A Semantic Similarity Algorithm Based on the Nearest Common Ancestor Node

Zhe ZHANG, Yun-xiao ZU* and Bin HOU

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author

Keywords: Sememe, The nearest common ancestor node, Semantic similarity.

Abstract. The word semantic similarity calculation is the basis of keyword search in the field of semantic web information retrieval. In view of the deficiency of semantic similarity computation in the existing algorithms, this paper discussed the influence of node depth on the semantic similarity, and presented an improved similarity calculation method based on the nearest common ancestor node, which can represent the depth of the two sememe nodes that need to compute the similarity. The simulation was carried out by using simple biological ontology and HowNet system architecture data, and the comparison with the original algorithm was done. The results prove the correctness and rationality of the improved method.

Introduction

In recent years, the word semantic similarity calculation in such as information retrieval, machine translation, personalized recommendation and data mining, and other fields has been more and more widely used, which put forward higher requirements on the accuracy of the calculation results of semantic similarity. At present, the semantic similarity calculation method can be divided into two categories. One method is based on the probability distribution of the information in the context of words. In this method, two words are replaced in different contexts, and the semantic similarity is judged according to the degree syntactic structure change before and after replacement. The greater the possibility that the syntactic semantic structure of the text not changing, the higher the words similarity. This method is more accurate, but it needs to be calculated by large-scale corpus. And the calculation is much more complex and easy to be influenced by the size of corpus and the noise of training data. Another method is based on a world knowledge system. This method calculates the semantic similarity of words mainly based on the hierarchical structure of a complete semantic dictionary, which is relatively simple and intuitive, but more dependent on the dictionary. As for the selection of dictionaries, we can select a comprehensive semantic dictionary, such as the English Dictionary, named as WordNet and the common knowledge database of Chinese words, HowNet. We can also build a semantic dictionary for a specific domain, which may cover a smaller range, but can be more detailed division within the field. In this paper, we mainly use the above second method to calculate the semantic similarity of words.

The Semantic Similarity of Chinese Words

Dekang LIN pointed out that the similarity of any two things depends on their commonalities and differences, and then gave the general formula of similarity calculation between two things from the perspective of information theory.

\[ Sim(A, B) = \frac{\log p(common(A,B))}{\log p(description(A,B))} \] (1)

Here, the denominator part, \( \log p(description(A,B)) \), is the amount of the information needed to completely describe the two things of \( A \) and \( B \), while the molecular part, \( \log p(common(A,B)) \), represents the magnitude of the information required to describe the commonalities of \( A \) and \( B \).
The Knowledge Base Structure of HowNet

HowNet is a common sense knowledge base, which takes the concepts represented by Chinese and English words as descriptive objects to reveal the relationship between different concepts and the attributes of concepts. In HowNet, each word will correspond to a number of concepts, and the concept is defined on the basis of sememes by knowledge base description language, which will be called the sense expression of the concept. The sememe is described by a variety of relationships, such as hyponymy, antonymy, and so on, afterwards they can form a tree diagram with the sememes as nodes. At present, the semantic similarity computation algorithms based on HowNet are all carried out according this tree. For the first time, Qun LIU et al. gave the method of computing lexical semantic similarity, which considered the path length between the elements. Dai et al. presented a similarity algorithm between Chinese and English words based on HowNet, while Xiao-lin WANG proposed an arithmetic of variable coefficients.

In HowNet, the words are described in terms of concepts, and a word can be expressed as several concepts, while the concept is described by the sememes. If we suppose the word "xg1849" has concepts which are "xg1829", "xg2869", and the word "xg1849" has concepts, "xg1829", "xg2870",⋯⋯,"xg3041", the semantic similarity value between the two words of "xg1849" and "xg1849" is defined as the maximum value of the absolute values of similarity between all concepts.

\[
\text{Sim}(W_1, W_2) = \max_{i=1,2,\cdots,n, j=1,2,\cdots,m} |\text{Sim}(C_{1i}, C_{2j})| \tag{2}
\]

As shown in Eq. 2, the similarity calculation of two words is transformed into the computation of similarity between two concepts. Because the concepts are expressed by sememes in HowNet, so the similarity calculation of sememes becomes the basis of semantic similarity calculation of words.

The Calculation of Sememes Similarity

The similarity computation of the sememes is usually based on the tree hierarchy of the dendrogram, and is calculated according to the path length between the two nodes.

\[
\text{Sim}(S_1, S_2) = \frac{\alpha}{\text{distance}(S_1, S_2) + \alpha} \tag{3}
\]

Here, distance \((S_1, S_2)\) represents the distance between the two points, and when \(S_1\) and \(S_2\) are not in a same tree, a larger constant value is adopted. As an adjustable parameter, \(\alpha\) can be defined as the path length between the two nodes when the similarity is 0.5.

