# A New Hybrid Improved Method for Measuring Concept Semantic Similarity in WordNet

Xiaogang Zhang, Shouqian Sun, and Kejun Zhang
College of Computer Science and Technology, Zhejiang University, Hangzhou, China

**Abstract:** *Computing semantic similarity between concepts is an important issue in natural language processing, artificial intelligence, information retrieval and knowledge management. The measure of computing concept similarity is a fundament of semantic computation. In this paper, we analyze typical semantic similarity measures and note Wu and Palmer's measure which does not distinguish the similarities between nodes from a node to different nodes of the same level. Then, we synthesize the advantages of measure of path-based and IC-based, and propose a new hybrid method for measuring semantic similarity. By testing on a fragment of WordNet hierarchical tree, the results demonstrate the proposed method accurately distinguishes the similarities between nodes from a node to different nodes of the same level and overcome the shortcoming of the Wu and Palmer's measure.*

**Keywords:** *Information content, Semantic similarity, WordNet taxonomy, Hyponym.*

*Received May 25, 2017; accepted April 25, 2018*
*https://doi.org/10.34028/iajit/17/4/1*

## 1. Introduction

Finding similarity between concepts is an important issue in many applications (e.g., natural language processing, artificial intelligence, information retrieval and knowledge management) [7, 21]. As the minimum unit of describing information and the basis for information resource matching, concept owned the linguistic independence and uniqueness in ontology, which was used to eliminate the polysemy and synonym for estimating textual semantic similarity [1, 24].

The measures of computing semantic similarity between concepts have been divided into two categories. One is based on statistical information of context, and the other is based on ontology. WordNet and the Wikipedia Category Graph (WCG) are both reference ontologies in computing concept semantic similarity [18]. WordNet is a common ontology developed by cognitive science laboratory of Princeton University, which has been used to describe concepts and their semantic relationships. As WordNet is versatility and owns rational semantic organizational form, it is widely used for word sense tagging, information extraction, text proofreading, knowledge reasoning and conceptual modelling tasks [10]. The WCG is the other resource in some research works, including in works of Aouicha *et al*. [2] and Zesch [26]. The WCG is different from WordNet owing to WCG is proposed by volunteers [22], and the categories of WCG do not include specifying the type in semantic relation [19]. In this paper, the researches for finding the method of concept semantic Similarity are based on WordNet ontology. This paper organization shows as follows. In section 2, we analyze some Information Content (IC) computing models and classical semantic similarity measures. In section 3, we study the existing problem of Wu and Palmer's [25] measure and propose a new method for measuring the concept semantic similarity. In section 4, we evaluate the proposed method in a given fragment of WordNet classification tree, then discuss and compare the results of the proposed method and Wu and Palmer's [25] method. In section 5, we summarize this paper and make a plan for future works.

## 2. Related Work

Representative measures for estimating semantic similarity between concepts included IC-based measures, path-based measures, feature-based measures and hybrid measures. The measure of IC-based computed concept similarity by examining the information content contained in the word pairs [11]. The measure of path-based computed concept similarity by the path semantic distance (the number of edges linking two concepts) between words, and then transformed the distance into similarity value [14]. The measure of feature-based estimated the semantic similarity between words according to the structural feature of taxonomy, which included nodes and edges [19]. The hybrid measure computed similarity between words by merging the advantages of other measures conceived [27]. The measures of based-path and based-IC are very important methods for measuring semantic similarity between concepts [4]. In which the measure of IC-based is the best measure among all proposed ones owing to the accurate similarity value and more effective than others. The key of based-IC measure is to compute the value of IC [13]. In WordNet, the taxonomy "is-a" is mainly used to measure the degree of similarity between concepts or words, which account for 70% in all of relationship [5]. In this paper, all methods are all on the basis of "is_a" relationship in WordNet taxonomy. The list of symbols used in this paper is shown in following Table 1.

