An Improved HowNet-based Algorithm for Semantic Similarity Computation

LIJUAN DIAO, HONG YAN, FUXUE LI, XIUMIN LI and GUOHUA LEI

ABSTRACT

Similarity technology has been widely used in many applications. However, most of the similarity researches focus on English language. In this paper, we develop the improved similarity technology and focus on its application to Chinese language. We use HowNet, which is a Chinese version of WordNet, as the bases. We analyze the structure and data resource of HowNet, and develop an improved information content model of concepts. Meanwhile, we extend the synonyms of concepts, and use them to develop a better similarity computation technique. The experiment results show that our methods achieve the best compared with other existing methods in the degree of similarity accuracy.

KEYWORDS

Semantic Similarity, Information Content Model, Semantic Distance.

INTRODUCTION

Similarity measure is widely used in many artificial intelligence domains. For example, it is used in content-based image retrieval [1], cluster, classification [2], and disambiguation in natural language [3], pattern recognition [4] and ontology learning [5]. There have been significant advances in semantic similarity technologies over the past two decades. In the beginning, distance-based semantic similarity is the major paradigm. For example, it is shortest path [6] and weighted links [7]. Rada proposes similarity computation in term of semantic distance of two concepts, in which semantic distance is the distance of the shortest path between two concepts in the classification tree. In their algorithms, similarity of two concepts also depends on the depth of the concepts in the classification tree. Based on the literatures of distance-based similarity measures, we can see that none of the algorithms have a comprehensive consideration of all factors. This restricts the potential performance of the algorithm.

According to the analysis of information content-based, information content is only taken into consideration in the process of semantic similarity calculation of these algorithms, and similarity result of concept pairs is not reasonable. Therefore, we need to combine the two aspects that are semantic distance and information content, when the semantic similarity algorithm is improved.

Qun Liu, Qinglei Liu and Feng Li have proposed semantic similarity algorithms [8] [9]. Semantic distance is only taken into account in the approach of Liu’s...
algorithm. Feng Li’s semantic similarity algorithm not only considers semantic distance of two sememes, but also considers the depth of two sememes in the whole sememe tree. The disadvantage of these two similarity algorithms are not included information content parameter in the process of similarity calculation.

In existing Chinese semantic similarity computation approaches, owing to only factors named semantic distance and the depth of two concepts are taken into consideration in the process of semantic similarity calculation, similarity result of concept pairs is not reasonable. According to advantage and disadvantage of distance-based and information content-based similarity computation, we propose the method of improved the similarity computation of Chinese concept based on HowNet and it is applied in the aspect of synonym expansion. This method contains the semantic distance and information content in the process of similarity computation. We present improved information content model that includes the depth and the number of nodes of whole hierarchy tree, and apply this model to compute the similarity. Meanwhile, we use two methods to verify the performance of the experimental results, which are the distribution of similarity and the correlation coefficient on similarity between every algorithm and artificial measured.

The rest of this paper is organized as follows. Section 2 presents the improved model of information content calculation process of HowNet-based Chinese semantic similarity used in this paper. Section 3 focuses on the techniques for improving semantic similarity calculation HowNet-based in Chinese concepts. Section 4 explains the conducted experiment and obtained outcome. Section 5 states the conclusion of the paper.

**IMPROVED INFORMATION CONTENT MODEL**

According to analysis of the above content, we present the improved the calculation model of information content, because it is very important factor to affect the value of semantic similarity. We consider the depth and the number of nodes of whole hierarchy tree and the depth and regional density of computational concept node in the whole hierarchy tree. Accordingly, we define the model of information content calculation as follows.

\[
IC(c) = k(1 - \frac{\log(hypo(c) + 1)}{\log(node_{max})}) + (1 - k)(\frac{\log(deep(c))}{\log(deep_{max})}
\]

Here, IC(c) is the information content of concept c. node_{max} is the number of all concepts in the classification tree. deep(c) denotes the depth of concept in the classification tree. deep_{max} is the whole depth of the classification tree and hypo(c) is the number of children conceptual nodes of the concept c respectively. In the information content model definition, k is adjusted parameter and is assumed value as 0.5.

We will use the information content model to compute the information contents of all concepts and then take use of the value of information contents to calculate similarity. Semantic similarity process is discussed in the following.
IMPROVED HOWNET-BASED SEMANTIC SIMILARITY ALGORITHM

We discuss semantic similarity definition and the relationship between semantic similarity and correlation. Then, we introduce the shortage of the existing similarity algorithms and present improved similarity algorithm base on the existing algorithms.

Semantic Similarity Definition

Similarity means the similar degree between two concepts and can be represented as a real number among the domain [0,1]. If two concepts are completely same, the similarity of two concepts is 1. While two concepts are completely different, the similarity of two concepts is 0.

Similarity refers as the similar degree between two concepts in the classification tree that only contains is_a semantic relation, while correlation refers as the relation degree between two concepts with the other semantic relation, for example, part_of, member_of and so on. We give an example of similarity and correlation, apple and banana, apple and vitamin. Therefore, we discuss the similarity between apple and banana, while the correlation between apple and vitamin.

