IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 27, 2020, accepted September 11, 2020, date of publication September 21, 2020, date of current version October 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025611

An Efficient Approach for Measuring Semantic Similarity Combining WordNet and Wikipedia

FEI LI[1,2], LEJIAN LIAO[1], LANFANG ZHANG[3], XINHUA ZHU[2], BO ZHANG[4], AND ZHENG WANG[5]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2]Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
[3]Faculty of Education, Guangxi Normal University, Guilin 541004, China
[4]School of Mathematics and Computer Science, Hezhou University, Hezhou 542899, China
[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798

Corresponding authors: Lanfang Zhang (lfzhang64@163.com) and Bo Zhang (zbomail@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Contract 61967003 and Contract 62062012, in part by the Natural Science Foundation of Guangxi of China under Contract 2020GXNSFAA159082, and in part by the Research Fund of Guangxi Key Laboratory of Multi-source Information Mining & Security under Grant 20-A-03-02.

ABSTRACT The measurement of semantic similarity between concepts is an important research topic in natural language processing. In the past, several approaches for measuring the semantic similarity between concepts have been proposed based on WordNet or Wikipedia. However, improvements in the measurement accuracy of most methods have led to a dramatic increase in time complexity, and the existing methods do not effectively integrate WordNet and Wikipedia. In this paper, we focus on designing an efficient semantic similarity method based on WordNet and Wikipedia. To improve the accuracy of WordNet edge-based measures, we propose an edge weight model for combining edge and density information, which assigns a weight to each edge adaptively based on the number of direct hyponyms of the subsumer. Second, to improve the computational efficiencies of the existing Wikipedia link vector-based measures, we propose a new Wikipedia link feature-based semantic similarity method that converts Wikipedia links into semantic knowledge and replaces the TF-IDF statistical weight model in the existing measures. In addition, we propose two new word disambiguation strategies to further improve the accuracy of Wikipedia link-based measures. Finally, to fully exploit the advantages of WordNet and Wikipedia, we propose two new aggregation schemas for combining WordNet "*is-a*" semantics and Wikipedia link semantics to replace the current aggregation schemas that combine WordNet "*is-a*" semantics with category semantics in Wikipedia. The experimental results show that our aggregation models are outstanding in terms of accuracy, efficiency and word coverage compared to state-of-the-art similarity measures.

INDEX TERMS Semantic similarity, edge weight model, word disambiguation strategy, WordNet, Wikipedia.

I. INTRODUCTION

The measurement of the semantic similarity between concepts or words is an important fundamental research topic in natural language processing that can be widely applied in the fields of intelligent retrieval [1]–[3], word sense disambiguation [4], [5], machine learning [6], information extraction [7], semantic annotation [8] and semantic similarity between sentences [9]. Although neural network-based word vectors such as word2vec [10] have achieved good results in semantic relatedness measurements, they are still inferior to knowledge ontology-based methods in semantic similarity measurements. For example, Qu *et al.* [11] reported that various WordNet-based methods outperform word2vec method on typical similarity datasets such as MC30 and RG65. However, how to use more semantic knowledge to balance computational accuracy and computational efficiency is an important challenge for semantic similarity research based on knowledge.

Most of the popular semantic similarity algorithms [12]–[21] are implemented and evaluated by using WordNet as an underlying reference ontology due to its

The associate editor coordinating the review of this manuscript and approving it for publication was Weiping Ding.

clear concept hierarchies. In these methods, edge-based and information-content-based (IC-based) approaches remain the research focus of semantic similarity. Edge-based semantic similarity metrics are intuitive and easy to understand and have low computational complexity [21]. However, the density non-uniformity problem in large lexical taxonomies severely hampers the performances of edge-based similarity metrics [14], which would cause that the same concept paths in areas with different densities represent the same semantic distances in a taxonomic ontology. In [21], a density-based path compensation model was proposed in which the area density is incorporated into edge-based approaches by a smooth parameter to solve this problem. However, this measure is a supervised machine learning method and depends strongly on the ontologies and the training data.

IC-based approaches [12], [15], [18], [19], [22], [23] can overcome the density non-uniformity of a large taxonomy by considering hyponyms of concepts in the taxonomy. However, information content computation requires us to count the numbers of all hyponyms for the measured concept in the taxonomy [12], [15], [18], [19]. Thus, the information-content-based similarity metric has high computational complexity, which may prevent the popularization and application of these approaches in a dynamic ontology that is frequently updated.

WordNet is an ontology that was manually constructed by psychologists, linguists and computer engineers of Princeton University. With the exponential growth of online information on the World Wide Web, the shortcomings of the limited coverage of WordNet began to emerge, which may limit its scope of application. To overcome this problem, in recent years several researchers [11], [24]–[29] utilized a new knowledge resource, namely, Wikipedia, to measure semantic similarity. Wikipedia is an online collaborative encyclopedia that is maintained by volunteers from all over the world and has the following advantages [24]: (1) It has broad concept coverage in many languages. (2) New concepts and terms are always updated timely. (3) It contains many of senses for each word. Therefore, Wikipedia can effectively overcome the coverage limitation of WordNet. As an encyclopedia, Wikipedia contains a variety of data, which include categories, a taxonomic hierarchy that is similar to that of WordNet, articles that correspond to titles of web documents (pages) and are used to introduce concepts, and links between pages.

However, with the rapid growth of Wikipedia, the spaces of the concept vectors in article-based measures and the stem vectors in Wikipedia category-vector-based measures are increasing rapidly and their vector weights are sparse, which causes the performances of these models to decline sharply. Wikipedia outlink vector-based measures are the most promising methods because the links, which are manually defined, are limited and are closer to human semantics. However, they still have shortcomings that require be overcome: first, when a link vector is constructed, they must assign each vector a suitable weight via the TF-IDF scheme, which is a time-consuming process; second, their disambiguation strategy of simply using all the senses does not perform sufficiently well.

To exploit the advantages of both WordNet and Wikipedia, Aouicha *et al.* [22] proposed an aggregation schema that exploits the WordNet "*is-a*" semantics and the Wikipedia category graph in a complementary way to increase the coverage capacity. However, the Wikipedia category graph is not a rigorous "*is-a*" hierarchy as that in WordNet. For example, in Wikipedia, *Computer systems* is categorized in the upper category *Technology systems*, in which they are an "*is-a*" relationship, whereas *Computer hardware* is categorized in the upper category *Computer systems*, in which they are a "*has-part*" relationship rather than "*is-a*" relationship. The Wikipedia category graph is designed to facilitate the management of pages in Wikipedia; thus, it is a hybrid structure composed of various semantic relations and the similarity measurement based on its structure is unreliable.

To overcome the above issues, this paper designs an efficient semantic similarity method that is based on WordNet and Wikipedia. Firstly, we propose an edge weight model for combining edge and density information. Secondly, to improve the computational efficiencies of the existing Wikipedia link vector-based measures, we propose a new Wikipedia link feature-based semantic similarity method that converts Wikipedia links into semantic knowledge and two new word disambiguation strategies. Finally, to fully exploit the advantages of WordNet and Wikipedia, we propose two new aggregation schemas for combining WordNet "is-a" semantics and Wikipedia link semantics. We evaluate our method on the widely used datasets of MC30, RG65, AG203, SimLex666 and Pedersen30, and compare it with various advanced methods. The contributions of this paper can be summarized as follows:

1) In WordNet edge-based measures, we utilize the number of direct hyponyms of the upper node to assign a weight to each edge to adapt to the changes of the density in the paths between concepts in WordNet, thereby improving the accuracy of edge-based measurements, which can compensate for the shortcoming that a single-layer structure with numerous direct hyponyms may be converted into a multi-layer structure during the development of a large taxonomy. In contrast to the current supervised learning method [21], this edge weight model is an unsupervised machine learning method and can adapt to the development and updating of WordNet.
2) In Wikipedia link-based measures, we convert Wikipedia links into semantic knowledge based on the description logic and propose a new Wikipedia link feature-based semantic similarity method for improving the computational efficiency and accuracy of the TF-IDF statistical weight model in the existing Wikipedia link-based measures [30]–[32].
3) We propose two new word disambiguation strategies that are based on volunteer awareness, which directly sort the outlinks within a disambiguation page

according to the order in which they occur in the disambiguation page developed by volunteers, rather than according to the number of links within the articles of selected outlinks as in the existing methods [22], [26].

4) To take full advantage of WordNet's "*is-a*" taxonomy and Wikipedia's semantic knowledge, we propose two new aggregation schemas for combining WordNet "*is-a*" semantics and Wikipedia link semantics, which are more reasonable than the current aggregation schemas [22] that combine WordNet "*is-a*" semantics with category semantics in Wikipedia, and substantially outperform existing schemas in terms of accuracy, efficiency and word coverage.

The remainder of this paper is organized as follows: Section II provides an overview of the popular similarity approaches that are related to our study. Section III proposes an edge weight model for increasing the accuracy of path-based similarity measures. Section IV proposes a Wikipedia link feature-based ratio model and two word disambiguation strategies. Section V proposes two aggregation schemas that combine the advantages of WordNet and Wikipedia. Section VI describes the experiments in detail. Section VII discusses the experiment results. Section VIII presents the conclusions of this work.

II. RELATED WORKS

Several studies have been reported on the use of WordNet or Wikipedia as a knowledge resource to measure the semantic similarity between concepts. In this section, we present several main methods.

A. WORDNET-BASED MEASURES

WordNet has a clear subsumption hierarchy; hence, many measurement approaches have exploited the topological parameters that are extracted from the "*is-a*" taxonomy to assess the similarity between concepts.

1) EDGE-BASED APPROACHES

Intuitively, the shortest path length between two concepts is closely related to the similarity between them and the most direct approach to measure the similarity is to count the shortest path length in the semantic net, which is the main strategy of edge-based methods. Rada *et al.* [17] adopted this strategy in their method for measuring semantic similarity. In this method, the shortest path length is converted to a similarity metric with the maximum path length (*max-path*) in the taxonomic ontology, as expressed in the following equation:

$$sim_{Rada}(c_1, c_2) = 2 \times max\text{-}path - pathLen(c_1, c_2) \quad (1)$$

where $pathLen(c_1, c_2)$ is the shortest path length between concepts c_1 and c_2 and it is equal to the number of "*is-a*" links from c_1 and c_2.

Leacock *et al.* [13] exploited the maximum depth (*max-path*) in the taxonomic ontology to scale the shortest path length and proposed a logarithmic function for similarity assessment, which is defined as follows:

$$sim_{Leacock}(c_1, c_2) = -log\frac{pathLen(c_1, c_2) + 1}{2 \times max\text{-}depth} \quad (2)$$

However, these two methods do not reflect a common intuition: if concept pairs have the same shortest path length but unequal depths in a taxonomic ontology, their similarities differ. Liu *et al.* [16] introduced the relative depth of the lowest common subsumer (LCS) between concepts and proposed two methods for measuring the semantic similarity of concepts. Their fundamental strategy was to simulate the process of human judgment, which was based on the ratio of the common and different features between two concepts in the taxonomic hierarchy. They presented the following two equations:

$$sim_{Liu\text{-}1}(c_1, c_2) = \frac{\alpha \times depth(LCS(c_1, c_2))}{\alpha \times depth(LCS(c_1, c_2)) + \beta \times shortest\text{-}pathLen(c_1, c_2)} \quad (3)$$

$$sim_{Liu\text{-}2}(c_1, c_2) = \frac{e^{\alpha \times depth(LCS(c_1,c_2))} - 1}{e^{\alpha \times depth(LCS(c_1,c_2))} + e^{\beta \times shortest\text{-}pathLen(c_1,c_2)} - 2} \quad (4)$$

where $LCS(c_1, c_2)$ is the least common subsumer between concepts c_1 and c_2, $depth(LCS(c_1, c_2))$ is the depth of their least common subsumer relative to the *root*, and α and β are the smoothing factors for depth and path, respectively $(0 \le \alpha, \beta \le 1)$.

However, according to Li *et al.* [14], humans may process information nonlinearly; hence, they exploited these two features and proposed a non-linear function for measuring the semantic similarity.

$$sim_{Li}(c_1, c_2) = e^{-\alpha \times shortest\text{-}pathLen(c_1,c_2)} \times \frac{e^{\beta \times depth(LCS(c_1,c_2))} - e^{-\beta \times depth(LCS(c_1,c_2))}}{e^{\beta \times depth(LCS(c_1,c_2))} + e^{-\beta \times depth(LCS(c_1,c_2))}} \quad (5)$$

where α and β are the smoothing factors, which scale the contributions of $pathLen(c_1, c_2)$ and $depth(LCS(c_1, c_2))$ $(0 \le \alpha, \beta \le 1)$.

2) INFORMATION-CONTENT-BASED APPROACHES

Information-content-based similarity measures commonly rely on the IC that is assigned to the concepts. The IC of a concept is the amount of information that is provided by the concept when it appears in a context [18].