Table 1. The list of symbols in similarity computing.

| | |
|---|---|
| **p(c)** | The probability which concept c appears in a given corpus. |
| **IC(c)** | The information content of concept c. |
| **hypo(c)** | The count of child nodes belonging to c. |
| **max_nodes** | The maximum number of the concepts in the classification tree. |
| **depth(c)** | The depth of concept c. |
| **max_depth(c)** | The maximum depth of the classification tree of including c. |
| **len($c_1$,$c_2$)** | The shortest path distance between $c_1$ and $c_2$(including itself). |
| **lso($c_1$,$c_2$)** | The deepest common parent of $c_1$ and $c_2$. |
| **subsumers(c)** | The number of nodes from the root to node c along the path of taxonomy. |
| **hypo(lso($c_1$, $c_2$) )** | The hyponym of the most specific common hypernym of node-pair$c_1$ and $c_2$. |
| **depth(lso($c_1$, $c_2$))** | The depth of the most specific common hypernym of node-pair $c_1$ and $c_2$. |

## 2.1. Typical IC Models

Computing IC is the core part of the semantic similarity measure, and it is usually divided into two categories according to different calculation object. One is based on statistical information, and the other is based on taxonomical structure.

### 2.1.1. IC Model Based on the Statistical Information

This kind of model calculated the IC value by counting the probability of a concept in a given corpus. Resnik [16] put forward a method that the probability of concept equalled the frequency of noun appearing in the Brown Corpus [3]. Resnik [16] used negative log likelihood to calculate IC. The corresponding calculating equation is as follows [16]:

$$IC(c) = -\,log\big(p(c)\big) \qquad (1)$$

Here, *c* is a concept node. The Equation (1) indicates that the frequency of a concept appears higher, the message transfers less. Each term appeared in the corpus is counted as an occurrence rate. Then, the function *Freq(c)* was computed as follows [16]:

$$Freq(c) = \sum_{\omega \in Word(c)} Count(\omega) \qquad (2)$$

Here, Word*(c)* is a set of words subsumed by *c*, and *Count(ω)* represents the frequency of the word ω in the given corpus. Where the function p(*c*) could be computed as follows [16]:

$$p(c) = \frac{Freq(c)}{N} \qquad (3)$$

Here, *N* is the total number of nouns appeared.

### 2.1.2. IC Model Based on Taxonomical Structure

Seco *et al*. [20] are the first one computing IC with ontology hierarchical structure. They discovered IC only related to with the taxonomical structure. If a concept includes more child nodes, the IC of this concept is fewer and the *IC* of its leaf node is larger. The IC of a concept only relied on the number of concepts which it subsumes. The Seco'smethod of calculating IC was as follows [20]:

$$IC(c) = 1 - \frac{log(|hypo(c)|+1)}{log(max\_nodes)} \qquad (4)$$

This method relies on the internal structure to calculate the IC value regardless of external information. But this method requires a precondition, which the taxonomical structure of ontology has been organized with a meaningful way.

Zhou *et al*. [28] introduced relative depth on hyponym, and proposed a new method to calculate IC values. They proposed a new formula as follows:

$$IC_{zhou}(c) = K\left(1 - \frac{log\big(hypo(c)\big)}{log(max\_nodes)}\right) + (1-K)\frac{log(depth(c))}{log(max\_depth(c))} \qquad (5)$$

But in Equation (5), the K has to be determined by the specific experiment debugging.

Later, Sanchez [18] proposed a new model, adopting subsumers of leaf node to calculate the value of IC. The equation was as follows:

$$IC_{David}(c) = -\,log\left(\frac{commonness(c)}{commonness(root)}\right) \qquad (6)$$

Where the function *commonness(c)* equals the sum of *commonness(n)*, and *commonness(n)* equals 1/subsumers(n). Wherein n is a leaf node and one of hyponym of node c.

In general, the method based statistical information is high efficiency and fits for large-scale data, but this method is low accuracy because it is subject to external interference. The method based on ontology structure is higher accuracy because this method only relied on the structure itself [23].

## 2.2. Typical Measures of Conceptsimilarity

Many semantic similarity methods have been proposed in last years. Wherein we focused on the measure based path and depth, IC-based measure. Typical measures include Rada *et al*. [15], Wu and Palmer [25], Leacock and Chodorow [8], Resnik [16], Jiang and Conrath [6] and Lin measure [9].