The Improved Semantic Similarity Algorithm Based on the Nearest Common Ancestor Node

The Improved Method for Computing Sememes Similarity

After analyzing of the above Eq. 3, we can find out that the existing algorithms only consider the distance between two sememes, while the effect of node depth on similarity is ignored. For example, in the sememes hierarchy architecture, the distance between \{“animals”, “plants”\} is the same as \{“vegetables”, “fruits”\}, both of which are 2, so they have the same value of similarity. However, in the personal intuitive understanding, it is clear that the former similarity should be lower than the latter, for the latter is a further subdivision of the sememe in the previous group. Thus it can be seen that the calculation results obtained by just considering the distance factors are often too rough and unreasonable.

On the basis of the existing algorithms, this paper further considers the influence of the node depth on the similarity between the sememes. The main idea is as follows: to find the nearest common ancestor node of the two sememes nodes, and use the depth of this node to represent the depth of the two sememes nodes need to compute the similarity. If the sememe node is deeper in the tree structure, it represents that the node is divided into a more detailed level, so the depth of the nearest common ancestor node is positively correlated with the similarity. Therefore, the improved formula to calculate the sememes similarity is obtained.
In this paper, we also have already discussed the method of computing the similarity between any two sememes and focused on how to synthesize the similarity between two concepts by using the sememes similarity. We refer to the idea of concepts similarity calculation proposed in Qun LIU’s paper, that is, the overall similarity can be obtained by the synthesis of partial similarity.

In HowNet, the description of a content word can be defined as a feature structure, which can be understood as a set of “Attribute-Value” pairs, and we call each pair as a “Feature”. In any feature structure, every "attribute" corresponding to each "feature" is unique. The feature structure not only describes the semantic information of words, but also gives the connection between concepts. The feature structure contains the following three parts.

1. Independent sememe description: the content is a basic sememe or a specific word, and we record this part of similarity as $Sim_1(S_1, S_2)$.

2. Relation sememe description: the content is a feature structure. For every feature, its attribute is a relation sememe, and its value is a basic sememe or a specific word. We named this part of similarity as $Sim_2(S_1, S_2)$.

3. Relation symbol description: just as the relation sememe description, the content is also a feature structure. But the difference is that the value of each feature is a set, and the element of that set is a basic sememe or a specific word. We recorded this part of similarity as $Sim_3(S_1, S_2)$.

Thus we can propose the formula of the overall semantic similarity of two concepts.

$$Sim(S_1, S_2) = \sum_{i=1}^{2} \beta_i \prod_{j=1}^{3} Sim_j(S_1, S_2)$$ (5)

Here, $\beta_i$ (1 ≤ i ≤ 3) are adjustable parameters, and $\beta_1 + \beta_2 + \beta_3 = 1$, $\beta_1 \geq \beta_2 \geq \beta_3$. It reflects that the effect on the overall similarity from $Sim_1$ to $Sim_3$ is more and more small, and the similarity of the main parts will influence the similarity of other parts.

Simulation and Results Analysis

Simulation Based on Biological Ontology

Ontology Construction. Protégé is a piece of software used for ontology editing and knowledge acquisition based on the Java language, which was developed at biological information research center of Medical College of Stanford University. This software is an ontology development tool and a knowledge-based editor, belonging to open source software. It is mainly used in the construction of ontology and is the core development tool of ontology construction in semantic web.
Protégé enables the developer constructing ontology concept classes, relationships, attributes, and instances, and the specific ontology description language is shielded. Users only need to construct domain ontology models at the conceptual level.

Based on the biological classification map, we constructed a simple biological ontology instance using Protégé development tools, and the parts of its tree structure graph are shown in Figure 1.

**Analysis of Simulation Results.** In this paper, the algorithm based on the nearest common ancestor node, showing in Eq. 4, was compared with the algorithm proposed by Qun LIU et al. [7]. In Eq. 3, which is Qun LIU's algorithm, the value of \( \alpha \) is 1.6. According to the experience gained in several attempts, the result is the most ideal when the value of \( \alpha \) in Eq. 4 is 0.3. The comparison of these two algorithms is shown in Table 1.

