneity in Category (HC)

HC refers to the differences in the category to which the entities belong and is defined as in Equation (6):

$$HC(e_1, e_2) = Diff(e_1.Cat, e_2.Cat), \quad (6)$$

where $e.Cat$ is the category information of entity e .

Heterogeneity in Shape (HSh)

HSh refers to the difference in the shapes of entities and is defined as in Equation (7):

$$HSh(e_1, e_2) = Diff(e_1.Shape, e_2.Shape), \quad (7)$$

where $e.Shape$ is the shape of entity e . HSp is different than HSh. HSp is the difference in spatial distance between entities, which is usually computed over the spatial coordinates of entities. HSp exists in all types of geometry object, including point, polyline, and polygon objects. HSh just stands for the difference in geometric shapes, which has nothing to do with the spatial distance between entities. HSh exists only in two types of geometry object (i.e., polyline and polygon).

Heterogeneity in Data-Type of Property (HPdt)

There are two types of property: Data property and object property, and their data-types may be string, numerical value (for example int, float), entity, etc. HPdt refers to the heterogeneity of data-types of two properties and can be defined as in Equation (8):

$$HPdt(e_1, e_2) = Diff(Pro_1.dt, Pro_2.dt), \quad (8)$$

where $Pro.dt$ is the data-type of property Pro .

Heterogeneity in range of property (HPr)

The range of property refers to the range of values for a property. HPr represents the difference in the ranges of two properties and is defined as in Equation (9):

$$HPr(e_1, e_2) = Diff(Pro_1.range, Pro_2.range), \quad (9)$$

where $Pro.range$ is the range of property Pro .

Heterogeneity in Property Value (HPv)

HPv refers to that the values corresponding to the same property of two entities are different and is defined as in Equation (10):

$$HPv(e_1, e_2) = (Diff(e_1.Pro.Value, e_2.Pro.Value) \cap (e_1.Pro \equiv e_2.Pro)), \quad (10)$$

where $e.Pro$ is the property of entity e , and $e.Pro.Value$ is the value corresponding to this property.

The heterogeneities involved in specific entity pairs are closely related to the type of entity (concept, properties, and instance). The heterogeneities in each type of entities are analyzed, as shown in Table 2. The differences in concepts are reflected in their lexicon and structure, and the possible heterogeneities in properties include heterogeneity in their lexicon, structure, data-type, and range, and the heterogeneities in instances are reflected in their lexicon, space, category, shape, and property value.

Table 2. Heterogeneities in each type of entities.

Entity Type	Heterogeneities							
	HL	HS	HSp	HC	HSh	HPdt	HPr	HPv
Concepts	✓	✓						
Properties	✓	✓				✓	✓	
Instances	✓		✓	✓	✓			✓

2.2. General Framework

2.2.1. Basic Ideas

Entity alignment can be divided into one-dimensional and multidimensional entity alignment according to the number of entity types for alignment. One-dimensional entity alignment is to align one type of entities from concept, property or instance. Most studies are on concept-level alignment [25], followed by those on instance-level alignment [26], and property-level alignment [27,28]. Multidimensional entity alignment is to simultaneously align two or all three types of entities, and there are few studies on this subject. Only Yu leveraged the similarity scores computed over multidimensional features and the approval voting strategy to construct an integrated framework, which could simultaneously align all three types of entities [29].

The basic ideas of entity alignment are tightly associated with the relationship among different types of entities. The three types of entities (concept, property, and instance) are not isolated from each other and show a relationship of comparatively strong cognitive logic. Properties are descriptions of the characteristics of things, which are the first level to cognize things. According to the similarities and differences in properties, things that have common properties belong to the same class, which is called the concept. Things of the same class can be repartitioned according to the differences in their properties, and then the hypernymic and hyponymic concepts can be defined. The hyponymic concept inherits all the properties of the hypernymic concept. An object, which conforms to all the properties of a certain class, is an instance belonging to this class, indicating that the instance inherits all the properties of concept to which it belongs. Thus, the relationship among the three types of entity can be summarized as follows: Properties are the definition of concept, concept is the abstraction of instances, and instances are the instantiation of the concept and properties.

According to the above relationship among the three types of entity, the basic ideas for entity alignment can be divided into schema-level and instance-level ideas. The former mainly aims at concept alignment and includes three types of approaches. The commonly employed type of approach is instance-based concept alignment. It is based on the cognitive logic that instances are the instantiation of concept, and it is assumed that if the instances belonging to two concepts are matched, these two concepts are matched. Thus, this approach discriminates matched concept pairs or nonmatched concept pairs by measuring the similarity between the instances belonging to the concept pairs to be matched [30]. The second type of approach is property-based concept alignment. It is based on the cognitive logic that properties are the definition of concept and is on the basis of the assumption that if the properties corresponding to two concepts are aligned, these two concepts are aligned. Li et al. used water body ontology as a case study and summarized 17 properties shared by all the concepts in the water body ontology. The similarity scores among these properties were computed over the shared members of the range domain of properties [28]. Then, the aligned concept pairs can be identified according to the similarity scores of properties. This method can actually perform alignment leveraging the semantic information of properties. The third type of approach directly performs concepts alignment based on their own information [29]. Instance-level alignment mainly relies on the information of instances to perform alignment.

Specific implementation techniques are consistent across different basic ideas and can be divided into three types: Element-based, structure-based, and hybrid techniques [31]. Element-based techniques consider entities in isolation, leveraging information about entity itself to perform alignment [32]. Structure-based techniques examine the structural information of entities, considering the linked entities of an entity in the structure of knowledge bases [33]. Hybrid techniques combine element-based and structure-based techniques.

In addition, performing entity alignment with background knowledge, which involves enough entities in common with entities to be aligned, can help to improve the result. For example, the near-synonyms, synonyms, and hypernymic–hyponymic relationships among the vocabularies provided by WordNet can support the calculation of the lexical similarity of entities [34–37]. Involving

human experts can also optimize the alignment process [38]. Prior to performing entity alignment, preprocessing steps can be adopted to remove entity pairs that are impossible to be matched or select potential matching entity pairs, thereby avoiding performing calculations over all entity pairs and reducing computational complexity. This is especially important when the number of entity pairs to be matched is large. For example, a blocking algorithm is performed to group possible matching entities into one block according to the similarity of literal description of entities to reduce the computational requirement [39].

2.2.2. Standard Workflow

The standard workflow for the entity alignment of GKBs is consistent across different basic ideas and implementation techniques. It includes four steps, as shown in Figure 2:

- Step 1. Similarity measurement. Determining suitable similarity metrics for each type of heterogeneities in entities.
- Step 2. Similarity combination. Selecting an effective method to combine multidimensional similarity scores.
- Step 3. Alignment judgement. Taking a decision for entity pairs to be matched based on a predefined threshold or leveraging an effective judging approach.
- Step 4. Result evaluation. Using suitable benchmarks and evaluation metrics to assess result quality.

The next two sections detail the algorithms from previous studies about the four steps.

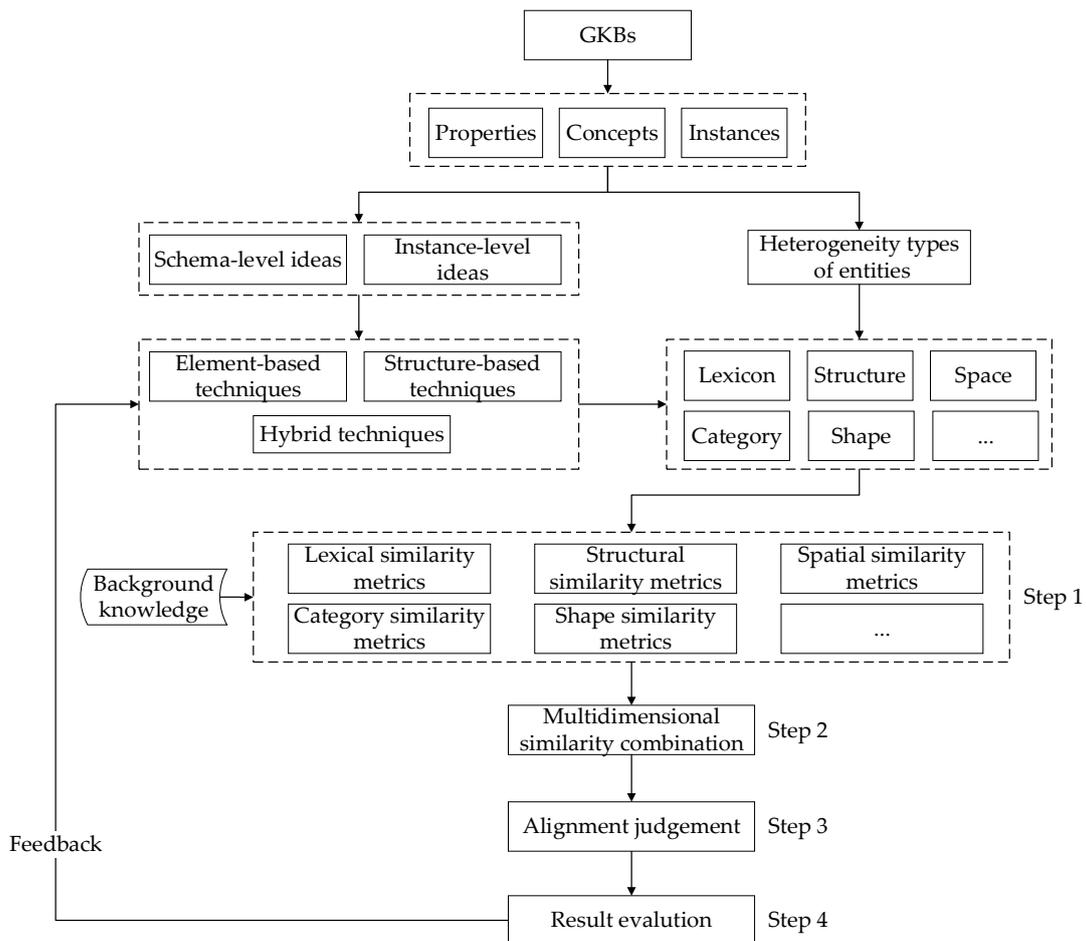


Figure 2. Standard workflow for entity alignment of GKBs.