Resnik [18] was the first to combine an ontology and a corpus. He stated that the similarity between concepts depends on the amount of shared information between them and proposed an IC-based similarity measure.

$$sim_{Resnik}(c_1, c_2) = IC(LCS(c_1, c_2)) \quad (6)$$

However, Resnik's method has a similar shortcoming to Rada's method: He only considered the information of the least common subsumer between concepts and did not consider the information that was contained in concepts. Jiang *et al.* [12] focused on this problem and proposed a distance-based method that relies on the information of the concepts and the least common subsumer between them. In their method, the length of a taxonomical link is quantified as the difference between the IC values of a concept and its subsumer. To compute the semantic distance between two concepts, they calculated the sum of the ICs of the individual concepts minus the IC of their LCS. Eq. (7) expresses this measure.

$$Distance_{Jiang}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2)) \quad (7)$$

Lin [15] also focused on this problem and proposed a new method: He exploited the ratio of the commonalities between concepts and their full information-needed as the similarity score, which is defined as follows:

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (8)$$

There are two main IC computation models: the corpora-based IC computation model and the intrinsic IC computation model. The former requires a large corpus of text documents for calculating the probability of a concept and it was mainly used in the early stages. The latter is based on the hierarchical structure of a taxonomic ontology.

Resnik [18] proposed the IC computation method and he used the probability of concept c in a specified environment. The IC value is calculated via Eq. (9):

$$IC(c) = -log(P(c)) \quad (9)$$

where $P(c)$ is calculated via Eq. (10):

$$P(c) = \frac{\sum_{w \in Word(c)} count(w)}{N} \quad (10)$$

where $Word(c)$ is the set of words that are subsumed by concept c, $count(w)$ is the frequency of word w in the corpus, and N is the total number of observed words in the corpus.

The IC of a concept is proportional to the information it contains. In the hierarchical structure, hyponym (descendant) nodes reflect the IC of a concept: the more hyponym nodes of a concept, the smaller the IC of the concept. Seco *et al.* [19] exploited this and proposed an IC computing model, which is defined as follows:

$$IC(c) = 1 - \frac{log(|hypon(c)|)}{log(max\text{-}nodes)} \quad (11)$$

where $hypon(c)$ is the set of all hyponym nodes of concept c (contain itself) and *max-nodes* is a constant that represents the total number of nodes in the taxonomic ontology.

According to Sánchez *et al.* [33], the leaf nodes that a concept contains more reasonably reflect its IC and proposed a new IC computing model.

$$IC(c) = -log(\frac{|leaves(c)| \,/\, |subsumers(c)|}{max\text{-}leaves}) \quad (12)$$

where $leaves(c)$ is the set of leaf nodes of concept c; $subsumers(c)$ is a set of hypernym (ancestor) nodes, which balances the contribution of $leaves(c)$; and *max-leaves* is a constant that represents the total number of leaf nodes in the taxonomic ontology. Since this model must count all hypernym nodes when it identifies the leaf nodes of a concept, it is highly time-consuming.

To overcome the limitations of Rada's method in path-based approaches, Zhou *et al.* [34] increased relative depth of the concept based on Seco's method to improve the IC measure. Eq. (13) represents this measure.

$$IC(c) = k(1 - \frac{log(|hypon(c)|)}{log(max\text{-}nodes)}) + (1-k)(\frac{log(depth(c))}{log(max\text{-}depth)}) \quad (13)$$

where $hypon(c)$, $depth(c)$, *max-nodes* and *max-depth* have the same meaning as in previous approaches and k is a smoothing factor.

3) FEATURE-BASED APPROACHES

Tversky [35] extracted semantic knowledge from multiple semantic relationships to construct the feature descriptions of concepts c_1 and c_2. The more shared features there are between concepts, the higher their semantic similarity. His ratio model is expressed as follows:

$$sim_{Tver}(c_1, c_2) = \frac{f(\psi(c_1) \cap \psi(c_2))}{f(\psi(c_1)\cap\psi(c_2)) + \alpha f(\psi(c_1) - \psi(c_2)) + \beta f(\psi(c_2) - \psi(c_1))} \quad (14)$$

where $f(\bullet)$ is a measure function on the feature space, which measures the contribution of (common or distinctive) features to the similarity between concepts; $\psi(c_1)$ and $\psi(c_2)$ are the feature description sets of concepts c_1 and c_2, respectively, and each contains features that are based on multiple semantic relationships; $\psi(c_1) \cap \psi(c_2)$ is the overlap of sets $\psi(c_1)$ and $\psi(c_2)$; $\psi(c_1) - \psi(c_2)$ denotes the semantic features that belong to concept c_1 but not to c_2 and $\psi(c_2) - \psi(c_1)$ is the inverse; and α and β are the weighting parameters, which satisfy $\alpha, \beta \geq 0$. In [20], for implementation in WordNet, the authors considered features of a concept c in the set of synsets, which is formed by its ancestors in the "*is-a*" taxonomy, along with the meronyms, holonyms and attributes of each ancestor and the hyponyms. The function $f(\bullet)$ is used to measure the cardinalities of various feature sets. The experiments are performed with $\alpha = 0.5$ and $\beta = 0.5$.

Rodríguez and Egenhofer [36] used the weighting parameters to linearly combine all similarities that are based on synsets (syns), neighbor concepts (those linked via semantic relations) and features (e.g., *meronymy* and *attribute*) to calculate the final similarity result, which is expressed as follows:

$$sim_{RE}(c_1, c_2) = \alpha S_{syns}(c_1, c_2) + \beta S_{features}(c_1, c_2) + \gamma S_{neighbors}(c_1, c_2) \quad (15)$$

where α, β, and γ are the weighting parameters, which are tuned according to the ontology, and S refers to the overlapping function, which is expressed as follows:

$$S(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| + \delta(c_1, c_2)|A \setminus B| + (1 - \delta(c_1, c_2))|B \setminus A|} \tag{16}$$

where A and B are the feature description sets that correspond to c_1 and c_2, respectively; $A \setminus B$ denotes the features that are in set A but not in set B; $B \setminus A$ is the inverse of $A \setminus B$; and the parameter δ depends on the depths of two concepts:

$$\delta(c_1, c_2) = \begin{cases} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, & depth(c_1) \le depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, & depth(c_1) \ge depth(c_2) \end{cases} \tag{17}$$

This method fully utilizes various semantic relationships to more accurately determine the similarity. However, the tuning of the weighting parameters mainly depends on the ontology.

Petrakis *et al.* [37] proposed the *X-similarity* function, in which structural ("*is-a*") and textual (gloss) features and neighbors are linked via semantic relations. Rather than using the multiple-semantic linear combination model, as Rodríguez and Egenhofer did, Petrakis *et al.* proposed a multiple-semantic complementary model that maximized the individual similarities, which was defined as (18), shown at the bottom of the page.

In Eq. (18), S is the ratio of the number of common features to the total number of features of both concepts. The semantic similarity that is based on neighbors depends on the maximum of each semantic relationship ("*is-a*" and "*part-of*"):

$$S_{neighbors}(c_1, c_2) = \max_{i \in SR} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \tag{19}$$

where each semantic relation type ("*is-a*" and "*part-of*" in WordNet) is computed separately and the neighbors come from all the synsets of all hypernyms up to the *root* of each hierarchy. The semantic similarity between gloss or synset sets is calculated as follows:

$$S(c_1, c_2) = \frac{|A \cap B|}{|A \cup B|} \tag{20}$$

where A and B are the gloss or synset sets of concepts c_1 and c_2, respectively.

By separately computing the similarities between semantic relationships and selecting the maximum similarity as the final result fully utilizes each feature. However, the accuracy of this method is not sufficient because it uses the same algorithm to calculate the similarities based on different relationships.

4) WEIGHTING-BASED APPROACHES

Weighting-based similarity measures have been proposed to estimate the semantic similarity between two concepts, and the core of these approaches is to explore a weight mechanism to weigh the degree of relevance of features in the semantic representation of a concept or the semantic distance between concepts.

Saif *et al.* [38] considered the semantic representation of a concept as a set of concepts that are extracted from its hypernym-concepts in a semantic taxonomy, and they then proposed four weight mechanisms to weigh the degree of relevance of features by using topological parameters (edge, depth, descendants, and density) in a semantic taxonomy. The weight mechanism with descendants has achieved the best results in their experiments, so we just show this one. This mechanism exploits the number of descendants of a concept and reflects the important parameters in semantic measures. The weight of concept c is expressed as follows:

$$w(c) = log(\frac{max\text{-}nodes}{|hypon(c)| + 1}) \tag{21}$$

where *max-nodes* and *hypon*(c) have the same meaning as in previous approaches. Finally, the semantic similarity between two concepts is equal to the cosine value between two semantic representations.

B. WIKIPEDIA-BASED MEASURES

According to the types of data that are used, Wikipedia-based measures can be divided into four groups:

1) CATEGORY STRUCTURE-BASED MEASURES

Category structure-based measures [22], [27], [39] only use Wikipedia's categories and category structure and, thus, have lower time complexity. However, since Wikipedia's category structure is not a rigorous "*is-a*" hierarchy, the accuracy is not high.

Strube *et al.* [39] were the first to measure semantic similarity using Wikipedia. Their approach, namely, WikiRelate!, is based on Wikipedia's category structure. They obtained web pages to which specified word pairs correspond and extracted the categories to which these pages belong. Finally, they exploited the paths that are formed by links between categories in Wikipedia's category structure to compute the semantic similarity. However, Wikipedia's categories do not follow rigorous "*is-a*" hierarchy; therefore, the accuracy of this method is not high.

$$sim_{\textbf{X-similarity}}(c_1, c_2) = \begin{cases} 1, & if\ S_{syns}(c_1, c_2) > 0 \\ \max\{S_{neighbors}(c_1, c_2), S_{gloss}(c_1, c_2)\}, & if\ S_{syns}(c_1, c_2) = 0 \end{cases} \tag{18}$$

Jiang *et al.* [27] exploited Wikipedia's category structure as an "*is-a*" hierarchy that is similar to that of WordNet to measure the semantic similarity. They identified the categories to which specified word pairs correspond by directly querying the category structure and proposed a k-approximate IC computation method for computing the similarity between categories. In their method, they directly match concepts to Wikipedia's category nodes rather than to the articles in Wikipedia. Relative to the articles in Wikipedia, the number of category nodes in Wikipedia is small. Therefore, this method has very low measurement coverage.

2) ARTICLE-BASED MEASURES

Article-based measures [40]–[42] utilize machine learning technologies to determine the similarity between concepts based on the content of web pages. These methods realize substantial accuracy improvement in semantic similarity measurement in early Wikipedia versions (2006 to 2008). However, due to its rapid growth, Wikipedia now has many articles and each article provides a substantial amount of information, which leads to huge space and time costs for these methods. Studies have shown that the accuracy of explicit semantic analysis [40] (ESA, which is a typical article-based measure) has declined sharply in similarity measurement due to its vector sparseness [27], [43]. Therefore, effectively reducing the space and time costs has become a key problem for these methods.

Gabrilovich *et al.* [40] proposed an explicit semantic analysis model (ESA) that utilized the meanings of texts in a high-dimensional space of concepts from Wikipedia to measure the similarity. This method substantially improved the similarity measurement accuracy in early Wikipedia versions. However, as we discussed, with the rapid growth of Wikipedia, its similarity measurement performance has declined sharply due to vector sparseness.

3) WIKIPEDIA CATEGORY-VECTOR-BASED MEASURES

Wikipedia category-vector-based measures [26], [29], [44] utilize the stems of all the articles in a Wikipedia category to build a category vector and convert the similarity between concepts to the cosine of the vectors of the categories to which their articles belong. These methods involve the category structure and article pages; therefore, they perform similarly to ESA [40] on an early Wikipedia version (2008). However, they have two main drawbacks: First, their space and time complexities are close to those of ESA and their efficiency is much lower compared to Wikipedia outlink vector-based measures. Second, these methods measure the similarity between all concepts that are in the same category as the maximum 1, namely, they consider all the concepts in the same category as synonyms, which is unreasonable.

To overcome ESA's drawbacks such as high-dimensional space and high-computational complexity, Li *et al.* [29] use the top k of concept vectors and Wikipedia Category Graph (WCG) to improve the ESA method. First, for each term in a word pair, the top k most relevant Wikipedia concepts are returned by the Naive-ESA algorithm to reduce the dimensional space of the ESA method. Second, for each different candidate concept in two relevant concept sets, they collect its categories set from WCG and use category vector to compute the similarity between concepts in two difference lists.

4) WIKIPEDIA OUTLINK VECTOR-BASED MEASURES

Wikipedia outlink vector-based measures [30], [31] use outlinks that occur in the web page of an article to construct a link vector for measuring semantic similarity. Since Wikipedia provides link data for each page in its database dumps, these methods need not parse the contents of a Wikipedia page and do not depend on any machine learning technologies; hence, they have low time costs and satisfactory generality. Further increasing the computational efficiency is an important direction for improving these methods.

Milne [30] and Milne and Witten [31] exploited outlinks in Wikipedia to construct a link vector for each concept and assigned a weight to each vector element. Finally, they defined the similarity between two concepts as the angle between two vectors. The weight w of link $b \rightarrow a$ is computed as follows:

$$w(b \rightarrow a) = |b \rightarrow a| \times log(\sum_{x \in A}^{|A|} \frac{|A|}{|x \rightarrow a|}) \tag{22}$$

where $b \rightarrow a$ denotes that the text of article b contains anchor text a, $|b \rightarrow a|$ is the number of times that the text of word b contains anchor text a, and A represents the set of all articles in Wikipedia.

