A New Hypred Improved Method for Measuring Concept Semantic Similarity in WordNet

Xiaogang Zhang¹,², Shouqian Sun² and Kejun Zhang³
¹College of Computer, Zhejiang University, Hangzhou, China
²College of Information Engineering, Tarim University, 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 foundation of semantic computation. In this paper, we analyze typical semantic similarity measures and note Wu & 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 overcomes the shortcoming of the Wu & Palmer’s measure.

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

Received May 25, 2017; accepted April 25, 2018
doi

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) [6, 9]. As the minimal unit of describing information and the basis for information resource matching, concept owns 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 [6]. WordNet is a common ontology developed by the 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 [12]. The WCG is the other resource in some research works, including in works of Taieb et al. [2, 7] and Zesch [26]. The WCG is different from WordNet owing to WCG is proposed by volunteers, and the categories of WCG do not include specifying the type in semantic relation [6]. 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 IC (Information Content) computing models and classical semantic similarity measures. In section 3, we study the existing problem of Wu and Palmer’s measure [25] 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 [13]. 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 [15]. The measure of feature-based estimated the semantic similarity between words according to the structural feature of taxonomy, which included nodes and edges [21]. 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 [16]. 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 symbol used in this paper is shown in following Table 1.

### Table 1. The list of symbols in similarity computing.

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

#### 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 put forward a method that the probability of concept equaled the frequency of noun appearing in the Brown Corpus [3]. Resnik used negative log likelihood to calculate IC. The corresponding calculating equation is as follows [18]:

$$IC(c) = -\log(p(c))$$  

(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 [18]:

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

(2)

Here, $\text{Word}(c)$ is a set of words subsumed by $c$, and $\text{Count}(\omega)$ represents the frequency of the word $\omega$ in the given corpus. Where the function $p(c)$ could be computed as follows [18]:

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

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

##### 2.1.2. IC Model Based on Taxonomical Structure

Seco et al. [22] 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 last node is larger. The IC of a concept only relied on the number of concepts which it subsumes. The Seco’s method of calculating IC was as follows [22]:

$$IC(c) = 1 - \frac{\log(\text{count}(\text{subsumers})))}{\log(max\_nodes)}$$  

(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:

$$IC_{zhou}(c) = K \left( \frac{\log(\text{count}(\text{subsumers}))}{\log(max\_nodes)} \right) + (1-K) \left( \frac{\log(depth(c))}{\log(max\_depth(c))} \right)$$  

(5)

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

Later, Sanchez et al. [19] proposed a new model, adopting subsumers of leaf node to calculate the value of IC. The equation was as follows:

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

(6)

Where the function $\text{commonness}(c)$ equals the sum of $\text{commonness}(n)$, and $\text{commonness}(n)$ 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 Concept Similarity

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. [17], Wu and Palmer [25], Leacock and Chodorow [10], Resnik [19], Jiang and Conrath [8] and Lin measure [11].

##### 2.2.1. The Measure Based Path and Depth

Rada et al. [17] 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) is as follows [17]:

\[
dis_{\text{rad}}(c_1, c_2) = \min_{t \in \text{path}(c_1, c_2)} |t|
\]

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:

\[
sim_{\text{W&P}}(c_1, c_2) = \frac{2 \times \text{depth}(\text{iso}(c_1, c_2))}{\text{len}(c_1, c_2) + 2 \times \text{depth}(\text{iso}(c_1, c_2))}
\]

Later, Leacock and Chodorow proposed a non-linear calculating model, which included two parameters \(\text{len}(c_1, c_2)\) and \(\text{max\_depth}(c)\). The calculating equation is as follows [10]:

\[
sim_{\text{L&C}}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \times \text{max\_depth}(c)}
\]

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 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, 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 [18]:

\[
sim_{\text{R}}(c_1, c_2) = -\log p(\text{iso}(c_1, c_2)) - \text{IC}(\text{iso}(c_1, c_2))
\]

Jiang and Conrath 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 [8]:

\[
\text{sim}(w_1, w_2) = \max_{j}(\text{sim}(s_{1j}, s_{2j}))
\]

Here, \(s_{1j}\) and \(s_{2j}\) are the significance of \(w_1\) and \(w_2\) (the concept of ontology). Jiang and Conrath 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 [8]:

\[
dis_{\text{IC}}(c_1, c_2) = \text{IC}(c_1) + \text{IC}(c_2) - 2 \times \text{IC(\text{iso}(c_1, c_2))}
\]