3. Algorithms of Entity Alignment

This section analyzes and reviews the algorithms of entity alignment in detail from the three aspects of similarity metrics, similarity combination, and alignment judgement.

3.1. Similarity Metrics

Although there are eight types of heterogeneities in entities from different GKBs, the first five types are commonly used. Previous studies targeted the concepts and instances as the main entity types to align and ignored the alignment of properties. Moreover, it is difficult to define essential properties, which can actually represent the connotation of concepts and the range domain of properties. Thus, the latter three types, which focus on the differences in geographic properties, are extremely rarely used or have never been leveraged for the aligning entity of GKBs. Thus, we only conducted a review on the similarity metrics for lexicon, structure, spatial position, category, and shape.

3.1.1. Lexical Similarity Metrics

Lexical similarity metrics are the most useful metrics for entity alignment, although there are many other types of similarity metrics for different types of heterogeneities [40]. The lexical corpuses of geographic entities are composed of their labels and comments. Santos et al. and Recchia and Louwse conducted comprehensive comparisons of various lexical similarity metrics. They pointed out that the differences in performance of the involved lexical similarity metrics were relatively small [40,41]. The lexical similarity metrics can be roughly grouped into three categories, namely character-based, token-based, and vector-based metrics. Character-based metrics calculate lexical similarity by measuring the difference of two original strings. Token-based metrics extract tokens from original strings, based on which the similarity between original strings is computed. Vector-based metrics compute similarity based on vector representations of original strings.

A basic character-based metric is edit distance (Levenshtein distance) [42]. Edit distance leverages the minimum number of editing operations required to transform a string *a* into another string *b*. Allowable editing operations include insertion, deletion, and substitution. This metric is commonly employed in lexical similarity computations for geographic entities [43–45]. However, due to the various variabilities of gazetteers, such as abbreviations and merges of strings, edit distance is not particularly suitable for measuring the similarity of feature names [46]. The Jaccard similarity coefficient is also a commonly used metric. It measures similarity by the ratio of the number of characters in common between two strings to the total number of characters. Sehgal et al. measured the similarity of location names by this metric [26]. However, this metric does not take the character transposition of strings into consideration.

The Jaro metric is a heuristic character-based metric. It computes similarity based on the number of characters that are present in both two strings and the number of character transpositions. Auer et al. used this metric to measure the similarity of labels of entities in LinkedGeoData and DBpedia [47]. A refined version of the Jaro metric is the Jaro–Winkler metric, which considers the weight assignment of characters based on their different positions in strings. This metric suggests that the strings, which begin with the same characters, should be given a higher similarity score. Therefore, the matched beginning part of the strings should be assigned greater weight. Stadler et al. used this metric to measure the similarity among the labels of entities from LinkedGeoData, DBpedia, and GeoNames [48]. Martins also employed this metric in the detection of duplicate gazetteer records [43]. There are also other character-based metrics employed in the alignment of geographic entities, such as the soft-term frequency-inverse document frequency (Soft-TFIDF) distance, Monge–Elkan distance, and Double Metaphone distance [29,43].

As for the token-based metrics, standard steps for extracting tokens from strings include tokenizing (extracting the recognizable substrings and marked symbols from strings), removing punctuations, stemming, removing words in a stop words list (stop words refer to words which are

either insignificant or too common), and completing abbreviations. After this procedure is completed, the token set of strings can be obtained [49]. Given the token sets $\text{Token}(S_1)$ for string S_1 and $\text{Token}(S_2)$ for string S_2 , there are many algorithms for calculating the similarity between them. A simple method is to compute the ratio of the shared tokens of the two sets to the total numbers of the two sets [50], but this method does not tolerate the spelling errors of token words. A more complex method is to calculate the edit distance between all token pairs of the two token sets [49]. This method has high accuracy, but when the number of token pairs is comparatively large, it will suffer from high computational complexity. To simplify this method, for all the token pairs, a match score of 1, 0.5, 0.25, and 0 is assigned for four types of situations corresponding to exact match, prefix or stemmed match, infix match (partial existing in anywhere within a string, except in its beginning), and complete mismatch, respectively [51].

Standard vector-based metrics compute cosine similarity between vector representations of strings. Wang et al. constructed virtual documents for entities by combining their labels and comments and represented the virtual documents with a vector. Every element of this vector was weight computed using the TF-IDF method [52], and the weight was assigned to the corresponding words of virtual documents [45]. Then, the lexical similarity between two entities was computed as the cosine value between the two vectors of their virtual documents. A more sophisticated method proposed by Ballatore et al. extracted the definitional terms (nouns, verbs, adjectives, etc.) for geographic entities from their comments to construct semantic vectors. A matrix was constructed, each cell of which contained a similarity between two terms from the vectors [53]. The similarity scores between term pairs were computed with WordNet. Then, the matrix would be used to compute the vector–vector similarity scores. When the comments of entities are rich, this method is able to attain higher accuracy. However, the accuracy will be greatly reduced with sparse text.

Word embedding, which is a research focus in the field of natural language processing (NLP), is a more advanced method to transform strings into vector representation. For a given document (word sequence), each different word in this document will be represented as a low-dimensional vector of real numbers by a mapping model. This method is based on the assumption that the representations in terms of the contexts of “similar words” are similar. Word2vector achieves the idea of word embedding and contains two model architectures: Continuous bag-of-words (CBOW) and Skip-Gram [54,55]. The currently commonly used toolkits: Word2vector (<https://code.google.com/archive/p/word2vec/source/default/source>) and gensim (<https://pypi.org/project/gensim/>) are the implementation of Word2vector model. In terms of entity alignment, word embedding-based methods initially transform original strings of entities into vector representations. Whether entity pairs are matching or nonmatching will be decided by comparing their vector representations. Instead of defining and extracting feature vectors for entities manually as in previous methods, this method can automatically generate vector representations for original strings of entities, thereby improving automation and reducing manual intervention. Santos et al. achieved a word embedding-based method using a deep neural network to generate representations for toponyms. These representations were processed by a feed forward network to judge entity alignment [56].

3.1.2. Structural Similarity Metrics

A knowledge base can be represented with a tree-like or graph-like structure. With the position of entity and the relationship with other entities in the tree-like or graph-like structure, the structural similarity between entity pairs can be measured. In the general alignment systems, structural similarity is an important metric. Due to the different perspectives and purposes for constructing different GKBs, there are great differences in their structures [39]. Therefore, it is inappropriate to assign greater weight to the structural factor in geographic entity alignment. Delgado and Finat used a string-based method, WordNet-based method, and general alignment system to perform the concept alignment of GKBs, including DBpedia, LinkedGeoData, and CityGML (<http://schemas.opengis.net/citygml/>).

Their results showed that the general alignment system, which put emphasis on structural factors, presented the worst precision and recall [57], thereby proving the aforementioned conclusion.

For GKBs with a tree-like structure, structural similarity is computed based on the relationship between sets of hypernymic nodes, hyponymic nodes, and sibling nodes of entities [58]. The descendant's similarity inheritance considered the explanatory roles of ancestors of entity for the identification of the specific entity [59]. The ancestor nodes include all the nodes on the path from root node to the closest hypernymic node of entity. Then, the structural similarity between entity pairs was computed over the weighted sum of similarities between all the ancestor node pairs of entity pairs. Weights were computed over the distance from each ancestor node to the entity [33].

Similarly, the algorithms based on similarities between hyponymic node pairs (e.g., ancestor's similarity contribution (ASC)) or sibling node pairs (e.g., sibling's similarity contribution (SSC)) also used the cumulative value of similarities among the corresponding nodes [22,59]. The metric based on the Jaccard coefficient was relatively simple, and it calculated structural similarity by the ratio of the number of shared nodes to the number of all nodes linked with the entity pairs [43].

Different than the GKB with a tree-like structure, the GKB with a graph-like structure just emphasizes the simple links between entities and lacks in clear ancestor–descendant and sibling relationship between entities. Thus, the structural similarity metrics for tree-like structures are not applicable to graph-like structures. A commonly used algorithm for calculating structural similarity in graph-like structures is the similarity flooding algorithm [60]. This algorithm transforms the native structures of two GKBs to be matched into two labeled directed graphs, which are then merged into one directed graph named a pairwise connectivity graph. Each node in this connectivity graph refers to a node pair from the two graphs before merging and is called a map pair. Each map pair is with an initial similarity score. Then, the connectivity graph will be transformed into the induced propagation graph when the initial similarity of map pair propagates through the graph over a number of iterative calculations according to the similarities of its adjacent map pairs. When the fixpoint, which means the similarities of all the map pairs remain unchanged, has been reached, the algorithm will converge to the final results. The final similarity of each map pair in the propagation graph is regarded as the structural similarity between two entities. Yu et al. and Kim et al. used the similarity flooding algorithm to measure structural similarity between entities in GKBs [29,61].