To improve the measurement accuracy of the existing Wikipedia outlink model in Eq. (22), Zhu *et al.* [32] utilized the outlinks and inlinks of concepts in Wikipedia to combine into a bidirectional link vector for concept semantic interpreter and uses a TF-IDF-based bidirectional weight method to uniformly calculate the strength of the mutual association between a given concept and its outlink or inlink concept.

C. WordNet-WIKIPEDIA-BASED MEASURES

Aouicha *et al.* [22] used the WordNet "*is-a*" taxonomy and the Wikipedia category graph and proposed an aggregation schema for computing semantic similarity. In their method, they proposed an IC computing model, which they applied to two graphs and calculated their concept semantic similarities separately. Finally, they used an aggregation strategy to obtain the final similarity. This strategy is expressed as follows:

$$sim_{Aou}(c_1, c_2) = \begin{cases} sim_{WNet}(c_1, c_2), & if\ sim_{WNet}(c_1, c_2) > \theta \\ sim_{Wiki}(c_1, c_2), & else \end{cases} \tag{23}$$

where $sim_{WNet}(c_1, c_2)$ and $sim_{Wiki}(c_1, c_2)$ refer to the semantic similarities that are provided by the WordNet "*is-a*" taxonomy and the Wikipedia category graph, respectively, under the IC computing model that they proposed and θ is a threshold.

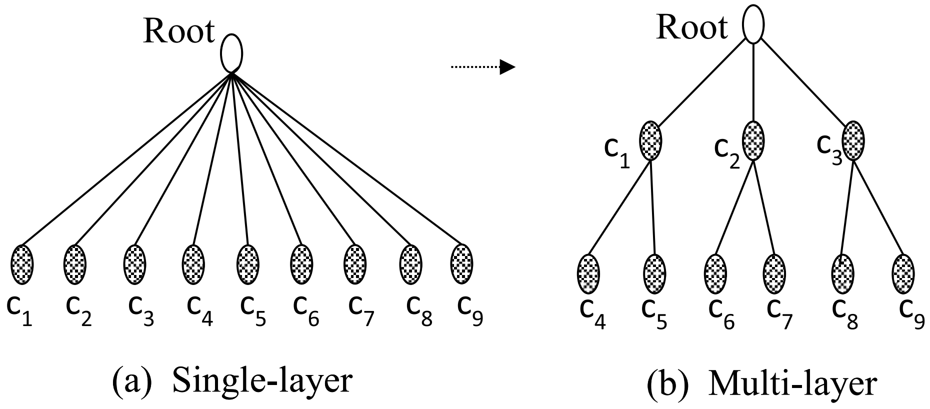

Root
c_1 c_2 c_3 c_4 c_5 c_6 c_7 c_8 c_9
(a) Single-layer
Root
c_1 c_2 c_3
c_4 c_5 c_6 c_7 c_8 c_9
(b) Multi-layer

FIGURE 1. Development of the taxonomy.

In this method, they considered the Wikipedia category graph as a taxonomy. However, the Wikipedia category graph is not a rigorous "*is-a*" hierarchy as that in WordNet. Therefore, the semantic similarity measurement accuracy of this method is not high.

Zhu *et al.* [45] utilized the knowledge graph DBpedia,[1] which is composed of structured information extracted from Wikipedia, to propose a semantic similarity method that combines the concepts' path in WordNet and the shared information content of concepts in DBpedia. Their method is a weighted path length (wpath) and expressed as follows:

$$sim_{wpath}(c_1, c_2) = \frac{1}{1 + length(c_1, c_2) \times k^{IC(LCS(c_1,c_2))}} \tag{24}$$

where the parameter k represents the contribution of the LCS's IC and $k \in (0, 1]$, $length(c_1, c_2)$ refers to the shortest path length between concepts c_1 and c_2 in WordNet. $IC(LCS(c_1, c_2))$ refers to the IC of two concepts' LCS in DBpedia, which is computed by entities that the concept has in DBpedia rather than by the hyponyms in WordNet, it is expressed as follows:

$$IC(LCS(c_1, c_2)) = -log\left(\frac{count((entities(LCS(c_1, c_2)))}{N}\right) \tag{25}$$

where N denotes the total number of entities in DBpedia, $entities(c)$ is the function to retrieve set of entities having type of c in DBpedia.

In the above weighted path similarity method, concept's IC is computed by entities of the concept in DBpedia, which is more reasonable than that computed by the hyponyms of the concept in WordNet. However, since DBpedia lacks sufficient disambiguation information, how to accurately align the concepts in DBpedia and WordNet becomes a problem. Moreover, the generation of DBpedia requires additional information extraction techniques and DBpedia cannot be updated as promptly as Wikipedia.

III. WORDNET-BASED EDGE WEIGHT SEMANTIC SIMILARITY COMPUTATION

As discussed previously, most path-based semantic similarity measures directly count the number of edges that connect two specified concepts to calculate the length of their shortest path. Intuitively, the edges between any two adjacent nodes are not necessarily of the same link strength because the density of the area they are in is different; therefore, it is necessary to weight the edge connection of the nodes. In this section, we propose an edge weight model for improving the accuracy of path-based similarity measures. In the process of assigning weights to edges, we utilize an information theory model and deduce an edge weight function.

Our proposed model is derived from a widespread phenomenon: In a large taxonomy, a single-layer structure with numerous direct hyponyms may be converted into a multi-layer structure during the development of the taxonomy. For example, with the development of WordNet from version 2.1 to 3.0, the maximum depth of the classification level has developed from 16 to 19 and the average number of direct hyponyms of each node has decreased from 3.2 to 2.7. Fig. 1(a) illustrates a single-layer structure with 9 direct hyponyms and after taxonomic development, it evolves into a double structure, as illustrated in Fig. 1(b). Thus, the shortest path length between concepts c_4 and c_9 changes from 2 to 4 and the depths of concepts c_4 and c_9 both change from 1 to 2. Therefore, the edge between a child concept and its parent concept does not necessarily have the same link strength in both structures and the link strength depends on the densities of concepts. To compensate for the shortcomings of edge-based similarity measures regarding this phenomenon, we propose an edge weight model.

We argue that the edge weights reflect the link strengths of concepts in the taxonomy and each edge weight is equal to the semantic distance between the concepts that are linked by the corresponding edge. Therefore, the edge weight can be defined as follows:

Definition 1: Let $e_{s \to p}$be an edge from concept s to concept p, where p is a direct hypernym (parent) of the concept s

[1] https://wiki.dbpedia.org/

TABLE 1. Proof for proposition 1.

Step	Reasoning	Explanation
1	$Weight(e_{s\to p}) = Distance(s,p) \Rightarrow$	According to Eq. (26)
2	$Weight(e_{s\to p}) =$ $IC(s) + IC(p) - 2 \times IC(LCS(s,p)) \Rightarrow$	According to Eq. (7)
3	$Weight(e_{s\to p}) =$ $IC(s) + IC(p) - 2 \times IC(p) \Rightarrow$	The LCS between the son s and its parent p is equal to its parent p
4	$Weight(e_{s\to p}) = IC(s) - IC(p) \Rightarrow$	By simplifying Step 3
5	$Weight(e_{s\to p}) =$ $1 - \frac{log(\lvert hypon(s)\rvert)}{log(max\text{-}nodes)} - (1 - \frac{log(\lvert hypon(p)\rvert)}{log(max\text{-}nodes)}) \Rightarrow$	Put Eq. (11) into Step 4
6	$Weight(e_{s\to p}) = \frac{log(\frac{\lvert hypon(p)\rvert}{\lvert hypon(s)\rvert})}{log(max\text{-}nodes)} \Rightarrow$	By simplifying Step 5
7	$Weight(e_{s\to p}) \approx$ $\frac{log(\frac{\lvert hypon(s)\rvert \times \lvert directhypon(p)\rvert}{\lvert hypon(s)\rvert})}{log(max\text{-}nodes)} \Rightarrow$	The hyponym number between s and its sibling concepts at the same level are approximately equal in a large ontology, so: $\lvert hypon(p)\rvert \approx \lvert hypon(s)\rvert \times \lvert directhypon(p)\rvert$
8	$Weight(e_{s\to p}) \approx \frac{log(\lvert directhypon(p)\rvert)}{log(max\text{-}nodes)} \Rightarrow$	By simplifying Step 7
9	$Weight(e_{s\to p}) \approx log(\lvert directhypon(p)\rvert)$ □	The *max-nodes* is a constant in an ontology and can be deleted

(son / child). Thus, the weight of edge $e_{s\to p}$ is defined as follows:

$$Weight(e_{s\to p}) = Distance(s, p) \tag{26}$$

Proposition 1: The weight of the edge between a child concept and its parent concept is approximately equal to the logarithm of the number of direct hyponyms of the parent concept. Formally, it can be defined as follows:

$$Weight(e_{s\to p}) \approx log(|directhypon(p)| + 1) \tag{27}$$

where $directhypon(p)$ refers to the set of direct hyponyms of concept p. According to Eq. (27), the weight of the edge between a child concept s and its parent concept p is approximately equal to the logarithm of the number of direct hyponyms of the parent concept. The direct hyponym number plus 1 is to ensure that the weight is not 0 when the direct hyponym number of the parent node p is 1. Eq. (27) is proved in Table 1.

Definition 2: Let c_1 and c_2 be any two concepts in a taxonomy. According to Eq. (27), we can calculate the shortest path length between concepts c_1 and c_2 based on our edge weight model, which is defined as follows:

$$\begin{aligned} pathLen_{weight}&(c_1, c_2) \\ &= \sum_{\forall e_{s\to p} \text{ in } path(c_1,c_2)} Weight(e_{s\to p}) \\ &\approx \sum_{\forall e_{s\to p} \text{ in } path(c_1,c_2)} log(|directhypon(p)| + 1) \end{aligned} \tag{28}$$

where $path(c_1, c_2)$ refers to the shortest path from c_1 to c_2 and p refers to the hypernym in edge $e_{s\to p}$. Note that there exists the same path length between synonyms of c_1 and c_2 in the taxonomy.

Definition 3: Let $LCS(c_1, c_2)$ be the lowest common subsumer between concepts c_1 and c_2. According to Eq. (27), we can calculate the depth of their lowest common subsumer relative to the root based on our edge weight model, which is defined as follows:

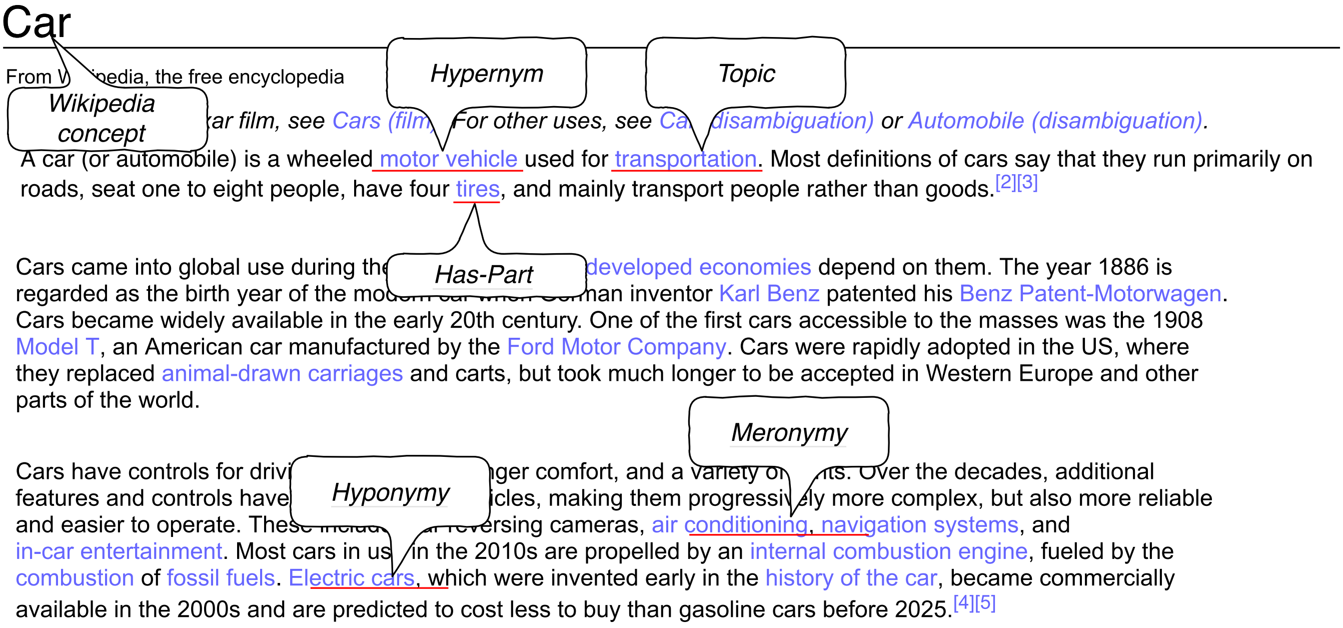
$$\begin{aligned} depth_{weight}&(LCS(c_1, c_2)) \\ &= \sum_{\forall e_{s\to p} \text{ in } path(LCS(c_1,c_2),root)} Weight(e_{s\to p}) \\ &\approx \sum_{\forall e_{s\to p} \text{ in } path(LCS(c_1,c_2),root)} log(|directhypon(p)| + 1) \end{aligned} \tag{29}$$

where *root* refers to the root of the taxonomy, $path(LCS(c_1, c_2), root)$ to the maximum path from the least common subsumer between concepts c_1 and c_2 to the *root*, and p refers to the hypernym in edge $e_{s\to p}$.