### 2.2.1. The Measure Based Path and Depth

Rada *et al*. [15] stated that the length of the minimum path of two concepts quantified their semantic distance. Namely, the similarity between words can be calculated by the minimum path distance linking their corresponding nodes. A simple measure to calculate their semantic distance defined by Equation (7) isas follows [15]:

$$dis_{rad}(c_1,c_2) = min_{\forall i}|path_i(c_1,c_2)| \quad (7)$$

Wu and Palmer's [25] measure is a typical method based on the shortest path. They thought the similarity between the concepts is smaller if the position of two concepts is lower in the classification tree. The equation of corresponding calculating is as follows:

$$sim_{W\&P}(c_1,c_2) = \frac{2\times depth(lso(c_1,c_2))}{len(c_1,c_2)+2\times depth(lso(c_1,c_2))} \quad (8)$$

Later, Leacock and Chodorow proposed a non-linear calculating model, which included two parameterslen (*c*₁, *c*₂) and max_*depth(c)*. The calculating equation is as follows [8]:

$$sim_{L\&C}(c_1,c_2) = -\log\frac{len(c_1,c_2)}{2\times max\_depth_{c\in WordNet}(c)} \quad (9)$$

We can see from the Equation (9), to a fixed classification tree, if the path distance between two concepts was further, the semantic similarity was smaller.

### 2.2.2. IC-Based Semantic Similarity Measure

Resnik [16] is the first one that introduced ontology to compute the similarity, namely using negative log likelihood to calculate IC. He evaluated the similarity of two concepts by the content of common part, so he considered the most specific common abstraction of $c_1$ and $c_2$ as semantic similarity of two concepts. The proposed model is as follows [16]:

$$sim_R(c_1,c_2) = -\,log\,p\big(lso(c_1,c_2)\big) = IC\big(lso(c_1,c_2) \quad (10)$$

Jiang and Conrath [6] used the concept of information content yet, and they made an improvement to the Equation (10). They took into account the greatest meaning of the word. The calculating equation is as follows [6]:

$$\mathrm{sim}(w_1,w_2) = max_{(i,j)}(sim(s_{1i},s_{2j})) \quad (11)$$

Here, $s_{1i}$ and $s_{2j}$ are the significance of $w_1$ and $w_2$ (the concept of ontology). Jiang and Conrath [6] computed the semantic distance through the IC sum of two concepts subtracting the IC of the most specific common abstraction. The measure equation is as follows [6]:

$$dist_{JC}(c_1,c_2) = IC(c_1) + IC(c_2) - 2\times IC\big(lso(c_1,c_2) \quad (12)$$

After a linear transformation, the equation of measuring semantic similarity is as follows [20]:

$$sim_{J\&C}(c_1,c_2) = 1 - \left(\frac{IC(c_1) + IC(c_2) - 2\times IC\big(lso(c_1,c_2)\big)}{2}\right) \quad (13)$$

Later, Lin believed that the similarity of two concepts should be measured by the ratio of common information and total information. The core of Lin's method was computing the commonality of two concepts. Lin's [9] equation is as follows:

$$sim_{Lin}(c_1,c_2) = \frac{2\times(-\,log\,p(c_0))}{-\,log\,p(c_1) - log\,p(c_2)} = \frac{2\times IC\big(lso(c_1,c_2)\big)}{IC(c_1)+IC(c_2)} \quad (14)$$

In a taxonomical tree, if $c_1\in C_1$ and $c_2\in C_2$, the commonality $c_0$ could expressed as $c_1\in C_0\cap c_2\in C_0$, where $C_0$ is the most specific class that subsumes both $C_1$ and $C_2$.

Based on stated above, it is noted that Rada's *et al.* [15] Leacock and Chodorow's [8] and Wu and Palmer's [25] methods took into account the shortest distance between concepts and the depth of common parent nodes, and the Resnik's [16], Jiang and Conrath's [6] and Lin's [9] measures regarded the IC value of the parent of two concepts as similarity of two concepts.