Improved HowNet-based Chinese Semantic Similarity Model

According to Liu Qun’s semantic similarity algorithm[13], we draw only single factor that is semantic distance and affects the value of similarity and concrete calculation model as follows.

\[ sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \]  

(2)

Here, \( d \) denotes semantic distance of two concepts or sememes and \( \alpha \) is an adjustable parameter. The result of this method cannot receive a reasonable evaluation.

Li Feng’s semantic similarity algorithm that is based on Liu’s algorithm considers the depth of two sememes in the sememe tree. The computation model formula is

\[ sim(p_1, p_2) = \frac{\alpha \times \min(depth_{p_1}, depth_{p_2})}{d + \alpha \times \min(depth_{p_1}, depth_{p_2})} \]  

(3)

Here, \( d \) is semantic distance of two sememes and \( \min(depth_{p_1}, depth_{p_2}) \) is minimum depth of two sememes respectively. Also, \( \alpha \) is an adjustable parameter.

Based on the shortage of the above analysis algorithm and the influence factors of similarity computation, we present the algorithm of semantic similarity of two concepts or sememes which considers semantic distance, the depth of concept and information content of concept. The specific model is given as follows.

\[ sim_{improved}(p_1, p_2) = 1 - \alpha \times \left( \frac{\log(len(p_1, p_2) + 1)}{\log(2 \times \max_{p \in \text{docset}}(depth(p)) - 1)} \right) - (1 - \alpha) \times \text{dist}(p_1, p_2) / 2 \]

\[ = 1 - \alpha \times \left( \frac{\log(len(p_1, p_2) + 1)}{\log(2 \times \max_{p \in \text{docset}}(depth(p)) - 1)} \right) - (1 - \alpha) \times (IC(p_1) + IC(p_2) - 2 \times IC(\text{Iso}(p_1, p_2))) / 2 \]  

(4)
The concept is described by sememe and sememe relation in HowNet. So, there are some sememes in each concept and the kind of sememe is different. There are two categories of sememe in terms of Li Feng’s algorithm in HowNet. One is direct sememe, that is sememe without sememe relation symbol. Another is indirect sememe, that is sememe with sememe relation symbol. The concrete conceptual calculation model is

\[ sim(c_1, c_2) = \beta \times sim(p_{11}, p_{21}) + (1 - \beta) sim(p_{12}, p_{22}) \]  

\[
\beta = \begin{cases} 
0 & \text{(if there is no indirect sememe of } c_1, c_2) \\
0.1 & \text{(if there is single indirect sememe of } c_1, c_2) \\
0.15 & \text{(if there are some indirect sememes of } c_1, c_2) 
\end{cases}
\]

Our method solves the questions that are unsolved in Liu and Li algorithms. For example, we consider some factors that are location depth, regional density, semantic distance, information content and so on. Therefore, the performance of improved semantic similarity model has been greatly improved in term of parts 4 experimental results.

**EXPERIMENTS AND ANALYSIS**

In consequence of improving algorithm in the experiment, this paper makes full use of the words contained in HowNet, WordNet and CiYin semantic resources. In this paper, we employ test data set that is the same as Liu’s and Li’s, in order to compare the results in each other.

In order to facilitate comparing, we put the values of similarity of four algorithms on a figure. So, we can see obviously effect. The result is given in figure1 as follows.

In order to verify the performance of improved algorithm in this paper on specific data, we use the second method that compute the value of correlation coefficients between semantic similarity of four algorithms and semantic similarity of artificial measured respectively. The correlation coefficient of the calculation formula is as follows.

![Figure 1. The comparison of similarity of four algorithms.](image-url)
TABLE 1. THE CORRELATION COEFFICIENT OF SIMILARITY BETWEEN FOUR ALGORITHMS.

<table>
<thead>
<tr>
<th>Method</th>
<th>Value of Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu Qun</td>
<td>0.760</td>
</tr>
<tr>
<td>Li Feng</td>
<td>0.842</td>
</tr>
<tr>
<td>Liu Qinglei</td>
<td>0.862</td>
</tr>
<tr>
<td>This paper</td>
<td>0.870</td>
</tr>
</tbody>
</table>

\[ \rho_{xy} = \frac{Cov(X,Y)}{\sqrt{D(X)\sqrt{D(Y)}}} \]  \hspace{1cm} (6)

\(Cov(X,Y) = E((X-E(X))(Y-E(Y)))\), \(D(X)\) and \(D(Y)\) denotes variance respectively.

From the above table 1, we can get the conclusion that performance of algorithm in this paper is better than the others, because the value of correlation coefficient between the similarity of algorithm and artificial measured is maximal among all the algorithms.

CONCLUSION

In this paper, we introduce a structure of HowNet and data resources that are WordNet and TongYiCiCiLin to help understand an important role of semantic similarity. We present improved model of information content calculation and use it in the similarity computational process. The approach of improved similarity in this paper focuses on using impact factors of similarity computation to improve the existing computational model of similarity and obtain better result. We plan to apply the method of similarity calculation to ontology learning and develop techniques to mine domain knowledge on the unstructured text.

ACKNOWLEDGEMENTS

The correspondence authors are Guohua Lei and Lijuan Diao. This research is supported by the education of Shandong province "in 12th Five-Year 2015 year" plan of higher education computer teaching project NO. ZBJ15005 and the scientific research fund of Binzhou Medical University with Grant No. BY2015KYQD07. The work is supported by College of Social Science project in Shandong Province NO. J15WB08.

REFERENCES