Our edge weighting method is a generic path computing model, which is an extension of the edge-counting model that is obtained by combining the edge counting model with information theory and can be applied with various edge-based similarity approaches in different taxonomies. The original structure of the algorithm formulas remains unchanged and we use our edge weight model only to replace the edge-counting-based path and depth computations in the measurement formulas.

IV. WIKIPEDIA LINK FEATURE-BASED SIMILARITY RATIO MODEL

To overcome the complex statistics shortcoming of Wikipedia outlink vector-based measures (see Section I and Section II-B for details), we convert Wikipedia outlinks into semantic knowledge and propose a novel Wikipedia



Car

From W pedia, the free encyclopedia

ar film, see Cars (film . For other uses, see Ca disambiguation) or Automobile (disambiguation).

A car (or automobile) is a wheeled motor vehicle used for transportation. Most definitions of cars say that they run primarily on roads, seat one to eight people, have four tires, and mainly transport people rather than goods.[2][3]

Cars came into global use during the developed economies depend on them. The year 1886 is regarded as the birth year of the mo an inventor Karl Benz patented his Benz Patent-Motorwagen. Cars became widely available in the early 20th century. One of the first cars accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company. Cars were rapidly adopted in the US, where they replaced animal-drawn carriages and carts, but took much longer to be accepted in Western Europe and other parts of the world.

Cars have controls for drivi ger comfort, and a variety o ts. Over the decades, additional features and controls have icles, making them progressively more complex, but also more reliable and easier to operate. The rsing cameras, air conditioning, navigation systems, and in-car entertainment. Most cars in us in the 2010s are propelled by an internal combustion engine, fueled by the combustion of fossil fuels. Electric cars, which were invented early in the history of the car, became commercially available in the 2000s and are predicted to cost less to buy than gasoline cars before 2025.[4][5]

FIGURE 2. **Wikipedia article for *Car*.**

link feature-based ratio model for measuring semantic similarity.

A. PROPOSED MODEL

The article is the basic unit of information in Wikipedia and each article typically describes a complete concept. In a Wikipedia concept article, there are plenty of manually defined anchor concepts (called outlinks) that link to other concept articles in Wikipedia. A concept is labeled as an outlink by a Wikipedia volunteer in an article because it has a semantic relationship with the host concept that corresponds to the article. For example, in Fig. 2, the outlink concept *motor vehicle* is the restriction target of the "*hypernymy*" relation in the host concept *car* and the outlink concept *tire* is the restriction target of the "*has-part*" relation in the host concept *car*. In addition, from a broad perspective, all outlinks in an article can be seen as typical semantic components of the article content, so the article content and its outlinks can form a "*has-part*" relationship similar to that in WordNet. According to the description logic [46] in the ontology, the semantics of a concept can be described by the outlink concepts in its Wikipedia article. Here, we present a proposition and define a formal representation for our model that is based on Wikipedia links.

Proposition 2: For any outlink concept l in a Wikipedia article P_c, there must be a semantic relation according to which l is the restriction target of the semantic relation in the host concept c of Wikipedia article P_c. Formally, this can be expressed as follows:

$$\forall l \in P_c \Rightarrow \exists r \in R_c(l \in Target(r, c)) \tag{30}$$

where R_c refers to the set of semantic relations of concept c in the real world and $Target(r, c)$ refers to the set of restriction targets of semantic relation r in concept c.

The proof of Proposition 2 is immediate via contradiction: Assume that for an outlink concept l_i, there is no corresponding semantic relationship according to which it is related to the host concept. A Wikipedia volunteer labels a concept as an outlink in the article only if the concept is associated with the host concept in his mind and the association between concepts can become a semantic relationship. Therefore, according to the assumptions, the volunteer must not label the concept l_i as an outlink.

Definition 4: Let c be a Wikipedia concept and P_c be a Wikipedia article for concept c. We define the semantic feature description of concept c as follows:

$$Des(c) = \{Target(r_i, c)|r_i \in R_c\} \simeq \{l_i|l_i \in P_c\} \tag{31}$$

where R_c refers to the set of semantic relations of concept c in the real world; $Target(r_i, c)$ refers to the set of the restriction targets of semantic relation r_i in concept c; and l_i refers to any outlink in Wikipedia article P_c.

According to Proposition 2 and Definition 4, we can apply the outlink-based semantic feature descriptions of concepts to Tversky's feature-based similarity ratio model in Eq. (14), use the function $f(\bullet)$ in Eq. (14) to measure the cardinalities of various feature sets, and let:

$$\begin{aligned}
&\alpha = 1, \beta = 1\\
&f(\psi(c_1)\cap\psi(c_2))+f(\psi(c_1)-\psi(c_2))=f(\psi(c_1))=|Des(c_1)|\\
&f(\psi(c_1)\cap\psi(c_2))+f(\psi(c_2)-\psi(c_1))=f(\psi(c_2))=|Des(c_2)|\\
&f(\psi(c_1)\cap\psi(c_2)) = |Des(c_1)\cap Des(c_2)|
\end{aligned}$$

We propose our Wikipedia link feature-based ratio model that based on logarithms as follows:

Definition 5: Let c_1 and c_2 be any two concepts in Wikipedia. The similarity between them is defined as (32), shown at the bottom of the page, where 1 is added in the numerator to avoid the scenario in which the set of common features is empty. $|Des(c)|$ represents the number of features of concept c, which is equal to the number of outlinks of the article for concept c in Wikipedia. Since there are typically fewer identical links between concept articles in Wikipedia, we use logarithms to optimize Tversky's formula to avoid the similarity being too small. The logarithm is a monotone function; hence, our model still accords with Tversky's feature theory.

B. WORD DISAMBIGUATION STRATEGY IN WIKIPEDIA

A word can represent multiple meanings. To compute the semantic similarity between two words from Wikipedia, we must identify the term or sense of interest for these two words. Since Wikipedia is edited by volunteers, the terms of a word are highly comprehensive (the average term count of a concept exceeds thirty in Wikipedia). Therefore, we cannot directly use the Cartesian product of the term lists of two words to compute the semantic similarity between them as in WordNet because this is too time-consuming and may have a negative impact.

Two similar strategies are available for word disambiguation in Wikipedia. Both strategies are divided into four steps and they have the same first, second and third steps. First, they obtain two term lists for the two words; for example, the word *rook* has term list L_1 = {rook (bird), montes rook, rook (surname), rook (chess), … } and the word *king* has term list L_2 = {king, fort king, king valley, king (surname), king (chess),… }. Then, they extract the elements with bracketed strings from these two lists to form two new lists, namely, $L_1{}'$ and $L_2{}'$; for example, *rook (bird)* belongs to new list $L_1{}'$. Next, they cyclically match the strings in parentheses from the two new lists. If a string in parentheses is the same as another word that is being compared, its corresponding element is retained in the new list; otherwise, it is removed from the new list. The final lists of new terms are used as the sets of senses for similarity calculations; for example, $L_1{}'$ = {rook (surname), rook (chess)} and $L_2{}'$ = {king (surname), king (chess)}. If either $L_1{}'$ or $L_2{}'$ is empty, one strategy is to select the first link in the disambiguation page as the term of interest (Called *Single match I*). Another strategy is to select the most commonly used term, namely, the term with the most outlinks, as the term of interest (Called *Single match II*). Both strategies attempt to extract one of the most commonly used terms from the disambiguation page to avoid the negative effects of extracting all the terms from the disambiguation page. However, these two strategies may lead to a problem: if two words are of high semantic similarity, these strategies may result in lower similarity because words typically have several common meanings in various contexts. To solve this problem, based on volunteer awareness, we propose two new strategies for word disambiguation in Wikipedia. We argue that the volunteers tend to list the more commonly used terms at the top of the page when they edit the disambiguation page. For example, Strube and Ponzetto [39] selected the first article linked in the disambiguation page to participate in the semantic calculation. Hadj Taieb *et al.* [25] selected the two first links existing in the ordered out-link set of the disambiguation page. Therefore, the order of the terms in the disambiguation page reflects the volunteer's view of the popularity of the corresponding terms. The proposed strategies are defined as follows:

Strategy 1: Proportional model. First, the terms within a disambiguation page are sorted according to the order in which they occur in the disambiguation page (called *volunteer awareness*). Then, suppose there are n terms in the disambiguation page for word w. We define the disambiguation term list L_w^θ with a proportional threshold $\theta(\theta \in (0, 1])$ as follows:

$$L_w^\theta = \{t_{w,i} | \theta = \frac{|\{t_{w,1}, t_{w,2}, \ldots, t_{w,i}\}|}{|\{t_{w,1}, t_{w,2}, \ldots, t_{w,n}\}|}\} \tag{33}$$

where $t_{w,i}$ refers to the ith term in the disambiguation page for word w.

Strategy 2: Number model. First, terms within a disambiguation page are sorted based on their volunteer awareness. Then, suppose there are n terms in the disambiguation page for word w. We define the disambiguation term list with a number threshold $m(m \in (0, 10])$ as follows:

$$L_w^m = \{t_{w,i} | i \le n \wedge i \le m\} \tag{34}$$

C. STUDY OF COMPLEXITY AND PORTABILITY

Wikipedia is a global multilingual encyclopedia that contains various versions of Wikipedia written in as many as 303 languages. Therefore, in addition to its good performance, an excellent semantic similarity system based on Wikipedia should be simple enough so as to be shared and migrated between various Wikipedia versions in different languages. In this section, we will discuss the complexity and portability of our proposed Wikipedia outlink feature model by comparing it against other Wikipedia-based similarity methods in terms of the complexity of treatment processes. First, we analyze and summarize the majority of the treatments that may appear in various semantic similarity systems based on Wikipedia, including pre-treatment and computing process, as shown in Table 2; and then, we analyze and give

$$sim_{Wikipedia}(c_1, c_2) = \frac{log(|Des(c_1) \cap Des(c_2)| + 1)}{log(|Des(c_1)|) + log(|Des(c_2)|) - log(|Des(c_1) \cap Des(c_2)| + 1)} \tag{32}$$

TABLE 2. Majority of treatments in Wikipedia-based semantic similarity systems.

No	Treatment	Pre-treatment	Computing process
(1)	Converting Wikipedia dump to an SQL database	$\surd$	
(2)	Filtering the articles contents	$\surd$	
(3)	Construction of the Wikipedia Category Graph (WCG)	$\surd$	
(4)	Filtering the WCG	$\surd$	
(5)	Providing the category, concept or word semantic description vectors using TF-IDF	$\surd$	
(6)	Access to the redirections		$\surd$
(7)	Disambiguation using a certain algorithm		$\surd$
(8)	Extracting of categories assigned to a word w_i		$\surd$
(9)	Computing the similarity using the semantic description vectors assigned to the words couple (w_1, w_2)		$\surd$
(10)	Computing the similarity using the WCG		$\surd$
(11)	Computing the similarity using the outlink feature		$\surd$

TABLE 3. Treatment processes of four Wikipedia-based semantic similarity systems.

System	Treatment processes	TF-IDF computing
Wikirelate! [39]	(1) → (3) → (4) → (6) → (7) → (8) → (10)	No
Explicit semantic analysis-based [11], [40]	(1) → (2) → (5) → (6) → (7) → (9)	Yes
Wikipedia category vector-based [22], [26]	(1) → (2) → (3) → (4) → (5) → (6) → (7) → (8) → (9)	Yes
Wikipedia link vector-based [30]–[32]	(1) → (5) → (6) → (7) → (9)	Yes
Our Wikipedia link feature model	(1) → (6) → (7) → (11)	No

the treatment processes of four Wikipedia-based semantic similarity systems using the treatment numbers in Table 2, as shown in Table 3.

The statistical results presented in Tables 2 and 3 show that the process of computing semantic similarity using our proposed Wikipedia link feature model is the simplest. The process involves only four treatment steps, which benefits from the fact that the Wikipedia link data involved in the system is already available in advance in the Wikipedia dump as well as the approach that our model directly treats the outlinks of the article into semantic relation-based features rather than a statistics-based link vector. More importantly, our model is cross-language because it does not involve article content filtering, which means that the system we have developed can be reused between various Wikipedia versions in different languages and is easily reproduced by other groups. In contrast, the systems based on Explicit Semantic Analysis (ESA) or Wikipedia category vector require filtering the content of the article when computing the semantic description vector, in which they require a language-dependent morphological analysis algorithm or stem extraction algorithm. Therefore, such systems must provide a separate version for each Wikipedia in different languages. Moreover, ESA and vector-based methods all require a statistical process to compute TF-IDF weight, which is a very time-consuming process. Although category structure-based Wikirelate! does not require computing TF-IDF, its measurement accuracy is far less than our model because Wikipedia's category structure is not a rigorous "*is-a*" taxonomy [47].