## 3. A New Improved Method for Measuring Concept Semantic Similarity

As discussed in section 2, the most critical issue which compute concept semantic similarity by the IC is how to get the accurate IC and introduce IC into the similarity measure.

### 3.1. The Shortcoming of Wu And Palmer's Measure

As was made clear in Equation (8), Wu and Palmer's [25] measure used the IC model of Equation (1), $\mathrm{IC(c)} = -\log(\mathrm{p(c)})$. The accuracy of Wu and Palmer's method was unsatisfactory [18]. We extend the method by introducing the Seco's IC model and propose a new improved method for measuring semantic similarity the following illustration is an example.

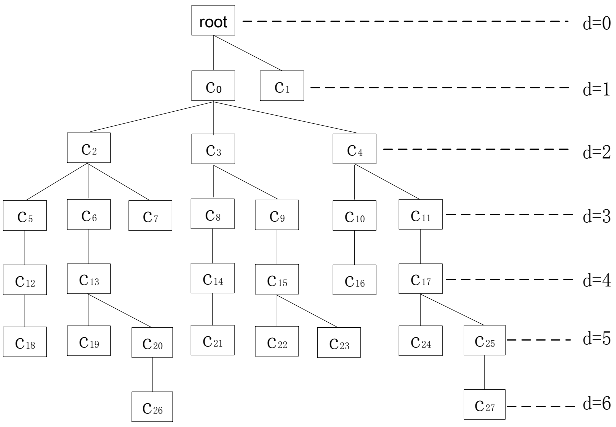

Figure 1. A fragment of "is-a" hierarchical taxonomy in WordNet.

As showed in Figure 1, supposing the root is in 0 level, so {$c_0$, $c_1$} are in 1 level, {$c_2$, $c_3$, $c_4$} are in 2 level, .... Using the Equation (8) of Wu and Palmer's measure, the similarity value was equal between nodes from a node to different nodes of the same level. For instance, $c_7$ and $c_{11}$ lie in the same level, $\mathrm{sim}_{W\&P}(c_7,c_3)$ equals $\mathrm{sim}_{W\&P}(c_{11},c_3)$, the value is 0.4; $c_{16}$ and $c_{17}$ lie in the same level, $\mathrm{sim}_{W\&P}(c_{16},c_{18})$equals $\mathrm{sim}_{W\&P}(c_{17},c_{18})$,the value equals 0.2222.

Obviously, Wu and Palmer's measure failed to distinguish the similarities between nodes from a node to different nodes of the same level, so the precision is unsatisfactory. In this paper, we will propose a new similarity measure to overcome the problem of Wu and Palmer's [25] measure.

### 3.2. Measure Improvement

The studies of Seco *et al*. [20] and Zhou *et al*. [28] indicated that IC is related to the degree of abstraction. In detail, a word with lower abstraction owned higher IC and vice versa. The parameter hyponym is significant to discriminate specific concepts because the set of hyponyms of a concept subsumed a great number of nodes, including direct and indirect descendent [21]. So in this paper we make use of the parameter hyponym to measure the abstraction of words.

Before proposing our similarity method, we define three definitions for semantic computation as follows.

- *Definition* 1: (Hyponym) defines the concept of$hypo(c) = \{c_i \in V, c_i < c\}$. $c_i$ is the descendent of c, V is a set of concepts of the classification tree.
- *Definition* 2: (The most specific common abstraction) defines the concept of *lso(c$_1$,c$_2$)*, which represents the most specific common hypernym of $c_1$ and $c_2$.
- *Definition* 3: (The maximum number of the concepts) define the concept of *max_nodes*, which represents the maximum number of concepts existed in the taxonomy.

Thus, we propose an improved hybrid method for measuring semantic similarity as follows:

$$sim_{new}(c_1,c_2) = \frac{2 \times log(max\_nodes) - log(hypo(lso(c_1,c_2)) + 1)}{2 \times log(max\_nodes) - log(hypo(c_1)+1) - log(hypo(c_2)+1)} \quad (15)$$

In Equation (15), hyponym are the most specific common abstraction of $c_1$ and $c_2$ The function hypo(lso($c_1$,$c_2$)) represents the number of hyponyms, which was used to discriminate the specificity of each concept because of number of offspring.

