CSCanada
**Management Science and Engineering**
Vol. 9, No. 1, 2015, pp. 25-29
**DOI:** 10.3968/6426

ISSN 1913-0341 [Print]
ISSN 1913-035X [Online]
www.cscanada.net
www.cscanada.org

# An Improved Semantic Similarity Algorithm on Hownet

LIAO Kaiji[a],*; BI Yingying[a]

[a]Research Center of Knowledge Management and Enterprise Information, South China University of Technology, Guangzhou, China.
*Corresponding author.

Received 12 November 2014; accepted 26 February 2015
Published online 26 March 2015

## Abstract
Semantic similarity algorithm is one of the basic researches in the field of natural language processing. This algorithm is widely used in information retrieval, machine translation based on examples and other fields. In this paper, based on the basis of HowNet lexical semantic similarity algorithm, introduced the concept of fuzzy mathematics degree of membership, the fixed weighting factor assigned into a coefficient of variation based on statistics through experimental verification of the results of this improved contribution.

**Key words:** Semantic similarity; HowNet; Fuzzy mathematics; Membership

Liao, K. J., & Bi, Y. Y. (2015). An Improved Semantic Similarity Algorithm on Hownet. *Management Science and Engineering*, *9*(1), 25-29. Available from: URL: http://www.cscanada.net/index.php/mse/article/view/6426
DOI: http://dx.doi.org/10.3968/6426

## INTRODUCTION

The words of natural language have a very complex relationship. In some areas, it is necessary to measure this complex relationship with a particular number, for example, the recommendation system, the Sentiment Analysis and so on. There are many basic researches among the area; the semantic similarity calculation of words is one of them. This paper focuses on the algorithm of words semantic similarity calculation.

Semantic similarity has been used in many areas such as recommendation (Gong, 2011), schema matching (Jeong, 2008), information retrieval, information extraction, text classification, word sense disambiguation, machine translation and so on. The research of semantic similarity calculation can be roughly divided into two categories: One is dependent on the semantic dictionary of hierarchical relationships structure between concepts, these methods are mainly based on the concept relationship between the hypernym-hyponym and synonymous words, using the distance of concept to calculate; the other is based on a large-scale corpus statistical method, the probability distribution of the vocabulary of context information as semantic similarity (Ge, 2010). Liu and Li (2002) proposed HowNet semantic similarity algorithm in the process of similarity calculation utilizing HowNet as semantic dictionary .This algorithm is widely used in the field of natural language processing. This method based on the theory of the overall similarity equals to the weighted average of the portion similarity. Firstly, this method broke down the whole into parts, and then any two parts are combined. Thirdly, by calculating the similarities of each combination to get the overall weighted average similarity.

Christopher et al. (2006) established an OWL concept semantic similarity measure model based on vague comparison relying on psychology similarity model, which is called FuzzyCon models. The feature set of concept is comprised by the significant implication set, the implicit implication set and the role of connected concepts. He uses fuzzy set to construct the fuzzy intersection and fuzzy set difference of concept feature set. The similarity function between concepts is used to construct the membership function of the fuzzy intersection and fuzzy set difference. Then he calculates the similarity according to the semantics of owl DL.

Pan (2010) researches the mode matching strategy based on WordNet. There are three types of similarity

calculation. They are the similarity of linguistics, the similarity of structure and the similarity of instance. He distinguishes these three similarities on a threshold into three categories: match, match to some extent and does not match. Finally, he conducted a fuzzy comprehensive evaluation according to the weight that was determined by the experts. Where in this process, he used the statistical approach to calculate the similarity for some factor, which we call similarity, belongs to a particular category.

The concept of "HowNet" is initiated and created by Dong Zhendong, who is a famous machine translator. HowNet's object is the concept of Chinese and English words. HowNet is a commonsense knowledge database that reveals the relationship of concepts as well as concept attributes. It focusses on the commonalities and individualities of concepts. Sememe is a minimum meaning units of describing senses. It is also the minimum meaning units of HowNet. There are eight relationships between sememes. They are hyponymy, antonymy, relations of justice, synonymy, property - host relationship, part - whole relationship material - Finished relations, event - role relationship. The first three are the most important between them. Sense is regarded as the minimum meaning units in traditional Lexicology. Sense is the meaning of a word. A word can have a sense or many senses while a sense can be interpreted by one or more sememes in HowNet.