The penetrating rank (P-Rank) algorithm is also a recursive algorithm for graph matching [62]. For two entities to be matched, this algorithm is based on two assumptions: (1) If they are linked with similar entities, then they are similar; (2) if they link to similar entities, then they are similar. Algorithms like co-citation and coupling are the variants of P-Rank [63,64]. These algorithms can obtain excellent results when there are dense links among entities in GKBs. Ballatore et al. applied the co-citation algorithm to compute the structural similarity of entities in OpenStreetMap [53,65].

3.1.3. Spatial Similarity Metrics

The spatial characteristic is the unique characteristic of geographic entities and makes them distinguishable from other types of entities. Spatial similarity is, thus, an important matching factor for geographic entity alignment [66], but it is not sufficient on its own [49]. It must be used in concert with other metrics emphasized in the general alignment system, such as lexical and structural similarity metrics [67].

Standard methods for computing spatial similarity are based on the spatial relationships (metric, topological, and sequential relationships) between entities. The metric relationship includes the distance, height, length, and proportion relationships, and the most commonly used is the distance relationship. A simple method is to calculate the Euclidean distance between entities based on the plane coordinates of spatial objects. Sehgal et al. and Yu et al. respectively used Euclidean distance and the inverse of it to calculate spatial similarity between two entities [26,29]. Safra et al. proposed a location-based algorithm, which computed Euclidean distance between points to find the corresponding entities for integrating geospatial datasets [68].

In order to make the spatial similarity more accurate, the distance between entities could be measured based on the spherical coordinate using a more complex haversine formula to simulate the ellipsoid surface of the earth [48]. The Hausdorff algorithm further considered the size of spatial entities [69]. The length relationship could be used to calculate the similarity between line objects [30,70]. The proportion relationship was mainly employed in measuring the proportion of coincidence in the spatial coverage of entities [51,71].

Integrating the metric relationship with other types of spatial relationships can help to calculate spatial similarity [72]. Beard and Sharma checked the topological relationship between entities before measuring the metric relationship [73]. Larson and Frontiera adopted a regression model to integrate topological and metric relationships to measure spatial similarity [74]. Li and Frederico proposed a method which comprehensively used metric, topological, and sequential relationships [75].

3.1.4. Category Similarity Metrics

Category similarity is often applied in toponym matching and location parsing. The traditional metrics for category similarity can be separated into three classes: Structure-based, content-based, and context-based metrics. Structure-based metrics are based on the category hierarchy of the classification system of toponyms. Given the category information of two toponyms, C_A and C_B , they are mapped to the category hierarchy. If C_A and C_B are on the same classification system, their closest common parent category C_P will be found; otherwise, C_A and C_B must be uniformly converted to a predefined classification system before C_P is found in this classification system. Category similarity is computed over C_P [51,71,76]. According to the different details of methods for calculating the similarity between two C_P , there are two types of methods. One of them directly represents the category similarity with the number of levels from the root node to C_P in the category hierarchy. The other type is performed based on the numbers of edges from C_A to C_P , from C_B to C_P , and from C_P to the root node in the category hierarchy.

Content-based metrics are based on the semantic information of categories of entities. According to information theory, the similarity between two categories depends on the information shared by them. Thus, category similarity can be calculated by quantitatively evaluating the common information content of two categories [77]. Kavouras et al. computed category similarity based on the semantic information, including cause, purpose, location, etc., which were extracted from the category description [78]. The matching-distance similarity measure (MDSM) algorithm combined the distinguishing features of categories with the semantic relationship between them to calculate their semantic distance, over which the computation of category similarity was performed [79].

Some improvements to the MDSM algorithm have been proposed. The MDSM algorithm is actually based on the assumption that the properties of category are equally important. However, the importance of different properties is completely different under different contexts. Thus, different weights were assigned to different properties of categories in different scenarios to improve this algorithm [80]. The synonym sets of category could also be used to improve this algorithm [81].

Context-based metrics calculate category similarity by means of background knowledge. For example, the similarity between geographic terminologies of thesaurus can be measured with the aid of a lexical database, such as WordNet [82].

3.1.5. Shape Similarity Metrics

Different than spatial similarity metrics for spatial distance between entities, shape similarity is to measure the difference between geometric shapes of entities. Shape matching is widely studied and applied in fields such as computer vision and pattern recognition [83], while this matching factor is rarely used in the alignment of geographic entities. The basic idea of shape matching is to represent the shape of entity into a normalized form, over which shape matching is performed indirectly [49]. The commonly used method is to match the nodes, which represent the shapes of entities. Safra et al. used a node matching algorithm to calculate the shape similarity of line objects [84]. Goodchild and

Hunter, as well as Fairbairn and Albakri, proposed a simpler method. The buffer zones for linear entities were initially constructed, and shape similarity between entities was indirectly computed over their buffer zones [85,86]. Du et al. also matched spatial objects leveraging the buffers of their geometries [87].

In addition to the abovementioned similarity metrics for lexicon, structure, space, category, and shape, we notice that some previous studies creatively used other features of geographic entities to complete alignment. Zhu et al. thought that if the respective spatial distribution patterns of all instances belonging to two concepts were similar, these two concepts were similar [88]. Based on this idea, the similarity between concepts was measured according to the similarity of local or global spatial distribution of instances using some metrics (e.g., Moran's I and the kernel density estimation). For the geographic concepts, which are described by description logics (DL), Janowicz et al. measured the similarity between concepts based on the degree of coincidence of their DL descriptions [89]. Kokla and Kavouras provided a formalized representation for geographic concepts based on concept lattices and integrated geographic concepts based on formal concept analysis (FCA) [90,91].

3.2. Similarity Combination

After the similarity scores on multidimensional features of geographic entities are measured, they will be combined organically, leveraging a suitable similarity combination algorithm [92]. Similarity combination includes two processes: (1) Selecting similarity scores computed over features, which can effectively improve the alignment results; (2) selecting a similarity combination model to combine the selected similarity scores.

The strategies for feature selection can be divided into three types. The first type is to select the similarities on all features directly. The second type adopts suitable selection principles or algorithms to select effective features, which can actually contribute to the alignment process. Effective feature must meet two principles: (1) This feature can effectively distinguish the matching entity pairs from the nonmatching entity pairs; (2) the number of matching entity pairs obtained using this single feature is close to the minimum number of entities in two knowledge bases to be matched [93]. The algorithms for feature selection include principal component analysis (PCA), analytic hierarchy process (AHP), expert scoring, etc. The third type, called the single factor judgement method, is to select the single feature with the greatest contribution and set the similarity threshold. When the similarity computed over this feature between two entities exceeds the threshold, they are matching pairs.

The models for combining similarities over multidimensional features include the feature vector model, geometric model, and mathematical model [94]. The feature vector model is to represent the multidimensional similarities into feature vectors, and the overall similarity is calculated over the cosine between the two vectors. In the geometric models, multidimensional similarities are represented as multidimensional coordinates, over which the aggregation of similarities is performed. The mathematical model aggregates multidimensional similarities leveraging methods of mathematical operation, which include: (1) The minimum or maximum value method, (2) average value method, (3) weighted sum method, (4) probabilistic method, (5) fuzzy aggregation method, and (6) rough sets method [95].

The most commonly used method for similarity combination is the weighted sum in the mathematical model [96]. This method assigns weight to each similarity metric and calculates the weighted sum of similarities. Thus, the multidimensional similarities are linearly combined with this method [97]. The key to this method is the weight assignment for each metrics.

The simplest method for weight assignment is the average value method, which assigns equal weight to each factor. The most direct method is to select the optimal weight assignment through multiple experiments [51,61]. Zhu et al. adopted the weight assignment method based on expert scoring. In this method, the judgement matrix was initially designed. Each element of this matrix was a relative importance score for each metric provided by the invited domain experts. The elements of the normalized eigenvector of this matrix were the weights for each metric [71]. Tran et al. proposed a

clustering-based weight estimation method. It used the K-means algorithm to divide the similarity matrix for each metric into two classes: One class with a higher mean value and the other class with a lower mean value. Then, all the similarities that belonged to the second class would be filtered out. The weight was calculated by the ratio of the number of rows that had a value in the matrix to the number of values in the first class [93]. McKenzie et al. ranked the respective results in terms of accuracy, which were computed over each single feature of name, category, position, and topic. The weight for each feature was assigned according to the ordinal ranking [98]. Li et al. proposed an entropy-based weight assignment method to determine the weights corresponding to each feature for merging multisource geo-ontologies. In this method, the information entropy for each feature was initially computed using axiomatic characterizations of information entropy. The weight for a single feature was computed by the ratio of its information entropy to the total information entropy of all single features [99].

3.3. Alignment Judgement

The standard method for alignment judgement is based on a threshold over the overall similarity, computed by combining the multidimensional similarities. If the overall similarity between two entities is beyond the predefined threshold, they are regarded as aligned correspondence. A threshold that is too high or too low will probably lead to alignment misjudgment for entity pairs, thereby making bad results. Thus, tuning the threshold is important for achieving optimal results.