## 4. Results and Discussion

In this section, in order to confirm the effect of the proposed method, we design an experiment to distinguish the similarity between nodes form a node to different nodes of the same level. We take the 10 nodes $\{c_2,c_3,\ldots,c_{11}\}$ of Figure 1 as an example to compute the semantic similarity.

### 4.1. Experiment Results

The experiment results of Figure 1 are listed in Table 2.

Table 2. Similarity comparison of Wu and Palmer and the proposed measure in figure.

| Number | Node-pair | Depth1 | Depth2 | Len | Hypo1 | Hypo2 | Depth (lso) | Hypo(lso) | Similarity value | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Wu & Palmer Method | The proposed Method |
| **1** | **($C_2$, $C_3$)** | 2 | 2 | 2 | 9 | 7 | 1 | 26 | 0.5 | 0.0319 |
| **2** | **($C_2$, $C_4$)** | 2 | 2 | 2 | 9 | 7 | 1 | 26 | 0.5 | 0.0319 |
| **3** | **($C_3$, $C_4$)** | 2 | 2 | 2 | 7 | 7 | 1 | 26 | 0.5 | 0.0290 |
| **4** | **($C_5$, $C_6$)** | 3 | 3 | 2 | 2 | 4 | 2 | 9 | 0.6667 | 0.5205 |
| **5** | **($C_5$, $C_7$)** | 3 | 3 | 2 | 2 | 0 | 2 | 9 | 0.6667 | 0.3700 |
| **6** | **($C_5$, $C_8$)** | 3 | 3 | 4 | 2 | 2 | 2 | 26 | 0.5 | 0.0163 |
| **7** | **($C_5$, $C_9$)** | 3 | 3 | 4 | 2 | 3 | 2 | 26 | 0.5 | 0.0174 |
| **8** | **($C_5$, $C_{10}$)** | 3 | 3 | 4 | 2 | 1 | 2 | 26 | 0.5 | 0.0149 |
| **9** | **($C_5$, $C_{11}$)** | 3 | 3 | 4 | 2 | 4 | 2 | 26 | 0.5 | 0.0184 |
| **10** | **($C_6$, $C_7$)** | 3 | 3 | 2 | 4 | 0 | 2 | 9 | 0.6667 | 0.4074 |
| **11** | **($C_6$, $C_8$)** | 3 | 3 | 4 | 4 | 2 | 2 | 26 | 0.5 | 0.0184 |
| **12** | **($C_6$, $C_9$)** | 3 | 3 | 4 | 4 | 3 | 2 | 26 | 0.5 | 0.0198 |
| **13** | **($C_6$, $C_{10}$)** | 3 | 3 | 4 | 4 | 1 | 2 | 26 | 0.5 | 0.0167 |
| **14** | **($C_6$, $C_{11}$)** | 3 | 3 | 4 | 4 | 4 | 2 | 26 | 0.5 | 0.0211 |
| **15** | **($C_7$, $C_8$)** | 3 | 3 | 4 | 0 | 2 | 2 | 26 | 0.5 | 0.0131 |
| **16** | **($C_7$, $C_9$)** | 3 | 3 | 4 | 0 | 3 | 2 | 26 | 0.5 | 0.0138 |
| **17** | **($C_7$, $C_{10}$)** | 3 | 3 | 4 | 0 | 1 | 2 | 26 | 0.5 | 0.0122 |
| **18** | **($C_7$, $C_{11}$)** | 3 | 3 | 4 | 0 | 4 | 2 | 26 | 0.5 | 0.0144 |
| **19** | **($C_8$, $C_9$)** | 3 | 3 | 2 | 2 | 3 | 2 | 7 | 0.6667 | 0.5995 |
| **20** | **($C_8$, $C_{10}$)** | 3 | 3 | 4 | 2 | 1 | 2 | 26 | 0.5 | 0.0149 |
| **21** | **($C_8$, $C_{11}$)** | 3 | 3 | 4 | 2 | 4 | 2 | 26 | 0.5 | 0.0184 |
| **22** | **($C_9$, $C_{10}$)** | 3 | 3 | 4 | 3 | 1 | 2 | 26 | 0.5 | 0.0159 |
| **23** | **($C_9$, $C_{11}$)** | 3 | 3 | 4 | 3 | 4 | 2 | 26 | 0.5 | 0.0198 |
| **24** | **($C_{10}$, $C_{11}$)** | 3 | 3 | 2 | 1 | 4 | 2 | 7 | 0.6667 | 0.5744 |
| **25** | **($C_2$, $C_8$)** | 2 | 3 | 3 | 9 | 2 | 1 | 26 | 0.4 | 0.0223 |
| **26** | **($C_3$, $C_6$)** | 2 | 3 | 3 | 7 | 4 | 1 | 26 | 0.4 | 0.0244 |
| **27** | **($C_4$, $C_9$)** | 2 | 3 | 3 | 7 | 3 | 1 | 26 | 0.4 | 0.0227 |