Wang (2011) proposed a variation coefficient sense calculation using HowNet through the relationship of words, sememes and senses according to the order from inside to outside from the level of senses. This method relies on the number of sememes in the senses and introduces a variable to replace the manually specified proportionality coefficient. It reduces the influence of subjective factors and gets the more accurate results. What's more, in the process of word similarity calculation he considered the sense speech and reputed that the similarity between two senses of different speech is small which can be ignored. This method greatly reduces the amount of computation to improve the computing speed.

This paper is based on HowNet. And the concept of fuzzy sets in fuzzy mathematics is introduced in the calculation of sense similarity. We use membership function to represent the weight proportion of senses. This will be more objective comparing to the manual specified weights.

## 1. SEMANTIC SIMILARITY ALGORITHM BASED ON HOWNET

There are many different algorithms about semantic similarity, such as Ana G's algorithm based on graphic that considers both the hierarchical and non-hierarchical structure of ontology (Maguitman et al., 2006). Also, Lixin Han proposed a method to consider the contextual similarity which makes the measure more accurate (Han et al., 2006). Janowicz et al. (2008) pointed that the explanation of similarity values was important.]So we firstly give an overview of fundamental theory.

This paper relies on the algorithms put forward by Liu Qun. Therefore, an overview of Liu Qun's study is necessary. As has been mentioned in the Introduction, this method considered that the overall similarity equals to the weighted average of the portion similarity. Firstly, this method broke down the whole into parts, and then any two parts are combined. Thirdly, by calculating the similarities of each combination to get the overall weighted average similarity. As a result, we must classify the calculation of portion similarity. In HowNet, Words similarity calculation is to first calculate the sememe similarity, and then merged into sense similarity, the last words similarity are calculated.

### 1.1 The Sememe Similarity

"HowNet" is a world knowledge database having network structure. The sememe similarity mainly uses the hypernym-hyponym relationships between sememes to shape a tree-structure of sememe hierarchy and calculates through the relationship of sememe node. Many scholars conducted researches in this area and proposed their own formula. The formula according to Liu Qun 's study is as follows:

$$\mathrm{Sim}(s1, s2) = \partial / (\partial + \mathrm{dist}(s1, s2)).$$

In this formula Sim ($s1$, $s2$) represents the sememe similarity; $S1$ and $S2$ represent the two different sememe; dist ($S1$, $S2$) represents the distance between them in sememe tree; $\partial$ is an adjustable parameter, indicating the path length when similarity is 0.5, and generally its value is 1.6.

This method considers only the hypernym-hyponym relationship between the sememe. It is a simplified calculation method.

### 1.2 The Sense Similarity

The description of senses is divided into four categories by Liu Qun: the first basic sememe description, other basic sememe description, relationship sememe description and relationship symbol description.

Assuming that sense $C1$ and $C2$ has m and n sememes differently. We describe this using the following formula:

$C1 = \{S11, S12, ..., S1m\}$, $C2 = \{S21, S22, ..., S2n\}$; $n \leq m$.

The first basic sememe similarity is recorded as Sim1 ($S1$, $S2$). And other basic sememe similarity is denoted as Sim2 ($S1$, $S2$). Relationship symbol is recorded as Sim3 ($S1$, $S2$). Relationship sememe is recorded as Sim4 ($S1$, $S2$).

Then the overall similarity of two senses are recorded as $\mathrm{Sim}(C1, C2) = \Sigma \beta i \mathrm{Sim}i\ (S1, S2)$.

Wherein, $\beta i$ $(1 \leq i \leq 4)$ is an adjustable parameter, designated generally based on experience, and there

$\beta1 + \beta2 + \beta3 + \beta4 = 1$, $\beta1 \geq \beta2 \geq \beta3 \geq \beta4$. Since the first basic sememe description reflects the main features of a concept, so the weight is generally 0. 5 or above.

### 1.3 Words Similarity

Words is formed by senses. So words similarity can be calculated by sense similarity. Qun Liu considered that word similarity was the maximum similarity of senses (concepts).

Suppose there are two words $W1$ and $W2$, where the number of W1 senses is $m$, and the number of W2 senses is $n$, the following formula describes this relationship:

$W1 = \{C11, C12, ..., C1m\}$; $W2 = \{C21, C22, ..., C2n\}$.

Word similarity of W1 and W2 is the maximum senses similarity of $C1i$ and $C2j$ combinations. Word similarity calculation uses a maximum matching method, as the following formula:

Sim ($W1$, $W2$) = Max (Sim ($C1i$, $C2j$)).

To specify, $i$ and $j$ is a regular number between 0~$m$ and 0~$n$.