In order to avoid tuning the threshold manually and to minimize human intervention, some previous studies focused on discriminating alignments automatically based on multiple similarity metrics. Supervised machine learning was commonly used for automatic geographic entity alignment [26,43,44,76]. With multiple similarity scores regarded as input, supervised machine learning models were trained by labeled entity pairs. The trained model could decide whether entity pairs to be matched are aligned based on their multiple similarity metrics. Santos et al. leveraged support vector machines (SVM), decision trees, and random forests to match toponyms, and their results showed that decision trees could achieve better results [40]. Martins also used machine learning techniques to detect duplicate gazetteers [43,44]. Li et al. and Chen et al. used an artificial neural network (ANN) model to match geographic concepts [28,100]. The supervised machine learning method actually performs a nonlinear combination for multiple similarity metrics [100]. Although it can learn an optimal scheme for similarity combination automatically, it requires large-scale training datasets, which are difficult to prepare, for a satisfactory trained model.

There are also some other algorithms which classify entity pairs as either matching or nonmatching automatically based on multiple similarity metrics. The voting-based method does not directly perform a numerical calculation on the similarity scores but generates respective matched results based on each metric. All the matched results are aggregated by voting, and the entity pairs with more affirmative votes are the final matched entity pairs [45]. Yu et al. aligned entities from OpenStreetMap and GeoNames by performing a voting-based method on multiple metrics, including spatial, lexical, and structural metrics [29]. Bock and Hettenhausen formulated entity alignment as an optimization problem, which was tackled with an iterative algorithm based on particle swarm optimization [101]. Each entity pair to be matched was represented as a particle in this algorithm, and the convergence of swarm was guided by proportional likelihood values assigned to each correspondence. Pareto ranking, which is a multiobjective evolutionary algorithm, was also used to find correspondences in heterogeneous geo-ontologies based on various similarity metrics [102].

These classical methods showed a good performance for entity alignment. However, they rely on multidimensional similarities computed over features of entity. Thus, they are actually a semi-automatic method. In order to avoid computing similarity and realize completely automatic entity alignment, Santos et al. adopted a deep neural network to align toponyms based on their original strings rather than similarity scores [56]. They initially generated representations from the sequences of bytes of original strings, leveraging gated recurrent units (GRUs), a type of recurrent neural network (RNN) architecture. The feed forward network processed these representations and

made an alignment decision. However, this method only leveraged lexicon features, thereby ignoring other features of geographic entities.

4. Evaluation of Entity Alignment

Evaluating alignment approaches of geographic entities is necessary to discover their weaknesses and strengths and choose the most suitable approach in a predefined context [103]. This step is to assess the correctness and effectiveness of entity alignment results using a suitable evaluation method and the same gold standard. Thus, the evaluation method and gold standard are two key factors in this process.

The evaluation methods can be divided into two classes: Cognitive plausibility-oriented and task-oriented methods [34]. The cognitive plausibility-oriented method is used to evaluate the simulation ability of the alignment algorithm for cognitive and behavior systems of humans [104,105]. The investigating method for cognitive plausibility is usually to compare the alignment results, which are provided by human subjects, and the alignment algorithm, for the same entity pairs to be matched. Subjects can easily decide whether entity pairs correspond based on their own cognition ability for entities. Thus, the stronger the simulating ability of an alignment algorithm for judging the process of subjects is, the higher its cognitive plausibility is, and the better the result computed by this algorithm is. Given the alignment result generated by human subjects R_S , which is usually regarded as the ground truth, and the corresponding computational result R_C , the cognitive plausibility of the alignment algorithm can be measured by calculating Spearman's correlation coefficient $\rho(R_S, R_C)$ between R_S and R_C .

The task-oriented method applies the alignment algorithm in a specific task and assesses its performance from what the degree of satisfaction is about the status of task completion with the indicators of precision, recall, and F1-measure [106], which are frequently used in information retrieval. The precision (P) is the ratio of the number of correctly aligned pairs R_T to the total number of discovered corresponding pairs N_A ; the recall (R) is the ratio of R_T to the number of desired aligned pairs N_T in the gold standard, and N_T is the true value; and F1-Measure is the harmonic mean of P and R. All the three metrics have been defined in Equations (13) and (15).

$$P = R_T / N_A \quad (11)$$

$$R = R_T / N_T \quad (12)$$

$$F1 = 2 \times P \times R / (P + R) \quad (13)$$

The well-defined and widely recognized gold standards, which are also known as benchmark datasets, are necessary to provide a common objective basis to make the results of different algorithms comparable [103]. In the Oxford Dictionary of English, the term 'gold standard' is explained as "A thing of superior quality which serves as a point of reference against which other things of its type may be compared" (https://en.oxforddictionaries.com/definition/gold_standard). In computer science, gold standard refers to human-generated datasets which consist of testing data and correct output results (i.e., ground truth) [107]. For a specific task, it can capture the behavior and cognitive patterns of humans, quantify the relevance between machine-generated and human-generated results, and thus be used as a benchmark to evaluate the performance of the computational method. An actually valid gold standard should be open, accessible, persistent, and unbiased for providing reliable and fair evaluation results, and it needs to meet certain requirements, such as coverage, quality, and precision [103,107].

In computer science, dozens of benchmark datasets have been created [108], such as XBenchmark [109] and STBenchmark [110]. In addition, some projects on assessing entity alignment have been carried out. The Semantic Evaluation at Large Scale (SEALS) project (<http://www.seals-project.eu/>) provides a software infrastructure for evaluating semantic web tools, including tools for entity alignment. The Ontology Alignment Evaluation Initiative (OAEI) (<http://oei>.

ontologymatching.org/) provides benchmark datasets annually for participants to evaluate different systems and algorithms for ontology alignment [111]. Each benchmark dataset is composed of reference ontology, target ontology, and reference alignment. The reference ontology contains 33 concepts, 24 object properties, 40 data properties, and 50 instances. The target ontology is composed of many kinds of alterations of reference ontology, including lexical changes, synonym substitution, compression of annotation, and flattening or expanding the hierarchical structure. Each type of alterations aims at a specific type of heterogeneity issues. The reference alignment is the desired result of entity alignment. All the participating alignment systems are employed to match the reference ontology and target ontology. The performance evaluation for these systems is implemented by comparing their results with the reference alignment. The OAEI can relatively comprehensively evaluate the performance, coverage, stability, and reusability of alignment systems.

Due to the spatial feature of geographic entities, the entity alignment algorithms for GKBs are different from the ones for general knowledge bases. Thus, the benchmarks for evaluating the alignment systems of computer science may not be applicable to GIScience. Meanwhile, there are very few benchmarks exclusively targeting geographic entity alignment. For example, PABench (POI Alignment Benchmark) contains 1580 entities and several test cases covering different situations of heterogeneities [112].

Some benchmark datasets, which were designed for evaluating the similarity metrics for geographic terms, can also be used to evaluate entity alignment algorithms. GeReSiD (<https://github.com/ucd-spatial/Datasets>) (geo relatedness and similarity dataset) contains 97 geographic terms from OpenStreetMap and 50 term pairs, for which the similarity ranking was provided by 203 human subjects [107]. The MDSM dataset created by Rodriguez and Egenhofer for assessing their MDSM algorithm is composed of 33 geographic terms, which cover natural and man-made features, and 108 term pairs. Seventy-two human subjects were asked to rank these pairs by their similarity scores [80].

5. Challenges and Future Research

As shown in the previous sections, massive outstanding achievements for the entity alignment of GKBs, in terms of similarity metrics, similarity combination, alignment judgement, and result evaluation, have been made. Meanwhile, this semantic technique has been widely used in the integration and conflation of geographic data or knowledge [113,114], toponym resolution [26,43,44,66], correlation and discovery of geographic information [58,71,115–118], web service chain composition [119,120], and personalized recommendations [121,122]. However, there are still some challenges, which need to be addressed in the future.

5.1. Quality Assessment of GKBs

The quality of GKBs has a major impact on the result of entity alignment. There are significant differences in the knowledge quality of multisource GKBs. Especially for GKBs generated from volunteered geographic information (VGI), this problem is more prominent [123]. Therefore, quality assessment for geographic knowledge in GKBs needs to be investigated. There have been some studies focused on quality assessment of geographic knowledge from GKBs [123–126]. The quality measures to describe the quality of GKBs include positional accuracy, thematic accuracy, topological accuracy, completeness, consistency, temporal accuracy, and semantic accuracy. Many methods have been developed to assess the positional accuracy, thematic accuracy, and topological consistency by comparing with a reference dataset. However, few methods exist to assess the rest of the quality measures. More methods to handle these quality measures should be developed in the future.

5.2. Feature Selection and Algorithms Optimization

The various heterogeneous types of entities in GKBs have been explained in the Section 2.1, and each of them focuses on different aspects of entities. Meanwhile, alignment tasks usually have different

requirements and constraints in terms of accuracy, completeness, and efficiency, thereby making feature selection a multiple criteria decision-making problem. Thus, it is very critical and challenging to determine the best composition of features for a specific task. A similar problem also exists in metrics selection, because there are many different metrics for each type of heterogeneity. The methods of AHP, PCA, machine learning, and ad hoc rules have been used for feature or metrics selection [127–130]. In addition, features that are not still employed should be considered. The data-type and range of properties should be further considered to align properties, and the instance alignment process should take the property value into consideration. Temporality is an intrinsic feature of the geographic entities, and spatial coverage and property values of geographic entity may change over time [67]. However, previous studies rarely take the temporal feature of entities into consideration [131], thereby sometimes leading to incorrect results of entity alignment.