In Table 2, hypo1 represents the number of hyponym of the first node of node-pair, and hypo2 represents the number of hyponym of the second node of node-pair. Depth1 represents the depth of the first node of node-pair, and depth2 represents the depth of the second node of node-pair. Hypo(lso) represents the hyponym of the most specific common hypernym of node-pair.

Depth(lso) represents the depth of the most specific common hypernym of node-pair.

## 4.2. Discussions

There are several aspects have to be addressed on the proposed method.

The first aspect involved the measuring stabilization. From the similarity value of rows 1-3 in Table 2, we concluded the similarity value was equal if all of the parameters are equal; the similarity value was different if one of parameters is different. Thus the proposed method owned good stabilization.

The second aspect related to the measuring sensitivity. For 4-25 rows, the results indicated the proposed method could distinguish the similarities between nodes from a node to different nodes of the same level in the classification tree. For example, $c_{10}$ and $c_{11}$ lay in the same level, using the Wu and Palmer's measure, $sim_{W\&P}$ $(c_8,c_{10})$equals$sim_{W\&P}(c_8,c_{11})$, the value equals 0.5.Wu and Palmer's measure failed to distinguish the similarity between nodes from a node to different nodes of the same level. Using the proposed method, $sim_{new}$ $(c_8,c_{10})$ equals 0.0149, $sim_{new}(c_8,c_{11})$equals 0.0184, so the proposed method overcame the shortcoming of Wu & Palmer's measure.

The third aspect dealt with the problem of different level nodes through the parameter hyponym, which can be used to discriminate the specificity of each node. For example, to the 25 row, the value of $(c_2, c_8)$ equals 0.0223; to the 26 row, the value of $(c_3, c_6)$ equals 0.0244; to the 27 row, the value of $(c_4, c_9)$ equals 0.0227. The results showed that the proposed method could distinguish the similarity of two nodes of different level because node pairs $(c_2, c_8)$, $(c_3, c_6)$ and $(c_4, c_9)$ subsumed different number of hyponym in the classification tree.

The fourth aspect, the proposed method is reasonable and consistent with earlier methods. For example, where $sim_{new}(c_5, c_6)$equals0.5205 and $sim_{new}(c_5, c_{10})$ equals 0.0149, in which the similarity of two nodes was larger in large branches than small branches. This mean that the proposed method was consistent with the previous studies of Zhou and Seco[20, 28], which indicated that a concept of lower abstraction owned higher IC.

The fifth aspect is the measuring complexity. Wu and Palmer's [25] method and the proposed method are all computing logarithm. The main factor of complexity is the number of parameters. Wu and Palmer's measure include three parameter, depth(c), $lso(c_1, c_2)$ and $len(c_1, c_2)$. The proposed method include three parameter hypo(c), $lso(c_1, c_2)$ and max_nodes. The complexity of method is similar.