V. SIMILARITY MODEL COMBINING WordNet AND WIKIPEDIA

As discussed previously, WordNet has a clear concept hierarchy; hence, most of the popular semantic similarity algorithms are implemented and evaluated by using it as an underlying reference ontology. However, with exponential growth of online information in World Wide Web, the shortcomings of the limited coverage of WordNet began to emerge. Moreover, some semantic deviations inevitably exist in a manual taxonomy such as WordNet, that is, not all locations of concepts in the "*is-a*" hierarchy of WordNet may always be the most appropriate ones compared with the cognitions of people, which may cause some deviations in the similarity measurements based on "*is-a*" relations. For example, the word pair *food* and *fruit* is given a low similarity of approximately 0.1 by existing algorithms based on "*is-a*" relations [2], [12]–[16], [18], whereas the human judgment yields a similarity of 0.77 (normalized) on the MC30 [48] dataset. Wikipedia is an online collaborative knowledge resource that has broad knowledge coverage and contains rich link semantics. The most direct approach for overcoming the above shortcomings of WordNet is to integrate WordNet and Wikipedia.

Although DBPedia, a knowledge graph extracted from Wikipedia, has more structured information than Wikipedia, such as the rich semantic relations and instances of concepts, it also loses some important information in Wikipedia, such as the text and links in concept pages, and disambiguation pages. Moreover, the generation of DBpedia requires additional information extraction techniques and DBpedia cannot be updated as promptly as Wikipedia. And our model is mainly to pursue the powerful word sense disambiguation function in Wikipedia and its real-time update, so in combination with WordNet, we use Wikipedia instead of DBPedia.

According to Tversky's cognitive psychology theory [35], semantic similarity typically reflects the commonality of the properties or components between concepts and can be measured by the "*is-a*" and "*has-part*" relations.

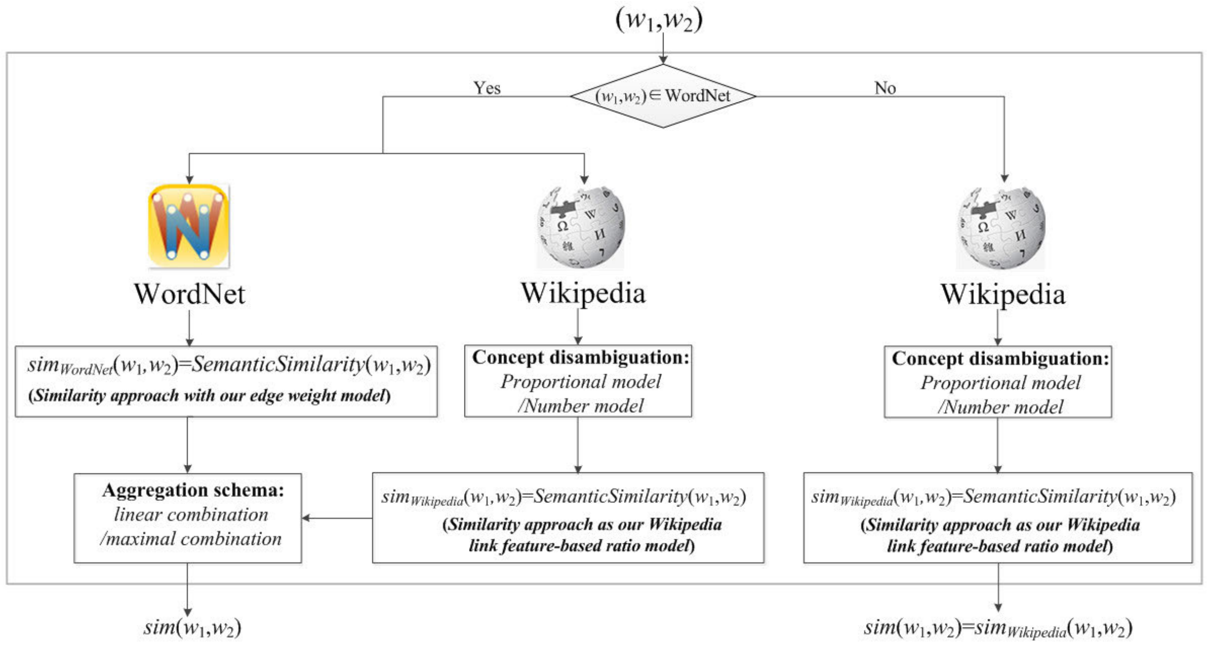

(w_1,w_2)
$(w_1,w_2)\in$ WordNet
Yes
No
WordNet
Wikipedia
Wikipedia
$sim_{WordNet}(w_1,w_2)$=$SemanticSimilarity(w_1,w_2)$
(Similarity approach with our edge weight model)
Concept disambiguation: Proportional model /Number model
Concept disambiguation: Proportional model /Number model
Aggregation schema: linear combination /maximal combination
$sim_{Wikipedia}(w_1,w_2)$=$SemanticSimilarity(w_1,w_2)$
(Similarity approach as our Wikipedia link feature-based ratio model)
$sim_{Wikipedia}(w_1,w_2)$=$SemanticSimilarity(w_1,w_2)$
(Similarity approach as our Wikipedia link feature-based ratio model)
$sim(w_1,w_2)$
$sim(w_1,w_2)=sim_{Wikipedia}(w_1,w_2)$

FIGURE 3. The aggregation architecture diagram combining WordNet and Wikipedia.

The hierarchy in WordNet reflects the "*is-a*" semantic relationship between concepts, and as described in Section IV-A, the outlinks in Wikipedia articles reflect the "*has-part*" relationship between concepts' articles. Based on the above analysis, we propose two aggregation schemas as follows:

First, we propose a linear combination model that represents the weighted average of the "*is-a*"-based computation and the "*has-part*"-based computation. Let $sim_{WN}(w_1, w_2)$ represent a similarity algorithm that uses WordNet as a reference ontology and let $sim_{Wik}(w_1, w_2)$ represent a similarity algorithm that uses Wikipedia as a reference ontology for a word pair (w_1, w_2). We define the semantic similarity computing approach as a linear combination as (35), shown at the bottom of the next page. where α is a smoothing factor, which is used to scale the contributions of $sim_{WN}(w_1, w_2)$ and $sim_{Wik}(w_1, w_2)$ $(\alpha \in [0, 1])$.

Second, we propose a maximal combination model that represents the semantic complement of the "*is-a*"-based computation and the "*has-part*"-based computation. We define the semantic similarity computing approach as a maximal combination as (36), shown at the bottom of the next page.

We propose the maximum complementation model instead of the minimum complementation model based on the observation that the semantic similarity using our WordNet edge weight and the semantic similarity based on wikipedia links are generally lower than the human judgment, so taking their maximum values can achieve performance enhancement.

In the above two aggregation schemas, we regard the similarity result that is obtained via Wikipedia link feature-based ratio model computing for a word pair as the final similarity if a word from the word pair does not exist in WordNet.

To facilitate understanding of the use of the similarity model that combines WordNet and Wikipedia, we present an aggregation architecture diagram as Fig. 3.

VI. EXPERIMENTS

A. KNOWLEDGE SOURCES AND DATASETS

In this paper, we exploit WordNet 3.0[2] as the taxonomic ontology and use the Java WordNet Interface (JWI)[3] to query related data for the experiments in WordNet. Moreover, WordNet is a domain-independent lexical resource and many experiments use domain-specific ontologies. To determine whether the edge weight model has wide coverage over the category graphs of various ontologies, we utilize a domain-specific knowledge source, namely, the SNOMED-CT clinical healthcare terminology,[4] in this study. The version of SNOMED-CT that we use in this study is from July 3, 2017 and we utilize the PyMedTermino[5] module to access SNOMED-CT. The data from Wikipedia that we use in this study are from March 2017 and we utilize the Java Wikipedia Library (JWPL)[6] to obtain the experimental data from Wikipedia.

Several evaluation datasets have been created. In this study, we use the famous Miller and Charles (MC30) [48], Rubenstein and Goodenough (RG65) [49] and Agirre *et al.* (AG203) [50] benchmarks as test beds for WordNet and Wikipedia and exploit the famous Pedersen *et al.* (Pedersen30) [51] benchmark as a test bed for SNOMED-CT. In addition, we have also established a large SimLex666

[2] http://wordnet.princeton.edu/wordnet/download/current-version/
[3] http://projects.csail.mit.edu/jwi/
[4] https://www.nlm.nih.gov/healthit/snomedct/international.html
[5] http://pythonhosted.org/PyMedTermino/index.html
[6] https://dkpro.github.io/dkpro-jwpl/

nominal dataset with 666 noun pairs as our test sets from the SimLex999[7] proposed by Hill *et al.* [52]. These benchmarks have become the de facto standards for evaluating the performances of similarity measures. The dataset of the Miller and Charles metric consists of 30 English noun pairs that were extracted from the original 65 pairs of the Rubenstein and Goodenough metric and the similarity of each pair was judged on a scale from zero (semantically unrelated) to four (highly synonymous) by 38 participants. With the same objective, Agirre *et al.* created a dataset from WordSim353,[8] which contains 203 pairs of terms from WordSim353, each of which has been re-scored according to similarity rather than relatedness. The Pedersen30 dataset consists of 30 pairs of clinical terms. The similarity of each term pair was judged by 9 medical coders and 3 physicians from the Mayo Clinic who were aware of the notion of semantic similarity. Finally, two sets of average values of human judgments are obtained according to the categories (Physician and Coder) of the experts who are involved in the test. The statistics of these datasets are listed in Table 4.

TABLE 4. Datasets used in evaluation of semantic similarity computing task.

Dataset	Year	#Pairs	POS	Scores
RG65	1965	65	Noun	[0, 4]
MC30	1991	30	Noun	[0, 4]
AG203	2009	203	Noun, Verb, Adjective	[0, 10]
SimLex666	2014	666	Noun	[0, 10]
Pedersen30	2007	30	Clinical term	[0, 4]

B. EVALUATION METRICS

Semantic measurements can usually be evaluated by two correlation coefficients: Pearson correlation coefficients and Spearman correlation coefficients. Pearson correlation coefficients mainly reflect how two variables are related in value and are suitable for the evaluation of the semantic similarity [12], [14], [16], [18]–[21], [23], [27], [28], [34], [37], [47], while Spearman correlation coefficients mainly reflect how the two variables are related in rank and are suitable for the evaluation of the semantic relatedness [22], [29], [32], [44], [50] that is a more general notion than semantic similarity and reflects the extent to which concepts co-occur in the context.

This paper focuses on the similarity in semantic measurements, so we used the Pearson correlation coefficient to correlate the scores that were computed via a similarity measure with the judgments that were provided by humans for the above four datasets. Pearson's r is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \bullet \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{37}$$

where x_i refers to i-th element in the list of human judgments; y_i refers to the corresponding i-th element in the list of measure results; $\bar{x}$ and $\bar{y}$ represent the average values of the human judgments and the measure results, respectively, on the dataset; and n is number of word pairs in the dataset.

C. EXPERIMENTS SETTINGS

To ensure the repeatability of the experiments, we describe the experimental and measurement processes as follows: 1)

1) For the measurements on the MC30, RG65, AG203 and SimLex666 datasets, we used each word in each word pairs from each dataset as an index word to query WordNet 3.0 or Wikipedia, and identified all synsets for each word in WordNet 3.0 or all senses for each word in Wikipedia. Then, we used a WordNet-based, Wikipedia-based or their aggregation approach to compute the similarity for each term pair of a word pair according to a similarity approach and used the following formula to compute the similarity $sim(w_1, w_2)$ for each word pair from the MC30, RG65, AG203 and SimLex666 datasets:

$$sim(w_1, w_2) = \max_{(c_1,c_2)\in Term(w_1)\times Term(w_2)} sim(c_1, c_2) \tag{38}$$

where (w_1, w_2) refers to a word pair from the MC30, RG65, AG203 or SimLex666 dataset, (c_1, c_2) refers to a term pair of word pair (w_1, w_2), and $Term(w_1)$ and $Term(w_2)$ are sets of terms that pertain to the taxonomic hierarchy and represent words w_1 and w_2, respectively.

2) For measurements on the Pedersen30 dataset, we obtained the *conceptId* for each term pair in SNOMED-CT and we used a similarity approach to compute their similarity. If multiple similarities were obtained under an environment of multiple inheritances, the maximum similarity was regarded as their final similarity. The average values of categories Physician and Coder of the expert judgments are

$$sim_{Linear}(w_1, w_2) = \begin{cases} sim_{Wik}(w_1, w_2), & if\,(w_1, w_2) \notin WordNet \\ \alpha \times sim_{WN}(w_1, w_2) + (1-\alpha) \times sim_{Wik}(w_1, w_2), & else \end{cases} \tag{35}$$

$$sim_{Max}(w_1, w_2) = \begin{cases} sim_{Wik}(w_1, w_2), & if\,(w_1, w_2) \notin WordNet \\ \max\{sim_{WN}(w_1, w_2), sim_{Wik}(w_1, w_2)\}, & else \end{cases} \tag{36}$$

[7] http://www.cl.cam.ac.uk/ fh295/simlex.html

[8] http://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/

TABLE 5. Pearson coefficients between measures and human judgments on the same datasets in WordNet and SNOMED-CT (The best performances are shown in bold).