Taking both the accuracy and efficiency of algorithms into account, how to optimize algorithms of similarity metrics, similarity combination, alignment judgement, and result evaluation is another issue [132]. The methods for finding the most suitable parameters of similarity measurement have been discussed [133,134]. Developing new algorithms with the benefit of new techniques in computer science can also facilitate the progress of this field. The methods of pareto ranking and particle swarm optimization have been employed for entity alignment [101,102]. The word embedding method in the field of NLP should be continuously studied and used for entity alignment. In addition, the majority of existing algorithms only align one type of entities, including concepts, properties, and instances, so developing holistic algorithms, which can complete a one-shot alignment for all types of entities, is necessary.

5.3. Alignment Techniques Integrated with Background Knowledge

GKBs are usually developed in specific contexts, which include relevant background knowledge, but the knowledge is often not directly represented in the developed GKBs. Moreover, some GKBs lack enough specifications. These problems may cause ambiguity in entities, thereby leading to incorrect alignments. Thus, integrating background knowledge into alignment is particularly necessary to achieve better results. The background knowledge is represented as multiple forms, such as domain corpora, domain ontologies or upper-level ontologies, existing aligned entity pairs, and web pages. There are some previous studies which performed alignment by means of lexicon, such as WordNet [34–37], but other forms of background knowledge have rarely been used. Thus, techniques which combine with multisource background knowledge need to be investigated.

5.4. Unified Infrastructure for Entity Alignment of GKBs

Most of the previous studies are experimental or in the prototype stage and cannot be applied in realistic scenarios. Thus, establishing a stable and unified infrastructure for the entity alignment of GKBs, which integrates a variety of alignment algorithms and can complete, store, share, and reuse alignments, is an important research mission in the future. This infrastructure will greatly facilitate the practical application of this field. In computer science, some systems, which provide many alignment methods and libraries of alignments, and can be used at design time and run time, have been designed [135–137]. This is of important reference value for designing similar infrastructures in GIScience.

5.5. Entity Alignment of Large-Scale GKBs

Under the era of geographic big data, the scales of GKBs have become increasingly large, and their structures are more complex, thereby bringing enormous growth in the computational complexity for alignment algorithms. However, existing methods rarely consider resource consumption, thus posing a challenge to the efficiency of entity alignment. There are two perspectives to tackle this problem.

From the perspective of data sources, the computation load of entity alignment is directly proportional to the size of GKBs and the number of entity pairs to be matched. Thus, a straightforward

solution is to partition GKBs into proper fragments, namely modularization of GKBs, and to reduce the number of entity pairs to be matched with some preprocessing steps [138,139]. Alignment algorithms are performed over split, smaller GKBs and preprocessed less entity pairs.

From the perspective of methods, the basic idea is to make alignment algorithms scalable and parallelized. In computer science, some previous studies parallelized original algorithms based on the parallel programming environment, such as the message passing interface (MPI) [18], or deployed alignment algorithms on distributed computing platforms, such as Hadoop, Spark et al. [140–142]. The parallel and distributed processing of alignment tasks for GKBs needs to be covered in the future.

5.6. Deep Learning-Based Entity Alignment of GKBs

Traditional methods abide by the standard workflow for aligning geographic entities. Therefore, the result of entity alignment is mainly determined by the selection of features and similarity metrics, and thus with varying influence of human intervention. Deep learning methods can complete end-to-end learning and learn intrinsic features of original data automatically, thereby reducing the influence caused by human intervention at a maximum level. Thus, in order to avoid the subjective factors of humans, we should try to develop novel methods based on deep learning to achieve automatic alignment. The key to this method is that it requires large-scale training datasets, which are difficult to prepare. This method is well worth investigating, because it can break through the standard workflow.

5.7. Benchmark Datasets for Entity Alignment of GKBs

Standardized large-scale benchmarks are important prerequisites for finding the cons and pros of entity alignment algorithms. Each benchmark usually consists of one initial knowledge base, one altered knowledge base, which is an alteration over the initial one, and a reference alignment, which is used to compare with the returned alignment [143]. The OAEI has provided many artificial benchmarks, but their scale and comprehensiveness remain inadequate for dealing with a variety of existing alignment matchers, and they lack variability. Moreover, artificial benchmarks are becoming infeasible for evaluating large-scale alignment tasks. Thus, some approaches focused on a semi-automatic generation of benchmark datasets have been proposed [144,145] (e.g., Swing [145], Spimbench [146], and Lance [147]). These generators take the seed knowledge base and parameters which describe the modification types to be applied as input and generate the modified knowledge base and the corresponding reference alignment [148].

The benchmarks provided by OAEI, however, are obviously inappropriate for evaluating the entity alignment of GKBs due to the spatial feature of geographic entities. Existing generators do not also cover the alterations over spatial feature. Thus, the construction of benchmarks for geographic entity alignment is still performed manually [80,107]. However, the scales of developed artificial datasets are too small and are not suitable for a large number of correspondences. Moreover, these datasets can only be used to assess concept-level alignment, so there is lack of benchmarks for evaluating instance-level alignment. Thus, benchmarks which cover multiple alterations over geographic entities and are applicable for evaluating all types of geographic entities alignment and benchmark generators which can produce large-scale test sets automatically need to be developed.

5.8. Applications of Entity Alignment of GKBs

The practical applications of this field remain slightly limited. Broadening the application fields of acquired alignments in real-life projects needs to be further improved. Entity alignment is actually an important prerequisite for many applications. With the development of deep learning, completely automatic entity alignment may be achieved, which can support the automatic construction of large-scale geographic knowledge graphs. The strong and deep knowledge reasoning ability of huge GKBs will facilitate the progress of some fields, including mining and analysis of geographic big

data, geographic knowledge services, and applying AI in the field of GIScience. Applying the entity alignment of GKBs in the practical applications of these fields needs to be further developed.

6. Conclusions

In this article, we provided a systematic and comprehensive analysis for research progress on the entity alignment of GKBs. We introduced the basic definitions and the multiple heterogeneities in entities, including differences in lexicon, structure, spatial position, category, shape, data-type of property, range of property, and property value. A general framework, which involved the basic ideas and standard workflow of this field, was also presented in this paper.

We provided a survey on the alignment algorithms of similarity metrics, similarity combination, alignment judgement, and result evaluation. For similarity metrics, we organized the insights from previous studies systematically from five aspects of lexical, structural, spatial, category, and shape similarity metrics. In terms of similarity combination, we introduced three models, a feature vector model, geometric model, and mathematical model. We also introduced some algorithms of alignment judgement, which can help to avoid tuning the threshold manually, including supervised machine learning, a voting-based model, etc. The insights for results evaluation were organized into two parts: Evaluation methods, including cognitive plausibility-oriented and task-oriented methods, and benchmark datasets. This review provides readers, especially new researchers in this field, with a general idea, and help them to understand the basics of this field.

On the basis of a systematic review, we presented key challenges facing this field, including the quality assessment of GKBs, feature selection and algorithm optimization, alignment techniques integrated with background knowledge, unified infrastructure, entity alignment of large-scale GKBs, alignment techniques based on deep learning, benchmark datasets, and application promotion. These insights will be helpful for promoting the progress and orienting future research for this field.

Author Contributions: All authors gave substantial contributions to this work. Conceptualization was conducted by all listed authors. Formal analysis and investigation were conducted by K.S. Writing—original draft preparation was conducted by K.S.; writing—review and editing were conducted by all authors. Supervision, project administration and funding acquisition were conducted by Y.Z. and J.S.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41631177, grant number 41771430; National Special Program on Basic Works for Science and Technology of China, grant number 2013FY110900.