<table>
<thead>
<tr>
<th>Node 1</th>
<th>Node 2</th>
<th>Eq. 3</th>
<th>Eq. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virus</td>
<td>Eukaryotes</td>
<td>0.444444</td>
<td>0.130435</td>
</tr>
<tr>
<td>Pisce</td>
<td>Amphibian</td>
<td>0.444444</td>
<td>0.428571</td>
</tr>
<tr>
<td>Australopithecinae</td>
<td>Homininae</td>
<td>0.444444</td>
<td>0.6</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>Nostocales</td>
<td>0.285714</td>
<td>0.069767</td>
</tr>
<tr>
<td>Cetacea</td>
<td>Tetrapoda</td>
<td>0.285714</td>
<td>0.272727</td>
</tr>
<tr>
<td>Monkey</td>
<td>Homininae</td>
<td>0.285714</td>
<td>0.375</td>
</tr>
<tr>
<td>DNA virus</td>
<td>Bryophyta</td>
<td>0.210526</td>
<td>0.047619</td>
</tr>
<tr>
<td>Gymnospermae</td>
<td>Vertebrata</td>
<td>0.210526</td>
<td>0.090909</td>
</tr>
<tr>
<td>Primates</td>
<td>Homo sapiens</td>
<td>0.210526</td>
<td>0.259259</td>
</tr>
</tbody>
</table>

We can see from Table 1 that every three rows are divided into one group, and the distances between pairs of nodes in every group are equal, therefore they have the same similarity calculated by Eq. 3. But according to the general personal judgment the approximate degree of "Australopithecinae" and "Homininae" should be higher than that of "Virus" and "Eukaryotes". The calculation result of Eq. 4 is more consistent with this cognition. In addition, when the depth of the nearest common ancestor node of the two nodes is shallower, the similarity value calculated by Eq. 4 is relatively less. This result also accords with the general perception, that is, the deeper the level of the node is, the more detailed it is divided, and the greater the similarity between the two nodes is. In this simulation, the biological ontology has 13 layers, and when the nearest common ancestor node of the two nodes is located in the sixth layer, the similarity values calculated by the two algorithms are the closest.

**Simulation Based on HowNet**

We applied the improved algorithm proposed in this paper to HowNet architecture, and compared it with the original algorithm. The simulation parameters are shown in Table 2, and the results of the two algorithms are shown in Table 3.
Table 2. Simulation parameters list.

<table>
<thead>
<tr>
<th></th>
<th>(\alpha)</th>
<th>(\beta_i)</th>
<th>(\gamma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qun LIU’s algorithm ([7])</td>
<td>1.6</td>
<td>0.17</td>
<td>0.13</td>
</tr>
<tr>
<td>Improved algorithm</td>
<td>0.5</td>
<td>0.14</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 3. Results of similarity calculation in HowNet.

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Qun LIU’s algorithm ([7])</th>
<th>Improved algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker</td>
<td>Teacher</td>
<td>0.722222</td>
<td>0.684498</td>
</tr>
<tr>
<td>Worker</td>
<td>Scientist</td>
<td>0.575926</td>
<td>0.665812</td>
</tr>
<tr>
<td>Teacher</td>
<td>Scientist</td>
<td>0.575926</td>
<td>0.712696</td>
</tr>
<tr>
<td>Red</td>
<td>Pink</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>Red</td>
<td>Yellow</td>
<td>0.861111</td>
<td>0.819689</td>
</tr>
<tr>
<td>Golden</td>
<td>Orange</td>
<td>0.861111</td>
<td>0.887845</td>
</tr>
<tr>
<td>Walk</td>
<td>Act</td>
<td>0.615385</td>
<td>0.705687</td>
</tr>
<tr>
<td>Move</td>
<td>Act</td>
<td>0.661111</td>
<td>0.705687</td>
</tr>
<tr>
<td>Walk</td>
<td>Run</td>
<td>0.615385</td>
<td>0.779194</td>
</tr>
</tbody>
</table>

In the simulation, we selected nouns, adjectives and verbs to calculate their similarity respectively. As shown in Table 3, the results of Qun LIU’s algorithm and the improved algorithm are generally close to each other. But the original algorithm cannot distinguish the similarities of some words pairs. For example, the similarity between \{“Worker”, “Scientist”\} is equal to that of \{“Teacher”, “Scientist”\}, and \{“Red”, “Yellow”\} has the same value of similarity as \{“Golden”, “Orange”\}. The improved algorithm can distinguish these differences and conform to the general personal judgment, that is, as a more detailed classification, the similarity of the latter is slightly higher than that of the former. In addition, this paper gave the basic sememe a greater weight in the overall similarity, so that when the basic sememe is the same, the similarity of the two words will be greater.

It is necessary to declare that, in the simulation based on HowNet, we only selected one of the most common meaning of each word rather than all the meanings.

Conclusions

Semantic similarity computation is the basis of keyword search. Based on the original algorithm, we improves the calculation method of words similarity. We further consider the influence of the node depth, and use the depth of the nearest common ancestor node of the two sememes nodes to represent the depth of the two nodes that need to compute the similarity. And the similarity of two words is positively correlated with the depth of their nearest common ancestor node.

The improved similarity calculation method was compared with the existing algorithm by simulation. The results show that the improved method can enhance the accuracy of similarity calculation effectively.

Acknowledgement

This work is supported by Beijing Key Laboratory of Work Safety Intelligent Monitoring (Beijing University of Posts and Telecommunications).

References