Algorithm	Edge/IC model	MC30	RG65	AG203	SimLex666	Pedersen30
IC-based						
Resnik (Eq. (6))	IC computed as Seco	0.81	0.85	0.64	0.55	0.74
Jiang (Eq. (7))	in Eq. (11)(pure IC)	0.84	0.85	0.64	0.49	0.80
Lin (Eq. (8))		0.81	0.82	0.66	0.58	0.79
Resnik (Eq. (6))	IC computed as Sánchez	0.84	0.86	0.63	0.52	0.76
Jiang (Eq. (7))	in Eq. (12)(IC and depth)	0.85	0.87	0.63	0.52	0.81
Lin (Eq. (8))		0.82	0.85	0.63	0.52	0.79
Hybrid						
Gao et al. [47]	Integrated edge and IC information	0.85	0.87	0.63	0.60	0.82
Aouicha et al. [22]	Integrated two soruces (WordNet and Wikipedia)	0.84	0.84	0.68	**0.65**	-
Weighting-based						
Saif et al. [38]		0.85	0.86	**0.71**	0.60	-
Edge-based						
Rada (Eq. (1))	path computed as edge	0.64	0.74	0.54	0.54	0.60
Leacock (Eq. (2))	counting(pure edge)	0.80	0.85	0.65	0.57	0.72
Liu-1 (Eq. (3))		0.80	0.84	0.67	0.57	0.75
Liu-2 (Eq. (4))		0.77	0.84	0.66	0.54	0.75
Li(Eq. (5))		0.80	0.86	0.66	0.56	0.75
Rada (Eq. (1))	path computed as our	0.78	0.84	0.62	0.59	0.67
Leacock (Eq. (2))	edge weight model	0.81	0.86	0.67	0.60	0.82
Liu-1 (Eq. (3))		**0.86**	0.87	0.69	0.62	0.80
Liu-2 (Eq. (4))		**0.86**	**0.88**	0.69	0.59	0.82
Li (Eq. (5))		0.85	**0.88**	0.69	0.63	**0.83**

regarded as the final human judgments for the Pedersen30 dataset.

3) We used the Pearson correlation coefficient to correlate the scores that were computed via a measure with the judgments that were provided by humans via Eq. (37).

D. EXPERIMENTAL RESULTS OF OUR EDGE WEIGHT SIMILARITY MODEL

We evaluate the performance of the proposed edge weight model as follows: First, we compare our edge weight model with the edge-counting-based path computing model using the five edge-based similarity algorithms that are defined in Eqs. (1)-(5) to measure the same datasets. This comparison is used to evaluate the performance of our model in enhancing the measurement accuracy of the edge-based similarity algorithms. Second, we compare our edge weight model with the measurements of the IC-based algorithms in combination with other IC models to evaluate whether the edge-based similarity algorithms in combination with our model can realize excellence performance. We also compare our edge weight model with hybrid methods that integrate edge and IC information [47] or multiple sources [22] to further evaluate the performance of our model.

Table 5 lists the Pearson coefficients of five edge-based measures that are combined with various path models, three IC-based measures that are combined with various IC computations and two hybrid measures on the MC30, RG65, AG203, SimLex666 and Pedersen30 (average of both Physician and Coder) datasets.

E. EFFICIENCY RESULTS OF OUR EDGE WEIGHT SIMILARITY MODEL

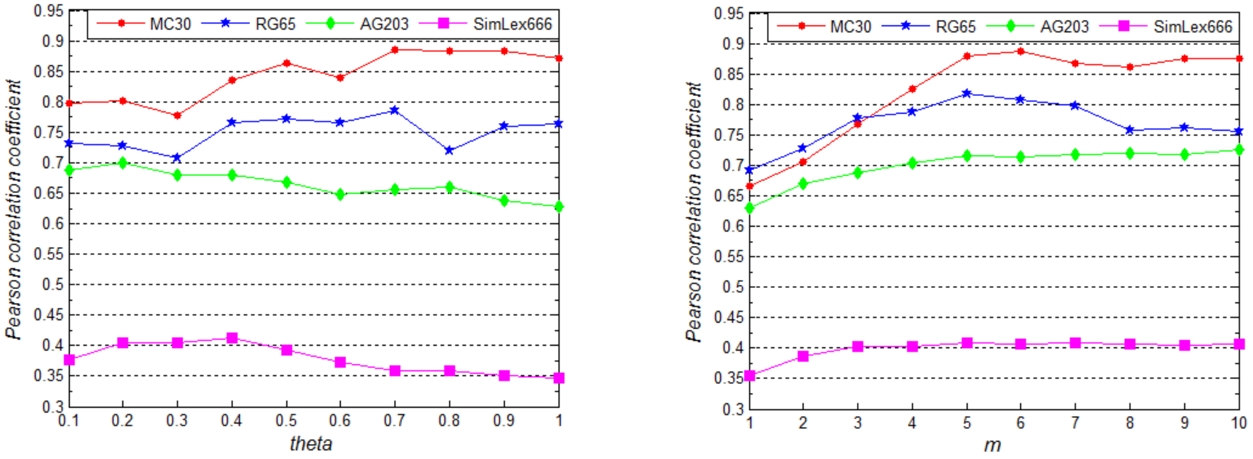
We select two hybrid methods that integrate edge and IC information as shown in Table 6. These two methods were selected for comparison due to their similar principle. Our edge weight model only is a path computing general model and itself cannot compute the similarity. Here we select Liu-1's method as our similarity computing model. Our edge weight path model can be seen as a special path IC, and it has the same paradigm with hybrid method that integrates edge and IC information. Moreover, the development trend of taxonomic ontology is online and real-time updating. To accommodate this trend, we assume that WordNet and SNOMED-CT are real-time dynamic ontologies rather than ontologies that are downloaded in advance. Thus, in hybrid similarity measures, we use following formula to calculate the total time for each measurement:

$$TotalTime = PreprocessingTime + ComputingTime \quad (39)$$

where the subsumption relationship is recursive and *PreprocessingTime* was used to explore the set of all hyponyms for the root node to perfectly characterize the concepts that are specializations of *root*. Finally, we count and store into a hash table the total number of hyponyms for each concept in the ontology. *ComputingTime* is used in the hybrid algorithms to determine the least common subsumer between two concepts and to compute the similarity scores of each word pair on MC30, RG65, AG203 and SimLex666 for WordNet 3.0 or on Pedersen30 for SNOMED-CT (2017)

TABLE 6. Comparison of the edge-based and IC-based measures in terms of efficiency.

Algorithm	Preprocessing	Computing	TotalTime	AverageTime	Dataset
Hybrid measure1	159.572	4.546	164.118	5.471	MC30
((Lin (Eq. (8)) + IC		6.354	165.926	2.553	RG65
computed as Zhou in		17.834	177.406	0.874	AG203
Eq. (13))		44.022	203.594	0.306	SimLex666
	1014.661	26.312	1040.973	34.699	Pedersen30
Hybrid measure2	159.572	4.504	164.076	5.469	MC30
(Gao et al. [47])		6.247	165.819	2.551	RG65
		17.562	177.134	0.873	AG203
		43.923	203.495	0.306	SimLex666
	1014.661	25.807	1040.468	34.682	Pedersen30
Edge-based measure	0	3.686	3.686	0.123	MC30
(path computed as our		5.728	5.728	0.088	RG65
edge weight model)		15.904	15.904	0.078	AG203
		38.734	38.734	0.058	SimLex666
		21.262	21.262	0.709	Pedersen30



(a) Proportional model

(b) Number model

FIGURE 4. Pearson correlation coefficient changes as different θ and m values.

TABLE 7. Computer configuration that is used in the experiment.

Computer type	CPU type	CPU frequency	Memory
Desktop PC	i5-2400	3.1GHz	4GB

according to the hyponym or depth hash table. The experimental results are presented in Table 6. The column entitled *Totaltime* in Table 6 corresponds to the total time for the benchmark, *AverageTime* corresponds to the average time for each word pair, and the units are in seconds. Table 7 describes the computer configuration that is used in our experiment.

F. PARAMETER VALUE FOR OUR WIKIPEDIA DISAMBIGUATION STRATEGY

Since Wikipedia is edited by volunteers, the terms of a concept are highly comprehensive. To reduce the time-consumption, we cannot directly measure the semantic similarity between concepts via Eq. (38); hence, it is necessary to reduce the number of terms of interest for a concept via a word disambiguation strategy. Both of our proposed word disambiguation strategies contain a parameter for which the value must be determined. We plot the Pearson correlation coefficient as a function of θ and m for the MC30, RG65, AG203 and SimLex666 datasets as follows:

According to Fig. 4, the Pearson correlation coefficients are at or near the maximum values with $\theta = 0.5$ for the Proportional model and $m = 5$ for the Number model.

G. EXPERIMENTAL RESULTS OF WIKIPEDIA LINK-BASED SIMILARITY

In this section, we evaluate the performance of the proposed Wikipedia link feature-based ratio model from two aspects: First, we compare our Wikipedia link feature-based similarity model with four word disambiguation strategies in Wikipedia (two strategies from the previous work and two new strategies that are proposed in this paper) to evaluate whether our disambiguation strategies perform effectively. We also compare our Wikipedia link feature-based similarity model that uses our word disambiguation strategies with popular approaches [26], [27], [29], [31], [39], [40] in Wikipedia to further evaluate the performance of our model. These results

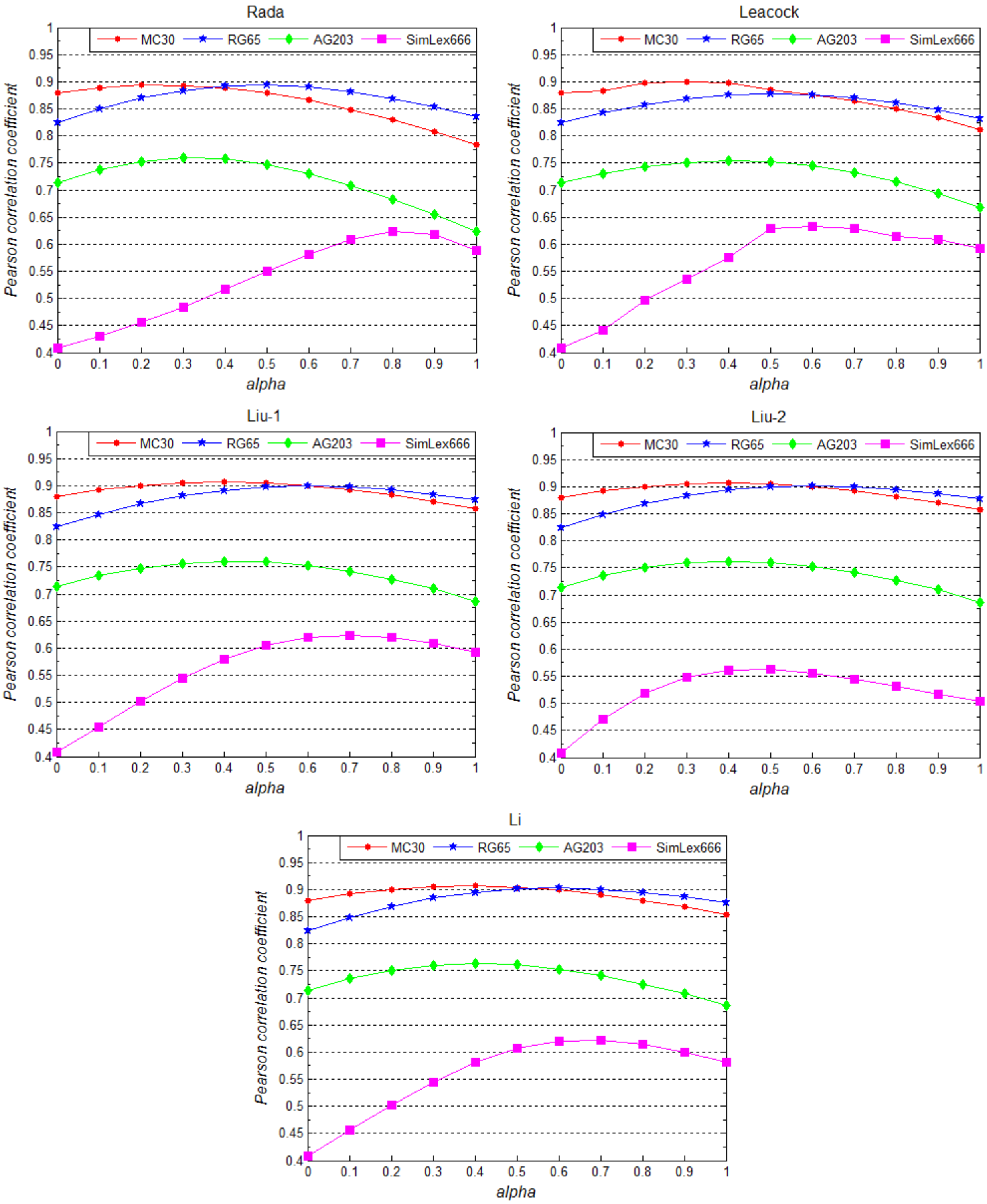


FIGURE 5. Variation of Pearson coefficients with the smoothing factor α.

are presented in Table 8. In the performance comparison experiments, we use the word disambiguation strategy with $\theta = 0.5$ for the Proportional model and $m = 5$ for the Number model.