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

References

1. Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. Dbpedia—A crystallization point for the Web of Data. *Web Semant. Sci. Serv. Agents World Wide Web* **2009**, *7*, 154–165. [CrossRef]
2. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago—A Large Ontology from Wikipedia and WordNet. *Web Semant. Sci. Serv. Agents World Wide Web* **2008**, *6*, 203–217. [CrossRef]
3. Wikidata. Available online: https://www.wikidata.org/wiki/Wikidata:Main_Page (accessed on 20 March 2018).
4. Bollacker, K.; Cook, R.; Tufts, P. Freebase: A Shared Database of Structured General Human Knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–26 July 2007; pp. 1962–1963.
5. Berners, T. Publishing on the semantic web. *Nature* **2001**, *410*, 1023–1024. [CrossRef] [PubMed]
6. DesignIssues: LinkedData. Available online: <https://www.w3.org/DesignIssues/LinkedData.html> (accessed on 20 March 2018).
7. Li, L.; Liu, Y.; Zhu, H.; Ying, S.; Luo, Q.; Luo, H.; Xi, K.; Xia, H.; Shen, H. A bibliometric and visual analysis of global geo-ontology research. *Comput. Geosci.* **2017**, *99*, 1–8. [CrossRef]
8. Liu, Y.; Li, L.; Shen, H.; Yang, H.; Luo, F. A Co-Citation and Cluster Analysis of Scientometrics of Geographic Information Ontology. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 120. [CrossRef]

9. Couclelis, H. Ontologies of geographic information. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1785–1809. [[CrossRef](#)]
10. Bittner, T.; Donnelly, M.; Smith, B. A spatio-temporal ontology for geographic information integration. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 765–798. [[CrossRef](#)]
11. Zong, N.; Nam, S.; Eom, J.H.; Ahn, J.; Joe, H.; Kim, H.G. Aligning ontologies with subsumption and equivalence relations in Linked Data. *Knowl.-Based Syst.* **2014**, *76*, 30–41. [[CrossRef](#)]
12. Euzenat, J.; Shvaiko, P. *Ontology Matching*, 1st ed.; Springer: Heidelberg, Germany, 2007.
13. Bhattacharya, I.; Getoor, L. Entity Resolution in Graphs. In *Mining Graph Data*, 1st ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005; pp. 311–344.
14. Whang, S.E.; Menestrina, D.; Koutrika, G.; Theobald, M.; Garcia-Molina, H. Entity resolution with iterative blocking. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, RI, USA, 29 June–2 July 2009; pp. 219–232.
15. Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* **2006**, *19*, 1–16. [[CrossRef](#)]
16. Li, C.; Jin, L.; Mehrotra, S. Supporting Efficient Record Linkage for Large Data Sets Using Mapping Techniques. *World Wide Web* **2006**, *9*, 557–584. [[CrossRef](#)]
17. Otero-Cerdeira, L.; Rodríguez-Martínez, F.J.; Gómez-Rodríguez, A. Ontology matching: A literature review. *Expert Syst. Appl.* **2015**, *42*, 949–971. [[CrossRef](#)]
18. Shvaiko, P.; Euzenat, J. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.* **2012**, *25*, 158–176. [[CrossRef](#)]
19. David, J. AROMA results for OAEI 2011. In Proceedings of the 6th International Conference on Ontology Matching, Bonn, Germany, 24 October 2011; pp. 122–125.
20. David, J.; Guillet, F.; Briand, H. Association Rule Ontology Matching Approach. *Int. J. Semant. Web Inf.* **2007**, *3*, 27–49. [[CrossRef](#)]
21. Cruz, I.F.; Antonelli, F.P.; Stroe, C. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *Proc. VLDB Endow.* **2009**, *2*, 1586–1589. [[CrossRef](#)]
22. Cruz, I.F.; Sunna, W.; Chaudhry, A. Semi-automatic Ontology Alignment for Geospatial Data Integration. In Proceedings of the Third International Conference on Geographic Information Science, Adelphi, MD, USA, 20–23 October 2004; pp. 51–66.
23. Hartung, M.; Kolb, L.; Groß, A.; Rahm, E. Optimizing Similarity Computations for Ontology Matching—Experiences from GOMMA. In Proceedings of the International Conference on Data Integration in the Life Sciences, Montreal, QC, Canada, 11–12 July 2013; pp. 81–89.
24. Kalfoglou, Y.; Schorlemmer, M. Ontology mapping: The state of the art. *Knowl. Eng. Rev.* **2005**, *18*, 1–31. [[CrossRef](#)]
25. Hess, G.N.; Iochpe, C.; Castano, S. An Algorithm and Implementation for GeoOntologies Integration. In Proceedings of the VIII Brazilian Symposium on Geoinformatics, Campos do Jordão, São Paulo, Brazil, 19–22 November 2006; pp. 109–120.
26. Sehgal, V.; Viechnicki, P.D.; Viechnicki, P.D. Entity resolution in geospatial data integration. In Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, Arlington, VA, USA, 10–11 November 2006; pp. 83–90.
27. Zhao, T. The framework of a geospatial semantic web-based spatial decision support system for Digital Earth. *Int. J. Digit. Earth* **2010**, *3*, 111–134. [[CrossRef](#)]
28. Li, W.; Raskin, R.; Goodchild, M. Semantic similarity measurement based on knowledge mining: An artificial neural net approach. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 1415–1435. [[CrossRef](#)]
29. Yu, L.; Qiu, P.; Liu, X.; Lu, F.; Wan, B. A holistic approach to aligning geospatial data with multidimensional similarity measuring. *Int. J. Digit. Earth* **2017**, *11*, 1–18. [[CrossRef](#)]
30. Volz, S. Data-driven matching of geospatial schemas. In Proceedings of the International Conference on Spatial Information Theory, Ellicottville, NY, USA, 14–18 September 2005; pp. 115–132.
31. Shvaiko, P.; Euzenat, J. *A Survey of Schema-based Matching Approaches*; Springer: Berlin, Germany, 2005; pp. 146–171.
32. Lin, F.; Sandkuhl, K. A Survey of Exploiting WordNet in Ontology Matching. In Proceedings of the Artificial Intelligence in Theory and Practice II, IFIP World Computer Congress, Milano, Italy, 7–10 September 2008; pp. 341–350.

33. Sunna, W.; Cruz, I.F. Structure-Based Methods to Enhance Geospatial Ontology Alignment. In Proceedings of the International Conference on Geospatial Semantics, Mexico City, Mexico, 29–30 November 2007; pp. 82–97.
34. Ballatore, A.; Wilson, D.C.; Bertolotto, M. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2099–2118. [[CrossRef](#)]
35. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Grounding Linked Open Data in WordNet: The Case of the OSM Semantic Network. In Proceedings of the International Symposium on Web and Wireless Geographical Information Systems, Banff, AB, Canada, 4–5 April 2013; pp. 1–15.
36. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Linking geographic vocabularies through WordNet. *Ann. GIS* **2014**, *20*, 73–84. [[CrossRef](#)]
37. Giunchiglia, F.; Maltese, V.; Farazi, F.; Dutta, B. *GeoWordNet: A Resource for Geo-spatial Applications*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 121–136.
38. Hu, Y. Geospatial Semantics. In *Comprehensive Geographic Information Systems*; Elsevier: Oxford, UK, 2017; pp. 80–94.
39. Zheng, J.G.; Fu, L.; Ma, X.; Fox, P. SEM+: Tool for discovering concept mapping in Earth science related domain. *Earth Sci. Inform.* **2015**, *8*, 95–102. [[CrossRef](#)]
40. Santos, R.; Murrieta-Flores, P.; Martins, B. Learning to combine multiple string similarity metrics for effective toponym matching. *Int. J. Digit. Earth* **2018**, *11*, 913–938. [[CrossRef](#)]
41. Recchia, G.; Louwerse, M.M. A Comparison of String Similarity Measures for Toponym Matching. In Proceedings of the ACM Sigspatial Comp'13, Orlando, FL, USA, 5–8 November 2013; pp. 54–61.
42. Levenshtein, V.I. Binary codes capable of correcting spurious insertions and deletions of ones. *Probl. Inf. Transm.* **1965**, *1*, 707–710.
43. Martins, B. A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records. In Proceedings of the International Conference on Geospatial Semantics, Brest, France, 12–13 May 2011; pp. 34–51.
44. Martins, B.; Galhardas, H.; Goncalves, N. Using Random Forest classifiers to detect duplicate gazetteer records. In Proceedings of the 7th Iberian Conference on Information Systems and Technologies (CISTI 2012), Madrid, Spain, 20–23 June 2012; pp. 1–4.
45. Wang, Z.; Li, J.; Zhao, Y.; Setchi, R.; Tang, J. A unified approach to matching semantic data on the Web. *Knowl.-Based Syst.* **2013**, *39*, 173–184. [[CrossRef](#)]
46. Hastings, J.; Hill, L. Treatment of duplicates in the alexandria digital library gazetteer. In Proceedings of the GeoScience, Boulder, CO, USA, 25–28 September 2002.
47. Auer, S.; Lehmann, J.; Hellmann, S. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In Proceedings of the International Semantic Web Conference, Chantilly, VA, USA, 25–29 October 2009; pp. 731–746.
48. Stadler, C.; Lehmann, J.; Ffner, K.; Auer, S. LinkedGeoData: A core for a web of spatial open data. *Semant. Web* **2012**, *3*, 333–354. [[CrossRef](#)]
49. 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](#)]
50. Aoe, J.I. *Computer Algorithms: String Pattern Matching Strategies*; John Wiley & Sons: Hoboken, NJ, USA, 1994.
51. Hastings, J. Automated conflation of digital gazetteer data. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1109–1127. [[CrossRef](#)]
52. Salton, G.; Yang, C.S. The Specification of Term Values In Automatic Indexing. *J. Doc.* **1973**, *29*, 351–372. [[CrossRef](#)]
53. Ballatore, A.; Bertolotto, M.; Wilson, D. A Structural-Lexical Measure of Semantic Similarity for Geo-Knowledge Graphs. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 471–492. [[CrossRef](#)]
54. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013; arXiv:1301.3781.
55. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
56. Santos, R.; Murrieta-Flores, P.; Calado, P.; Martins, B. Toponym matching through deep neural networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 324–348. [[CrossRef](#)]
57. Delgado, F.; Finat, J. An evaluation of ontology matching techniques on geospatial ontologies. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2279–2301. [[CrossRef](#)]