The six aspect, variance is an important standard for estimating the discretization of a set of data. The variance of the proposed method is 0.037, so the proposed method owns a good performance in discretization.

Finally, there are two insufficient in the proposed method. Firstly, in this paper, the proposed method only concentrated on hierarchical structure of WordNet without considering the network structure of WordNet. Secondly, the proposed method focuses only on the single inheritance node without considering the multiple inheritances node in WordNettaxonomy. So there are some studies need be done for these aspects in future.

## 5. Conclusions

In this paper, through analyzing IC computing models and concept semantic similarity measures, we put forward a new hybrid method to improve Wu and Palmer's [25] problem which didn't distinguish the similarities between nodes from a node to different nodes of the same level in taxonomy. By an experiment on Wu and Palmer's. [25] measure and the proposed measure in a fragment of WordNet hierarchical taxonomy, the results show the proposed method solves the problem of Wu and Palmer's measure [25].

In general, the proposed method owned two features. First, the proposed method was based on WordNet intrinsic structure, and took into account the path and depth factor, so this measure is a hybrid method. Second, the proposed method converted calculating the minimum distance of node-pair into seeking the hyponyms of node-pair and their most specific common hypernym, in which the proposed method improved the accuracy and not increasing the workload.

In future, we will improve this method by considering the spatial structure of WordNet hierarchical taxonomy and proof-test this method in common data set of Miller and Charles [12], and Rubenstein and Goodenough [17].

## Acknowledgements

The authors acknowledge the National Natural Science Foundation of China (Grant: 61562072).

## References

[1] Adhikari A., Dutta B., Dutta A., Mondal D., and Singh S., "An Intrinsic Information Content-Based Semantic Similarity Measure Considering the Disjoint Common Subsumers of Concepts of an Ontology," *Journal of the Association for Information Science and Technology*, vol. 69, no. 8, pp. 1023-1034, 2018.

[2] Aouicha M., Taieb M., and Ben Hamadou A., "Taxonomy-Based Information Content and Wordnet-Wiktionary-Wikipedia Glosses for Semantic Relatedness," *Applied Intelligence*, vol. 45, no. 2, pp. 475-511, 2016.

[3] Bailey R., "Frequency Analysis of English Usage: Lexicon and Grammar, and: Word Frequencies in British and American English," *Dictionaries*

*Journal of the Dictionary Society of North America*, vol. 5, no. 1, pp. 128-134,1983.

[4] Cai Y., Zhang Q., Lu W., and Che X., "A Hybrid Approach for Measuring Semantic Similarity Based on IC-Weighted Path Distance in Wordnet," *Journal of Intelligent Information Systems*, vol. 51, no. 1, pp. 23-47, 2018.

[5] Devitt A. and Vogel C., "The Topology of Wordnet: Some Metrics," *in Proceedings of GWC-04, 2nd Global WordNet Conference*, Brno, pp. 106-111, 2004.

[6] Jiang J. and Conrath W., "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *in Proceedings of International Conference on Research in Computational Linguistics*, Taiwan, pp. 19-33, 1997.

[7] Karra W. and Slimani T., "A New Approach for Arabic Named Entity Recognition," *The International Arab Journal of Information Technology*, vol. 14, no. 3, pp. 332-338, 2017.

[8] Leacock C. and Chodorow M., "C-Rater: Automated Scoring of Short-Answer Questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 389-405, 2003.

[9] Lin D., "An Information-Theoretic Definition of Similarity," *in Proceedings of 15th International Conference on Machine Learning*, Morgan Kaufmann, pp. 296-304, 1998.

[10] Lofi C., "Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approachesm," *Information and Media Technologies*, vol.10, no. 3, pp. 493-501, 2015.

[11] Lu W., Cai Y., Che X., and Lu Y., "Joint Semantic Similarity Assessment with Raw Corpus and Structured Ontology for Semantic-Oriented Service Discovery," *Personal and Ubiquitous Computing*, vol. 20, no. 3, pp. 311-323, 2016.

[12] Miller G. and Charles W., "Contextual Correlates of Semantic Similarity," *Language Cognition and Neuroscience*, vol. 6, no. 1, pp. 1-28, 1991.