H. EXPERIMENTAL RESULTS OF PROPOSED AGGREGATION MODELS

We evaluate the performances of the two proposed aggregation schemas from two aspects: First, in Fig. 5, we plot how

TABLE 8. Pearson coefficients comparison of various measures with Wikipedia on the same datasets (The best performances are shown in bold).

Method	Algorithm	MC30	RG65	AG203	SimLex666
Category structure-based	Strube et al. [39]	0.46	0.54	0.49	0.35
	Jiang et al. [27]	0.81	0.66	-	-
ESA-based	Gabrilovich et al. [40]	0.54	0.33	-	0.32
	Li et al. [29]	0.48	0.48	-	-
Category-vector-based	Hadj Taieb et al. [26]	0.83	0.78	-	-
Link vector-based	Milne et al. [31]	0.81	0.69	0.54	0.36
Our link feature model	Single match I	0.85	0.72	0.63	0.38
	Single match II	0.85	0.73	0.65	0.38
	Proposed strategy 1	0.86	0.77	0.67	0.39
	Proposed strategy 2	**0.88**	**0.82**	**0.72**	**0.41**

Note: Except for our model and Link vector-based [31], which is evaluated in this paper, the results of the other methods are quoted in the original paper.

TABLE 9. The best coefficients of various measures using WordNet and Wikipedia on the same datasets.

Algorithm	Knowledge source	MC30	RG65	AG203	SimLex666
Resnik (Eq. (6)) [18]	WordNet	0.84	0.86	0.63	0.52
Jiang (Eq. (7)) [12]	WordNet	0.85	0.87	0.63	0.52
Lin (Eq. (8)) [15]	WordNet	0.82	0.85	0.63	0.52
Gao et al. [47]	WordNet	0.85	0.87	0.63	0.60
Hadj Taieb et al. [26]	Wikipedia 2008	0.83	0.78	-	-
Jiang et al. [27]	Wikipedia 2013	0.81	0.66	-	-
Milne et al. [30]	Wikipedia 2017	0.81	0.69	0.54	0.36
Zhu et al. [32]	Wikipedia 2017	0.87	0.84	0.72	0.44
Qu et al. [11]	Wikipedia 2018	0.83	0.88	0.73	-
Aouicha et al. [22]	WordNet & Wikipedia	0.84	0.84	0.68	0.58
Word2Vec [53]	Google News	0.78	0.76	0.76	0.46
Glove [54]	Wikipedia	0.74	0.74	0.74	0.54
Wikipedia2vec[11]	Wikipedia	0.81	0.76	0.76	0.42
wpath(graph) [45]	WordNet & DBpedia	0.79	0.78	0.62	0.58
Our edge weight model*	WordNet	0.86	0.88	0.69	0.59
Our link feature model**	Wikipedia	0.88	0.82	0.72	0.41
Our aggregation model***	WordNet & Wikipedia	**0.91**	**0.90**	**0.76**	**0.62**

* with Liu-2 (Eq. (4))
** with disambiguation strategy 2
*** maximal aggregation with Leacock's approach

the Pearson correlation coefficient changes with the value of the smoothing factor α in the linear combination schema to obtain the common value of α so that the results that are obtained using our model can be fairly compared with those of other popular approaches. Second, we compare our two aggregation models with other high-performing approaches on WordNet and Wikipedia to evaluate the performances of our models; the results are listed in Tables 9 and 10.

According to Fig. 5, as the value of parameter α is increased, the Pearson's correlation coefficient initially rises and subsequently falls in all cases. To fit all solutions, we use the same parameter value ($\alpha = 0.5$) to compare our linear combination model with other high-performing approaches, for which all the Pearson correlation coefficients are at or near the maximum values in all datasets. The recent trend in semantic similarity computation is to use a word embedding vector based on an artificial neural-network. We also compare our model with three advanced word embedding vector models, called word2vec[9] [53], GloVe[10] [54] and Wikipedia2vec,[11] which use the representation of words as continuous vectors. For word2vec and Wikipedia2vec, we use 300-dimensional pre-trained embeddings. For GloVe, we use 300-dimensional uncased pre-trained embeddings. The cosine distance between vectors is used to calculate their semantic similarity. Since detailed results have been presented in Table 5 and Table 8, we present only the best results from WordNet and Wikipedia in Table 9 and our two aggregation models in Table 10. Table 11 shows the word pairs that are significantly improved by our maximal aggregation model combined with Liu-1's method in AG203, which improves semantic deviations exist in the "*is-a*" hierarchy of WordNet.

VII. DISCUSSIONS

A. DISCUSSION ON WordNet EDGE WEIGHT MODEL

From the experimental results in Tables 5 and 6, we draw several conclusions. First, the results in Table 5 demonstrate that our edge weight model can substantially increase the accuracy of various edge-counting-based similarity measures on both the WordNet and SNOMED-CT taxonomies.

[9] The Word2Vec word embeddings used in the experiments were downloaded at https://code.google.com/archive/p/word2vec/.

[10] The GloVe word embeddings used in the experiments were downloaded at https://nlp.stanford.edu/projects/glove/.

[11] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

TABLE 10. The Pearson coefficients of our two aggregation models using disambiguation strategy 2.

Algorithm		Knowledge source	MC30	RG65	AG203	SimLex666
Our linear aggregation model	With Leacock's approach	WordNet and Wikipedia 2017	0.88	0.88	0.75	0.63
	With Liu-1's approach		0.91	0.90	0.76	0.61
	With Liu-2's approach		0.90	0.90	0.76	0.56
	With Li's approach		0.90	0.90	0.76	0.61
Our maximal aggregation model	With Leacock's approach		0.91	0.90	0.76	0.62
	With Liu-1's approach		0.92	0.91	0.75	0.61
	With Liu-2's approach		0.92	0.92	0.75	0.57
	With Li's approach		0.92	0.92	0.74	0.60

For example, in combination with our edge weight model, five edge-based algorithms are used to obtain a competitive human correlation, especially Liu-2's and Li's algorithms, which performed on the same level as state-of-the-art IC-based measures and hybrid measures on all datasets above and even outperformed them on the RG65, AG203 and Pedersen30 datasets. Our model can substantially improve the accuracies of edge-based methods, which is mainly due to three advantages of our edge weight model in semantic similarity measurements: (1) we use an edge weighting strategy to improve the performance in distinguishing the semantic distances between concepts; (2) we combine an edge counting model and information theory to overcome the irregular density problem of large taxonomies; and (3) our edge weight model can also be regarded as an IC prediction method, in which a concept's IC is predicted by the local density at its location.

Second, in terms of computational efficiency, the results in Table 6 demonstrate that the edge-based method achieves the highest computational efficiency in concept semantic similarity measurement because it does not require any preprocessing; the IC-based method has a moderate computational efficiency because it requires prior counting of all hyponyms of concepts in the taxonomy; and the hybrid method that integrates edge and IC information has the lowest computational efficiency because it must consider the depths of concepts when calculating concepts' information contents. Our edge weight model performs similarly to edge-based methods in terms of computational efficiency because we regard edges as the main information source for concepts and consider only direct hyponyms of the lowest common subsumer between concepts in calculating the density, rather than all hyponyms.

Finally, in comparison with the weight-based method proposed by Saif *et al.* [38], our model combined with Li or Liu-1 is equal to or exceeds it on MC30, surpass it on the RG65 and SimLex666 datasets, but defeat it on AG203. Overall, our model is slightly superior to Saif's method in terms of measurement accuracy. However, our model only calculates the direct hyponyms of the super-concept when considering the density, while Saif's method in Eq. (21) calculates all the hyponyms of the super-concept when considering the density. Therefore, our model has a significant advantage over Saif's method in computational efficiency.

B. DISCUSSION ON WIKIPEDIA LINK FEATURE MODEL

From the results in Fig. 4 and Table 8, we can draw several important conclusions: (1) the overall performance of our link feature model using proposed disambiguation strategy 2 outperforms various existing Wikipedia similarity methods on the four datasets, including category structure-based measures [16], [27], ESA-based measures [29], [40], category vector-based measure [26] and link vector-based measure [31]. More importantly, the excellent performance of our model is achieved with the lowest complexity as analyzed in Section IV-C, which fully demonstrates that our method to convert Wikipedia links into semantic knowledge is reasonable and feasible; (2) under the same disambiguation strategy 2, the Pearson correlation coefficients of our link feature model on the four datasets are significantly larger than those of link vector-based measure proposed by Milne *et al.* [31], which shows that the links manually labeled by volunteers on the Wikipedia page are processed into semantic knowledge more reasonable than processing into TF-IDF weight vectors in the similarity calculations; and (3) in terms of disambiguation strategy comparison, we propose two strategies based on volunteer awareness that are significantly better than the existing two simple matching strategies, in which compared with existing simple matching strategies, our proposed strategy 2 improves the average human correlation of our model by about 10%.

C. DISCUSSION ON PROPOSED AGGREGATION MODELS

The results presented in Tables 9 and 10 show that proposed two similarity aggregation schemas combining WordNet and Wikipedia defeat various state-of-the-art similarity methods on the four datasets, including WordNet-based excellent measures [12], [15], [18], [47], Wikipedia-based excellent measures [11], [26], [27], [30], [32], WordNet-Wikipedia-based measure [22], WordNet-DBpedia-based measure [45] and word embedding vector-based measures [53], [54]. The achievement of these excellent results is mainly due to the following aspects: (1) aggregated WordNet edge weight model and Wikipedia link feature model perform well in both computational efficiency and measurement accuracy, which have been revealed in Tables 3, 5, 6 and 8, respectively; (2) proposed disambiguation strategy based on volunteer awareness is simple and feasible, and significantly improves the measurement accuracy of our Wikipedia link feature model;

TABLE 11. Word pairs significantly improved by our maximal aggregation model with Liu-1's method in AG203 ("-" indicates that the corresponding word pair does not exist in WordNet).

Word1	Word2	Human (normalized)	Liu-1 (only WordNet)	Our linear aggregation model with Liu-1	Our maximal aggregation model with Liu-1
announcement	news	0.756	0.175	0.444	0.712
food	fruit	0.752	0.076	0.401	0.727
opera	performance	0.688	0.030	0.233	0.440
train	car	0.631	0.433	0.376	0.411
bread	butter	0.619	0.406	0.408	0.825
king	rook	0.592	0.825	0.485	0.650
bishop	rabbi	0.669	0.650	0.449	0.372
glass	metal	0.556	0.104	0.238	0.372
space	chemistry	0.488	0.052	0.239	0.425
drink	car	0.304	0.022	0.152	0.282
listing	proximity	0.256	0.046	0.140	0.234
century	year	0.759	0.233	0.516	0.799
five	month	0.338	0.077	0.206	0.336
coast	forest	0.315	0.090	0.204	0.317
chance	credibility	0.388	0.076	0.197	0.318
seven	series	0.356	0.055	0.222	0.389
experience	music	0.347	0.110	0.206	0.302
Wednesday	news	0.222	0.024	0.125	0.226
month	hotel	0.181	0.000	0.062	0.124
morality	marriage	0.369	0.083	0.205	0.326
atmosphere	landscape	0.369	0.152	0.260	0.368
media	radio	0.742	-	0.774	0.774
Harvard	Yale	0.813	-	0.670	0.670
Arafat	Jackson	0.250	-	0.062	0.062
drink	eat	0.687	-	0.765	0.765
Mexico	Brazil	0.744	-	0.512	0.512
street	children	0.494	-	0.066	0.066
Mars	water	0.294	-	0.459	0.459
media	gain	0.288	-	0.135	0.135
media	trading	0.388	-	0.312	0.312
stock	live	0.373	-	0.201	0.201

and (3) our aggregation model effectively breaks through the ceiling of measurement accuracy based on a single WordNet or Wikipedia by integrating the "*is-a*" taxonomy in WordNet and the link feature in Wikipedia.

The results presented in Tables 9 also show that our edge weight model only slightly improves the methods [12], [15], [18], [47] of information content in WordNet, our link model model only slightly improves the Wikipedia-based methods [11], [26], [27], [30], [32], but these improvements are significant because their advantages are achieved with significant efficiency gains, seen more details in Table6 for our edge weight model, Table 2 and Table 3 for our link model. More importantly, our aggregation model, which combines the proposed WordNet edge weight and Wikipedia link models, significantly surpasses all other methods in four datasets. Especially on the small datasets MC30 and RG65 composed of common word pairs, the best human correlation of our similarity aggregation models reaches 0.92 and has exceeded the average correlation (0.9015) between individual human subjects reported in Resnik's replication [18] of the Miller and Charles experiment, which fully demonstrates that our aggregation model makes the potential of WordNet and Wikipedia in similar calculations reach the limit.