58. Reza, K.; Ali, A.; Majid, H. A mixed approach for automated spatial ontology alignment. *J. Spat. Sci.* **2010**, *55*, 237–255. [[CrossRef](#)]
59. Cruz, I.F.; Sunna, W. Structural Alignment Methods with Applications to Geospatial Ontologies. *Trans. GIS.* **2008**, *12*, 683–711. [[CrossRef](#)]
60. Melnik, S.; Garcia-Molina, H.; Rahm, E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In Proceedings of the International Conference on Data Engineering, San Jose, CA, USA, 26 February–1 March 2002; pp. 117–128.
61. Kim, J.; Vasardani, M.; Winter, S. Similarity matching for integrating spatial information extracted from place descriptions. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 56–80. [[CrossRef](#)]
62. Zhao, P.; Han, J.; Sun, Y. P-Rank: A comprehensive structural similarity measure over information networks. In Proceedings of the ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 553–562.
63. Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Assoc. Inf. Sci. Technol.* **1973**, *24*, 265–269. [[CrossRef](#)]
64. Kessler, M.M. Bibliographic coupling between scientific papers. *J. Assoc. Inf. Sci. Technol.* **1963**, *14*, 10–25. [[CrossRef](#)]
65. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81. [[CrossRef](#)]
66. Kang, H.; Sehgal, V.; Getoor, L. GeoDDupe: A Novel Interface for Interactive Entity Resolution in Geospatial Data. In Proceedings of the International Conference Information Visualization, Zurich, Switzerland, 4–6 July 2007; pp. 489–496.
67. Hess, G.N.; Iochpe, C.; Ferrara, A.; Castano, S. Towards Effective Geographic Ontology Matching. In Proceedings of the GeoSpatial Semantics, Second International Conference, Mexico City, Mexico, 29–30 November 2007; pp. 51–65.
68. Safra, E.; Kanza, Y.; Sagiv, Y.; Beerli, 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](#)]
69. Janée, G.; Frew, J. Spatial search, ranking, and interoperability. In Proceedings of the 27th Annual International ACM SIGIR Conference, Sheffield, UK, 29 July 2004.
70. Walter, V.; Fritsch, D. Matching spatial data sets: A statistical approach. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 445–473. [[CrossRef](#)]
71. Zhu, Y.; Zhu, A.X.; Song, J.; Yang, J.; Feng, M.; Sun, K.; Zhang, J.; Hou, Z.; Zhao, H. Multidimensional and quantitative interlinking approach for Linked Geospatial Data. *Int. J. Digit. Earth* **2017**, *10*, 923–943. [[CrossRef](#)]
72. Bruns, H.T.; Egenhofer, M.J. Similarity of Spatial Scenes. In Proceedings of the Symposium on Spatial Data Handling, Delft, The Netherlands, 12–16 August 1996; pp. 31–42.
73. Beard, K.; Sharma, V. Multidimensional ranking for data in digital spatial libraries. *Int. J. Digit. Libr.* **1997**, *1*, 153–160. [[CrossRef](#)]
74. Larson, R.R.; Frontiera, P. Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In Proceedings of the Research and Advanced Technology for Digital Libraries, European Conference, Bath, UK, 12–17 September 2004; pp. 45–56.
75. Li, B.; Frederico, F. TDD: A Comprehensive Model for Qualitative Spatial Similarity Assessment. *Spat. Cogn. Comput.* **2006**, *6*, 31–62. [[CrossRef](#)]
76. Zheng, Y.; Fen, X.; Xie, X.; Peng, S.; Fu, J. Detecting nearly duplicated records in location datasets. In Proceedings of the ACM Sigspatial International Symposium on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November 2010; pp. 137–143.
77. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; pp. 448–453.
78. Kavouras, M.; Kokla, M.; Tomai, E. Comparing categories among geographic ontologies. *Comput. Geosci.* **2005**, *31*, 145–154. [[CrossRef](#)]
79. Rodriguez, M.A.; Egenhofer, M.J.; Rugg, R.D. Assessing Semantic Similarities among Geospatial Feature Class Definitions. In *Interoperating Geographic Information Systems*; Springer: Berlin, Germany, 1999; pp. 189–202.

80. Rodriguez, M.A.; Egenhofer, M.J. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 229–256. [[CrossRef](#)]
81. Rodriguez, M.A.; Egenhofer, M.J. Determining semantic similarity among entity classes from different ontologies. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 442–456. [[CrossRef](#)]
82. Chen, Z.; Song, J.; Yang, Y. An Approach to Measuring Semantic Relatedness of Geographic Terminologies Using a Thesaurus and Lexical Database Sources. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 98. [[CrossRef](#)]
83. Veltkamp, R.C.; Hagedoorn, M. State of the Art in Shape Matching. In *Principles of Visual Information Retrieval*; Springer: London, UK, 2001; pp. 87–119.
84. Safra, E.; Kanza, Y.; Sagiv, Y.; Doytsher, Y. Ad hoc matching of vectorial road networks. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 114–153. [[CrossRef](#)]
85. Goodchild, M.F.; Hunter, G.J. A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 299–306. [[CrossRef](#)]
86. Fairbairn, D.; Albakri, M. Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 349–370. [[CrossRef](#)]
87. Du, H.; Alechina, N.; Jackson, M.; Hart, G. A Method for Matching Crowd-sourced and Authoritative Geospatial Data. *Trans. GIS* **2017**, *21*. [[CrossRef](#)]
88. Zhu, R.; Hu, Y.; Janowicz, K.; Mckenzie, G. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Trans. GIS* **2016**, *20*, 333–355. [[CrossRef](#)]
89. Janowicz, K.; Schwarz, M.; Wilkes, M.; Panov, I.; Espeter, M. Algorithm, implementation and application of the SIM-DL similarity server. In Proceedings of the International Conference on Geospatial Semantics, Mexico City, Mexico, 29–30 November 2007; pp. 128–145.
90. Kokla, M.; Kavouras, M. Fusion of top-level and geographical domain ontologies based on context formation and complementarity. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 679–687. [[CrossRef](#)]
91. Kavouras, M.; Kokla, M. A method for the formalization and integration of geographical categorizations. *Int. J. Geogr. Inf. Sci.* **2002**, *16*, 439–453. [[CrossRef](#)]
92. Peukert, E.; Maßmann, S.; König, K. Comparing Similarity Combination Methods for Schema Matching. In Proceedings of the 40th Annual Conference of the German Computer Society (GI-Jahrestagung), Leipzig, Germany, 1 October 2010; pp. 692–701.
93. Tran, Q.V.; Ichise, R.; Ho, B.Q. Cluster-based similarity aggregation for ontology matching. In Proceedings of the International Conference on Ontology Matching, Bonn, Germany, 24 October 2011; pp. 142–147.
94. Schwering, A. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Trans. GIS* **2008**, *12*, 5–29. [[CrossRef](#)]
95. Jan, S.; Shah, I.; Khan, I.; Khan, F.; Usman, M. Similarity Measures and their Aggregation in Ontology Matching. *Int. J. Comput. Sci. Telecommun.* **2012**, *3*, 52–57.
96. Do, H.H.; Rahm, E. COMA: A system for flexible combination of schema matching approaches. In Proceedings of the VLDB Endowment, Hong Kong, China, 20–23 August 2002; pp. 610–621.
97. Hu, Y.H.; Ge, L. Learning Ranking Functions for Geographic Information Retrieval Using Genetic Programming. *J. Res. Pract. Inf. Technol.* **2009**, *41*, 39–52. [[CrossRef](#)]
98. 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](#)]
99. Li, J.; He, Z.; Zhu, Q. An entropy-based weighted concept lattice for merging multi-source geo-ontologies. *Entropy* **2013**, *15*, 2303–2318. [[CrossRef](#)]
100. Chen, Z.; Song, J.; Yang, Y. Similarity Measurement of Metadata of Geospatial Data: An Artificial Neural Network Approach. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 90. [[CrossRef](#)]
101. Bock, J.; Hettenhausen, J. Discrete particle swarm optimisation for ontology alignment. *Inf. Sci.* **2012**, *192*, 152–173. [[CrossRef](#)]
102. Bharambe, U.; Durbha, S.S. Adaptive Pareto-based approach for geo-ontology matching. *Comput. Geosci.* **2018**, *119*, 92–108. [[CrossRef](#)]
103. Daskalaki, E.; Flouris, G.; Fundulaki, I.; Saveta, T. Instance matching benchmarks in the era of Linked Data. *J. Web Semant.* **2016**, *39*, 1–14. [[CrossRef](#)]
104. Keßler, C. What is the difference? A cognitive dissimilarity measure for information retrieval result sets. *Knowl. Inf. Syst.* **2012**, *30*, 319–340. [[CrossRef](#)]