After a linear transformation, the equation of measuring semantic similarity is as follows [22]:

\[
sim_{\text{J&C}}(c_1, c_2) = 1 - \frac{\text{IC}(c_1) + \text{IC}(c_2) - 2 \times \text{IC(\text{iso}(c_1, c_2))}}{2}
\]

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 equation is as follows [11]:

\[
sim_{\text{Lin}}(c_1, c_2) = \frac{2 \times (-\log p(c_1)) - \log p(c_2)}{\text{IC}(c_1) + \text{IC}(c_2)}
\]

In a taxonomical tree, if \(c_1 \in C_1\) and \(c_2 \in C_2\), the commonality \(c_0\) could expressed as \(c_0 \in C_0 \cap C_1, c_0 \in C_0\), 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 [17] Leacock and Chodorow’s 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 [18]. Jiang and Conrath’s and Lin’s 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 measure used the IC model of Equation (1), \(IC(c) = -\log(p(c))\). The accuracy of Wu and Palmer’s method was unsatisfactory [20]. 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.](image-url)
the same level, \( \text{sim}_{\text{W&P}}(c_{16},c_{19}) \) equals \( \text{sim}_{\text{W&P}}(c_{17},c_{18}) \), 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 measure.

### 3.2. Measure Improvement

The studies of Seco et al. [22] and Zhou [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 descendant [6]. 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 \( \text{hypo}(c) = \{c_i \in V, c_i < c\} \). \( c \) is the descendant of \( c \), \( V \) is a set of concepts of the classification tree.

- **Definition 2:** (The most specific common abstraction) defines the concept of \( \text{iso}(c_1,c_2) \), which represents the most specific common hyponym of \( c_1 \) and \( c_2 \).

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

\[
\text{sim}_{\text{new}}(c_1, c_2) = \frac{2 \times \log(\text{max_nodes}) - \log(\text{hypo}(\text{iso}(c_1,c_2))) + 1}{2 \times \log(\text{max_nodes}) - \log(\text{hypo}(c_1)) - \log(\text{hypo}(c_2)) + 2}
\]

In Equation (15), hyponym is the most specific common abstraction of \( c_1 \) and \( c_2 \). The function \( \text{hypo}(\text{iso}(c_1,c_2)) \) 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 \( \{c_1,c_2,\ldots,c_{10}\} \) 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>
<thead>
<tr>
<th>Node-pair</th>
<th>Depth1</th>
<th>Depth2</th>
<th>Len</th>
<th>Hypo1</th>
<th>Hypo2</th>
<th>Depth (iso)</th>
<th>Hypo(iso)</th>
</tr>
</thead>
<tbody>
<tr>
<td>((c_1,c_2))</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_3))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_4))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_5))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_6))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_7))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_8))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_9))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
<tr>
<td>((c_1,c_{10}))</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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(iso) represents the hyponym of the most specific common hyponym 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_{t0} and c_{t1} lay in the same level, using the Wu and Palmer’s measure, sim_{W&P}(c_{t0},c_{t0})=simm_{W&P}(c_{t0},c_{t1}), 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_{t0},c_{t0})=0.0149, sim_{new}(c_{t0},c_{t1})=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_{t2}, c_{t8}) equals 0.0223; to the 26 row, the value of (c_{t8}, c_{t8}) equals 0.0244; to the 27 row, the value of (c_{t4}, c_{t9}) equals 0.0227. The results showed that the proposed method could distinguish the similarity of two nodes of different level because node pairs (c_{t2}, c_{t8}), (c_{t8}, c_{t8}) and (c_{t4}, c_{t9}) subsumed different number of hypernym in the classification tree.

The fourth aspect, the proposed method is reasonable and consistent with earlier methods. For example, where sim_{new}(c_{t3}, c_{t1})=0.5205 and sim_{new}(c_{t5}, c_{t10})=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 et al., which indicated that a concept of lower abstraction owned higher IC.

The fifth aspect is the measuring complexity. Wu and Palmer’s method and the proposed method are all computing logarithm. The main factor of complexity is the number of parameters. Wu and Palmer’s measure 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 WordNet taxonomy. 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 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.

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 hypernoms 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 [14], and Rubenstein and Goodenough [19].

Acknowledgements

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

References


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

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

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