The results presented in Table 10 also show that the two aggregation models proposed by us have the same performance on the four datasets as a whole, in which our maximal aggregation model performs better on the small datasets MC30 and RG65 and our linear aggregation model are more stable on the big datasets AG203 and SimLex666. Moreover, Table 11 gives some examples to show how our aggregation model improves the similarity method based on WordNet taxonomy through Wikipedia link features. First, our aggregation model expands the word coverage of the similarity measures in WordNet. For example, in the end of Table 11, our aggregation model implements the measurement of 10 word pairs that do not exist in WordNet. Second, our aggregation model improves semantic deviations existing in WordNet taxonomy by integrating Wikipedia link features. For example, Table 11 shows that our aggregation model significantly narrows the gap between the measured value and the human value in the similarity measurement of 20 word pairs. Furthermore, there are some word pairs, such as "train" & "car" and "Arafat" & "Jackson", whose measurement cannot be improved by our aggregation model. This means that our aggregation model needs to be integrated with other knowledge sources. Finally, although the maximum aggregation model works better than the linear aggregation model in most cases, there are also opposite situations such as in the measurement of the "king" & "rook" pair.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we propose an edge weight model for overcoming the density non-uniformity of edge-based measures. Our model can adapt to variations in the density of edges without requiring an additional parameter and has wide coverage over various edge-based measures on multiple ontologies. Then, we propose a Wikipedia link feature-based ratio model and two word disambiguation strategies. This model ignores Wikipedia's extensive textual content and is highly efficient, and these disambiguation strategies are based on volunteer awareness and can improve computing accuracy. Finally, we propose two aggregation models for further improving the computing accuracy. The results of extensive experiments demonstrate that our model realizes high performance, high efficiency and high coverage, and has substantial application prospects in various application fields. In the future, we are planning to introduce Support Vector Machine (SVM) in our model to determine the best application scenarios for different aggregation models, and to further combine our model with the DBpedia knowledge graph to obtain more semantic evidence.

REFERENCES

[1] I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach," *Expert Syst. Appl.*, vol. 44, pp. 386–399, Feb. 2016, doi: 10.1016/j.eswa.2015.09.028.

[2] F. M. Anuar, R. Setchi, and Y.-K. Lai, "Semantic retrieval of trademarks based on conceptual similarity," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 2, pp. 220–233, Feb. 2016, doi: 10.1109/TSMC.2015.2421878.

[3] A. Otegi, X. Arregi, O. Ansa, and E. Agirre, "Using knowledge-based relatedness for information retrieval," *Knowl. Inf. Syst.*, vol. 44, no. 3, pp. 689–718, Sep. 2015, doi: 10.1007/s10115-014-0785-4.

[4] D. O. , S. Kwon, K. Kim, and Y. Ko, "Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph," in *Proc. 27th Int. Conf. Comput. Linguistics, COLING*, Santa Fe, NM, USA, Aug. 2018, pp. 2704–2714. [Online]. Available: https://aclanthology.info/papers/C18-1229/c18-1229

[5] G. Zhu and C. A. Iglesias, "Exploiting semantic similarity for named entity disambiguation in knowledge graphs," *Expert Syst. Appl.*, vol. 101, pp. 8–24, Jul. 2018, doi: 10.1016/j.eswa.2018.02.011.

[6] Q. Liu, B. Liu, Y. Zhang, D. S. Kim, and Z. Gao, "Improving opinion aspect extraction using semantic similarity and aspect associations," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 2986–2992. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11973

[7] C. Ru, J. Tang, S. Li, S. Xie, and T. Wang, "Using semantic similarity to reduce wrong labels in distant supervision for relation extraction," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 593–608, Jul. 2018, doi: 10.1016/j.ipm.2018.04.002.

[8] F. Pech, A. Martinez, H. Estrada, and Y. Hernandez, "Semantic annotation of unstructured documents using concepts similarity," *Sci. Program.*, vol. 2017, Jan. 2017, Art. no. 7831897, doi: 10.1155/2017/7831897.

[9] M. A. H. Taieb, M. B. Aouicha, and Y. Bourouis, "FM3S: Features-based measure of sentences semantic similarity," in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.*, Bilbao, Spain, Jun. 2015, pp. 515–529, doi: 10.1007/978-3-319-19644-2_43.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[11] R. Qu, Y. Fang, W. Bai, and Y. Jiang, "Computing semantic similarity based on novel models of semantic representation using wikipedia," *Inf. Process. Manage.*, vol. 54, no. 6, pp. 1002–1021, Nov. 2018, doi: 10.1016/j.ipm.2018.07.002.

[12] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Res. Comput. Linguistics Int. Conf., ROCLING*, Taipei, Taiwan, Aug. 1997, pp. 19–33. [Online]. Available: https://aclanthology.info/papers/O97-1002/o97-1002

[13] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," *WordNet, Electron. Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.

[14] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, Jul. 2003, doi: 10.1109/TKDE.2003.1209005.

[15] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn. ICML*, Madison, WI, USA, Jul. 1998, pp. 296–304.

[16] X.-Y. Liu, Y.-M. Zhou, and R.-S. Zheng, "Measuring semantic similarity in wordnet," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Aug. 2007, pp. 3431–3435.

[17] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 17–30, Jan. 1989, doi: 10.1109/21.24528.

[18] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell., IJCAI*, Montréal QC, Canada, Aug. 1995, pp. 448–453. [Online]. Available: http://ijcai.org/Proceedings/95-1/Papers/059.pdf

[19] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," in *Proc. 16th Eureopean Conf. Artif. Intell., ECAI*, Valencia, Spain, Aug. 2004, pp. 1089–1090.

[20] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "Ontology-based approach for measuring semantic similarity," *Eng. Appl. Artif. Intell.*, vol. 36, pp. 238–261, Nov. 2014, doi: 10.1016/j.engappai.2014.07.015.

[21] X. Zhu, F. Li, H. Chen, and Q. Peng, "An efficient path computing model for measuring semantic similarity using edge and density," *Knowl. Inf. Syst.*, vol. 55, no. 1, pp. 79–111, Apr. 2018, doi: 10.1007/s10115-017-1078-5.

[22] M. B. Aouicha, M. A. Hadj Taieb, and A. B. Hamadou, "Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness," *Int. J. Speech Technol.*, vol. 45, no. 2, pp. 475–511, Sep. 2016, doi: 10.1007/s10489-015-0755-x.

[23] L. Meng, J. Gu, and Z. Zhou, "A new model of information content based on concept's topology for measuring semantic similarity in wordnet," *Int. J. Grid Distrib. Comput.*, vol. 5, no. 3, pp. 81–94, 2013.

[24] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *J. Artif. Intell. Res.*, vol. 34, pp. 443–498, Mar. 2009, doi: 10.1613/jair.2669.

[25] M. A. H. Taieb, M. B. Aouicha, M. Tmar, and A. B. Hamadou, "Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring," in *Proc. Int. Conf. Data Knowl. Eng.*, Fujian, China, Nov. 2012, pp. 128–140, doi: 10.1007/978-3-642-34679-8_13.

[26] M. A. Hadj Taieb, M. Ben Aouicha, and A. Ben Hamadou, "Computing semantic relatedness using wikipedia features," *Knowl.-Based Syst.*, vol. 50, pp. 260–278, Sep. 2013, doi: 10.1016/j.knosys.2013.06.015.

[27] Y. Jiang, W. Bai, X. Zhang, and J. Hu, "Wikipedia-based information content and semantic similarity computation," *Inf. Process. Manage.*, vol. 53, no. 1, pp. 248–265, Jan. 2017, doi: 10.1016/j.ipm.2016.09.001.

[28] Y. Jiang, X. Zhang, Y. Tang, and R. Nie, "Feature-based approaches to semantic similarity assessment of concepts using wikipedia," *Inf. Process. Manage.*, vol. 51, no. 3, pp. 215–234, May 2015, doi: 10.1016/j.ipm.2015.01.001.

[29] P. Li, B. Xiao, W. Ma, Y. Jiang, and Z. Zhang, "A graph-based semantic relatedness assessment method combining wikipedia features," *Eng. Appl. Artif. Intell.*, vol. 65, pp. 268–281, Oct. 2017, doi: 10.1016/j.engappai.2017.07.027.

[30] D. Milne, "Computing semantic relatedness using Wikipedia link structure," in *Proc. new zealand Comput. Sci. Res. Student Conf.*, 2007, pp. 63–70.

[31] D. N. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proc. AAAI Workshop Wikipedia Artif. Intell., Evolving Synergy*, 2008, pp. 25–30.

[32] X. Zhu, Q. Guo, B. Zhang, and F. Li, "An efficient approach for measuring semantic relatedness using wikipedia bidirectional links," *Appl. Intell.*, vol. 49, no. 10, pp. 3708–3730, 2019.

[33] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, Jul. 2012, doi: 10.1016/j.eswa.2012.01.082.

[34] Z. Zhou, Y. Wang, and J. Gu, "A new model of information content for semantic similarity in WordNet," in *Proc. 2nd Int. Conf. Future Gener. Commun. Netw. Symposia*, Dec. 2008, pp. 85–89.

[35] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, p. 327, Jul. 1977.

[36] M. A. Rodriguez and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 442–456, Mar. 2003, doi: 10.1109/TKDE.2003.1185844.

[37] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, "X-similarity: Computing semantic similarity between concepts from different ontologies," *JDIM*, vol. 4, no. 4, pp. 233–237, 2006. [Online]. Available: http://www.dirf.org/jdim/abstractv4i4.htm#v4n4a5

[38] A. Saif, U. Z. Zainodin, N. Omar, and A. S. Ghareb, "Weighting-based semantic similarity measure based on topological parameters in semantic taxonomy," *Natural Lang. Eng.*, vol. 24, no. 6, pp. 861–886, Nov. 2018.

[39] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *Proc. 21st Nat. Conf. Artif. Intell. 18th Innov. Appl. Artif. Intell. Conf.*, Boston, MA, USA, Jul. 2006, pp. 1419–1424. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-223.php

[40] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. 20th Int. Joint Conf. Artif. Intell., IJCAI*, Hyderabad, India, Jul. 2007, pp. 1606–1611. [Online]. Available: http://ijcai.org/Proceedings/07/Papers/259.pdf

[41] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," in *Proc. 20th Int. Conf. World Wide Web WWW*, 2011, pp. 337–346, doi: 10.1145/1963405.1963455.

[42] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, "Wikiwalk: Random walks on wikipedia for semantic relatedness," in *Proc. Workshop Graph-Based Methods Natural Lang. Process.*, Aug. 2009, pp. 41–49. [Online]. Available: http://www.aclweb.org/anthology/W09-3206

[43] M. T. Pilehvar and R. Navigli, "From senses to texts: An all-in-one graph-based approach for measuring semantic similarity," *Artif. Intell.*, vol. 228, pp. 95–128, Nov. 2015, doi: 10.1016/j.artint.2015.07.005.

[44] M. B. Aouicha, M. A. H. Taieb, and A. B. Hamadou, "LWCR: Multi-layered wikipedia representation for computing word relatedness," *Neurocomputing*, vol. 216, pp. 816–843, Dec. 2016, doi: 10.1016/j.neucom.2016.08.045.

[45] G. Zhu and C. A. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 72–85, Jan. 2017.

[46] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[47] J.-B. Gao, B.-W. Zhang, and X.-H. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory," *Eng. Appl. Artif. Intell.*, vol. 39, pp. 80–88, Mar. 2015, doi: 10.1016/j.engappai.2014.11.009.

[48] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Lang. Cognit. Processes*, vol. 6, no. 1, pp. 1–28, Jan. 1991.

[49] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965, doi: 10.1145/365628.365657.

[50] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics NAACL*, 2009, pp. 19–27. [Online]. Available: http://www.aclweb.org/anthology/N09-1003

[51] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *J. Biomed. Informat.*, vol. 40, no. 3, pp. 288–299, Jun. 2007, doi: 10.1016/j.jbi.2006.06.004.

[52] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (genuine) similarity estimation," *Comput. Linguistics*, vol. 41, no. 4, pp. 665–695, 2014, doi: 10.1162/COLI_a_00237.

[53] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst., 27th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-a%nd-phrases-and-their-compositionality

[54] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empiri Cal Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1162.pdf

FEI LI is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include natural language processing and information extraction.

LEJIAN LIAO received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 1994. He is currently a Professor with the School of Computer Science and Technology, Beijing Institute of Technology. He has published numerous articles in several areas of computer science. His research interests include machine learning, natural language processing, and intelligent networks.

LANFANG ZHANG is currently a Professor with the Faculty of Education, Guangxi Normal University, China. His research interests include natural language processing, information extraction, and intelligent assistant teaching systems.

XINHUA ZHU is currently a Professor with the School of Computer Science and Information Engineering, Guangxi Normal University, China. He is also the Principal Investigator for several National Natural Science Foundation Projects. His research interests include natural language processing, semantic computing, and intelligent assistant teaching systems.

BO ZHANG received the M.S. degree in computer application from Guangxi Normal University, Guilin, China, in 2010. He is currently an Associate Professor with Hezhou University. His research interests include natural language processing and distance education technology.

ZHENG WANG received the bachelor's degree from Shandong University, in 2016, and the master's degree from The University of Hong Kong, in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU) under the supervision of C. Long and G. Cong. His research interests include data mining, database, and deep learning.