105. Janowicz, K.; Keßler, C.; Panov, I.; Wilkes, M.; Espeter, M.; Schwarz, M. A Study on the Cognitive Plausibility of SIM-DL Similarity Rankings for Geographic Feature Types. In Proceedings of the Agile, Washington, DC, USA, 4–8 August 2008; pp. 115–134.
106. Goutte, C.; Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359.
107. Ballatore, A.; Bertolotto, M.; Wilson, D.C. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica* **2014**, *18*, 747–767. [[CrossRef](#)]
108. Fundulaki, I.; Ngonga-Ngomo, A.C. Instance Matching Benchmark for Spatial Data: A Challenge Proposal to OAEI. In Proceedings of the International Semantic Web Conference, Kobe, Japan, 17–21 October 2016; pp. 233–234.
109. Duchateau, F.; Bellahsene, Z. Designing a Benchmark for the Assessment of XML Schema Matching Tools. *Open J. Datab.* **2014**, *1*, 3–25. [[CrossRef](#)]
110. Alexe, B.; Tan, W.C.; Velegrakis, Y. STBenchmark: Towards a benchmark for mapping systems. *VLDB Endow.* **2008**, *1*, 230–244. [[CrossRef](#)]
111. Euzenat, J.; Ferrara, A.; Hollink, L.; Isaac, A.; Joslyn, C.; Malaisé, V.; Meilicke, C.; Nikolov, A.; Pane, J.; Sabou, M.; et al. Results of the Ontology Alignment Evaluation Initiative 2009. In Proceedings of the 4th ISWC Workshop on Ontology Matching, Chantilly, VA, USA, 25 October 2009; pp. 73–95.
112. Berjawi, B.; Duchateau, F.; Favetta, F.; Miquel, M.; Laurini, R. PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching. In Proceedings of the Seventh International Conference on Advanced Geographic Information Systems, Applications, and Services, Lisbon, Portugal, 22–27 February 2015; pp. 7–16.
113. Janowicz, K.; Pehle, T.; Pehle, T.; Hart, G. Geospatial semantics and linked spatiotemporal data—Past, present, and future. *Semant. Web* **2012**, *3*, 321–332. [[CrossRef](#)]
114. Stock, K.; Cialone, C. An Approach to the Management of Multiple Aligned Multilingual Ontologies for a Geospatial Earth Observation System. In Proceedings of the International Conference on GeoSpatial Semantics, Brest, France, 12–13 May 2011; pp. 52–69.
115. Cruz, I.F.; Sunna, W.; Makar, N.; Bathala, S. A visual tool for ontology alignment to enable geospatial interoperability. *J. Visual Lang. Comput.* **2007**, *18*, 230–254. [[CrossRef](#)]
116. Mata, F. Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching. In Proceedings of the Second International Conference on GeoSpatial Semantics, Mexico City, Mexico, 29–30 November 2007; pp. 98–113.
117. Vaccari, L.; Shvaiko, P.; Pane, J.; Besana, P.; Marchese, M. An evaluation of ontology matching in geo-service applications. *GeoInformatica* **2012**, *16*, 31–66. [[CrossRef](#)]
118. Duckham, M.; Worboys, M. Automated Geographical Information Fusion and Ontology Alignment. In *Spatial Data on the Web*; Springer: Berlin, Germany, 2007; pp. 109–132.
119. Lutz, M. Ontology-Based Descriptions for Semantic Discovery and Composition of Geoprocessing Services. *GeoInformatica* **2007**, *11*, 1–36. [[CrossRef](#)]
120. Vaccari, L.; Shvaiko, P.; Marchese, M. A geo-service semantic integration in spatial data infrastructures. *Int. J. Spat. Data Infrastruct. Res.* **2009**, *4*, 24–51.
121. Ding, R.; Chen, Z. RecNet: A deep neural network for personalized POI recommendation in location-based social networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1–18. [[CrossRef](#)]
122. Zhu, Y.; Zhu, A.X.; Feng, M.; Song, J.; Zhao, H.; Yang, J.; Zhang, Q.; Sun, K.; Zhang, J.; Yao, L. A similarity-based automatic data recommendation approach for geographic models. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1403–1424. [[CrossRef](#)]
123. Senaratne, H.; Mobasher, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [[CrossRef](#)]
124. Moreri, K.K.; Fairbairn, D.; James, P. Volunteered geographic information quality assessment using trust and reputation modelling in land administration systems in developing countries. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1–29. [[CrossRef](#)]
125. Barron, C.; Neis, P.; Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Trans. GIS* **2015**, *18*, 877–895. [[CrossRef](#)]

126. Bordogna, G.; Carrara, P.; Criscuolo, L.; Pepe, M.; Rampini, A. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Inf. Sci.* **2014**, *258*, 312–327. [[CrossRef](#)]
127. Marie, A.; Gal, A. Boosting Schema Matchers. In Proceedings of the OTM Confederated International Conferences On the Move to Meaningful Internet Systems, Monterrey, Mexico, 9–14 November 2008; pp. 283–300.
128. Mochol, M.; Jentzsch, A.; Euzenat, J. Applying an Analytic Method for Matching Approach Selection. In Proceedings of the International Workshop on Ontology Matching, Athens, GA, USA, 5 November 2006; pp. 37–48.
129. Huza, M.; Harzallah, M.; Trichet, F. OntoMas: A Tutoring System dedicated to Ontology Matching. In *Enterprise Interoperability II*; Springer: London, UK, 2007; pp. 377–388.
130. Mochol, M.; Jentzsch, A. Towards a Rule-Based Matcher Selection. In Proceedings of the International Conference on Knowledge Engineering: Practice and Patterns, Acitrezza, Italy, 29 September–2 October 2008; pp. 109–119.
131. Parent, C.; Spaccapietra, S.; Zimányi, E. Conceptual Modeling for Traditional and Spatio-Temporal Applications. In *The MADS Approach*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006; pp. 188–189.
132. Shvaiko, P.; Euzenat, J. Ten Challenges for Ontology Matching. In Proceedings of the OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Monterrey, Mexico, 9–14 November 2008; pp. 1164–1182.
133. Lee, Y.; Sayyadian, M.; Doan, A.; Rosenthal, A.S.; Sayyadian, M.; Doan, A.; Rosenthal, A.S. eTuner: Tuning schema matching software using synthetic scenarios. *VLDB J.* **2008**, *16*, 97–122. [[CrossRef](#)]
134. Duchateau, F.; Bellahsene, Z.; Coletta, R. A Flexible Approach for Planning Schema Matching Algorithms. In Proceedings of the OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Monterrey, Mexico, 9–14 November 2008; pp. 249–264.
135. Noy, N.F.; Musen, M.A. The PROMPT suite: Interactive tools for ontology merging and mapping. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 983–1024. [[CrossRef](#)]
136. Ghazvinian, A.; Noy, N.F.; Jonquet, C.; Shah, N.; Musen, M.A. What Four Million Mappings Can Tell You about Two Hundred Ontologies. In Proceedings of the International Semantic Web Conference, Chantilly, VA, USA, 25–29 October 2009; pp. 229–242.
137. Euzenat, J. Alignment infrastructure for ontology mediation and other applications. Proceedings of 1st ICSOC international workshop on Mediation in semantic web services, Amsterdam, Netherlands, 12 December 2005; pp. 81–95.
138. Do, H.H.; Rahm, E. Matching large schemas: Approaches and evaluation. *Inform. Syst.* **2007**, *32*, 857–885. [[CrossRef](#)]
139. Ehrig, M.; Staab, S. QOM—Quick Ontology Mapping. In Proceedings of the International Semantic Web Conference, Hiroshima, Japan, 7–11 November 2004; pp. 683–697.
140. Kirsten, T.; Kolb, L.; Hartung, M.; Groß, A.; Köpcke, H.; Rahm, E. Data Partitioning for Parallel Entity Matching. *Comput. Sci.* **2010**, *3*, 1–8. [[CrossRef](#)]
141. Bianco, G.D.; Galante, R.; Heuser, C.A. A fast approach for parallel deduplication on multicore processors. In Proceedings of the ACM Symposium on Applied Computing, TaiChung, Taiwan, 21–24 March 2011; pp. 1027–1032.
142. Hungsik, K.; Dongwon, L. Parallel linkage. In Proceedings of the ACM, Lisbon, Portugal, 6–10 November 2007; pp. 283–292.
143. Euzenat, J.; Meilicke, C.; Stuckenschmidt, H.; Shvaiko, P.; Trojahn, C. Ontology Alignment Evaluation Initiative: Six Years of Experience. In *Journal on Data Semantics XV*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 158–192.
144. Giunchiglia, F.; Yatskevich, M.; Avesani, P.; Shivaiko, P. A Large Scale Dataset for the Evaluation of Ontology Matching Systems. *Knowl. Eng. Rev.* **2009**, *24*, 137–157. [[CrossRef](#)]
145. Ferrara, A.; Montanelli, S.; Noessner, J.; Stuckenschmidt, H. Benchmarking Matching Applications on the Semantic Web. In Proceedings of the Extended Semantic Web Conference on the Semantic Web: Research and Applications, Heraklion, Crete, Greece, 29 May–2 June 2011; pp. 108–122.

146. Saveta, T.; Daskalaki, E.; Flouris, G.; Fundulaki, I.; Herschel, M.; Ngonga Ngomo, A.C. Pushing the Limits of Instance Matching Systems: A Semantics-Aware Benchmark for Linked Data. In Proceedings of the International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 105–106.
147. Saveta, T.; Daskalaki, E.; Flouris, G.; Fundulaki, I.; Herschel, M.; Ngomo, A.C.N. LANCE: Piercing to the Heart of Instance Matching Tools. In Proceedings of the International Semantic Web Conference, Bethlehem, PA, USA, 11–15 October 2015; pp. 375–391.
148. Euzenat, J.; Roşoiu, M.E.; Trojahn, C. Ontology matching benchmarks: Generation, stability, and discriminability. *J. Web Semant.* **2013**, *21*, 30–48. [[CrossRef](#)]



© 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/>).