[13] Pirró G., "A Semantic Similarity Metric Combining Features and Intrinsic Information Content," *Data and Knowledge Engineering*, vol. 68, no. 11, pp. 1289-1308, 2009.

[14] Pirró G. and Euzenat J., "A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness," *International semantic web conference*, Berlin, pp. 615-630, 2010.

[15] Rada R., Mili H., Bicknell E., and Blettner M., "Development and Application of A Metric on Semantic Nets," *IEEE Transactions on Systems Man and Cybernetics*, vol. 19, no. 1, pp. 17-30, 1989.

[16] Resnik P, "Using Information Content to Evaluate Semantic Similarity in A Taxonomy," *in Proceeding of international Joint Conference on Artificial Intelligence*, San Francisco, pp. 448-453, 1995.

[17] Rubenstein H. and Goodenough J., "Contextual Correlates of Synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627-633, 1965.

[18] Sánchez D. and Batet M., "Semantic similarity estimation in the Biomedical Domain: An Ontology-Based Information-Theoretic Perspective," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 749-59, 2011.

[19] Sánchez D., Ribalta A., Batet M., and Serratosa F., "Enabling Semantic Similarity Estimation Acrossmultiple Ontologies: An Evaluation in the Biomedical Domain," *Journal of Biomedical Informatics*, vol. 45, no. 1, pp. 141-155, 2012.

[20] Seco N., Veale T., and Hayes J., "An Intrinsic Information Content Metric for Semantic Similarity in Wordnet," *in Proceeding of European Conference on Artificial Intelligence, Ecai', Including Prestigious Applicants of Intelligent Systems*, Spain, pp. 1089-1090, 2004.

[21] Taieb M., Ben Aouicha M., and Ben Hamadou A., "Ontology-Based Approach for Measuring Semantic Similarity," *Engineering Applications of Artificial Intelligence*, vol. 36, pp. 238-261, 2014.

[22] Taieb M., Ben Aouicha M., and Ben Hamadou A., "Computing Semantic Relatedness Using Wikipedia Features," *Knowledge-Based Systems*, vol. 50, no. 50, pp. 260-278, 2013.

[23] Taieb M., Ben Aouicha M., and Ben Hamadou A., "A New Semantic Relatedness Measurement Using Wordnet Features," *Knowledge and Information Systems*, vol. 41, no. 2, pp. 467-497, 2014.

[24] Varelas, G., Voutsakis E., Raftopoulou P., and Petrakis E., "Semantic Similarity Methods in Wordnet and Their Application to Information Retrieval on the Web," *in Proceeding of ACM International Workshop on Web Information and Data Management*, Germany, pp. 10-16, 2005.

[25] Wu Z. and Palmer M., "Verb Semantics and Lexical Selection," *in Proceedings of Annual Meeting on Association for Computational Linguistics*, United States, pp. 133-138, 1994.

[26] Zesch T., "Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources," Thesis, Technische Universität, 2010.

[27] Zhang Y., Shang L., Huang L., Porter A., Zhang G., Lu J., and Zhu D., "A Hybrid Similarity Measure Method for Patent Portfolio Analysis," *Journal of Informetrics*, vol. 10, no. 4, pp. 1108-1130, 2016.

[28] Zhou Z., Wang Y., and Gu J., "A New Model of Information Content for Semantic Similarity in Wordnet," *in Proceeding of International*

*Conference on Future Generation Communication and Networking Symposia*, China, pp. 85-89, 2008.

**Xiaogang Zhang** PhD student, College of Computer Science and Technology, Zhejiang University. Major research: Data Mining, Semantic Computation.

**Shouqian Sun**, College of Computer Science and Technology, Zhejiang University. Major research: Application of ergonomics and design, creative design services, intelligent sports equipment technology.

**Kejun Zhang**, College of Computer Science and Technology, Zhejiang University. Major research: Data Mining, applying Machine Learning, Evolutionary Algorithms, and Computational Linguistics techniques to the extraction of knowledge from music.