## 2. INTRODUCING THE CONCEPT OF FUZZY SETS

It is very subjective for the senses similarity calculation using the method mentioned above. Also, it gives high weight for the first basic sememe which may lead to a greater gap between the calculated results and the actual situation. Lin, Xue Fang etc. considered that the distinction role of the first basic sememe has gradually weakened as the deeper of senses explanation. Furthermore, the definition of first basic sememe is general and board. They are usually located in the root of sememe tree. In this situation, it is inappropriate to give the first basic sememe higher weight (Li, Xue, & Ren, 2009).

Such as the description for "doctor": DEF = human, occupation, cure, medical.

From the above description, we can know this sense are "people" category, followed by the "occupation" related and he implements "cure" for some objects as well as related to "medical". We know a whole definition of doctor with the depth of definition layers. Comparing to the first basic sememe, the last three sememes become the decisive factor of the concept doctor distinguished with other human. If we give the first basic sememe higher weight in this situation, similar senses similarity will be larger than the actual situation. On the other hand, the subjective factor of HowNet have large influence on the first basic sememe. For the definition of "Gem", "stone" is the first basic sememe. This is very different from the popular understanding of "Gem" which is regarded as valuables. If we give the first basic sememe higher weight in this situation, similar senses similarity will be smaller than the actual situation.

DEF = stone, treasure.

To overcome the shortcoming of artificial weight decision, we introduce the membership intermediate of fuzzy mathematics. Fuzzy set is the basis of fuzzy mathematics. Zadeh, who is the famous expert of Cybernetics, extended the the range of ordinary set theory in characteristic function from {0,1} to the closed interval [0,1], by which we mean classes in which there may be grades of membership intermediate between full membership and non membership (Zadeh, 1968) .That meant the forming of fuzzy set definition (Yang, Gao, & Ling, 2011).

Supposing that there is a mapping in the domain $U$

$A$: $U$ [0,1]

$U$ A ($u$)

$A$ is called a fuzzy on $U$. $A$ ($u$) is called the membership function of $A$ (or $U$ membership degree of $A$).

According this, we define $\beta i = ki / (m + n)$,

Wherein, $ki$ is the sememe number of class i; m, n is the total sememe number of respectively $C1$ and $C2$.

In this formula, we can find that as the depth of sense definition, the node level becomes greater. The first basic sememe weight becomes smaller and the influence on the sense similarity becomes smaller as a result and vice versa.

We use the maximum matching principle for the similar sememe combination, which is matching the two sememes that have maximum similarity. In order to facilitate the calculation, this article does not consider the antisense and justice relationships (Jiang et al., 2008), and the following is agreed (Lei & Gu, 2010):

a) The similarity of sememe and space takes a smaller constant 0.01;

b) For the case of direct description using words, if the two words are the same, the similarity is 1, and 0 otherwise. The similarity between words and sememe takes a small constant 0.01;

c) For the symbolic sememe, only when the symbol is the same can we calculate the similarity. Otherwise it is set to a small constant 0.01;

d) For the relationship sememe, only when the relationship is the same can we calculate the similarity. Otherwise it is set to a small constant 0.01;

e) For the sememe that is not the same tree, the similarity is set to a small constant 0.01.

Compared with fixed Coefficient sense similarity algorithm, this method improves the objectivity and accuracy. And the results match with the people's intuition.

## 3. EXPERIMENTAL RESULTS AND ANALYSIS

For comparison, we use the experimental data of literature (Liu & Li, 2002; Li et al., 2007). After getting the similarity

results of the introduction of the coefficient of variation, we get the experimental data in the following table.

**Table 1**
**Word Similarity of Different Algorithm**

| Word1 | Word2 | Literature (Liu & Li, 2002) | Literature (Li et al., 2007) | Improved algorithm |
|---|---|---|---|---|
| Man | Father | 1 | 1 | 1 |
| Man | Monk | 0.833 | 0.921 | 0.815 |
| Man | Manager | 0.657 | 0.53 | 0.474 |
| Man | Happy | 0.013 | 0.191 | 0.064 |
| Man | Radio | 0.164 | 0.159 | 0.081 |
| Man | Carp | 0.208 | 0.357 | 0.179 |
| Man | Apple | 0.166 | 0.313 | 0.148 |
| Man | Work | 0.164 | 0.148 | 0.09 |
| China | United Nations | 0.136 | 0.125 | 0.271 |
| Pink | Crimson | 0.074 | 0.7 | 0.576 |
| Troops | Expedition | 0.105 | 0.307 | 0.383 |
| Run | Jump | 0.444 | 0.762 | 0.444 |
| China | America | 0.94 | 0.625 | 0.75 |
| Invention | Create | 0.615 | 0.849 | 0.615 |

(a) We can find that the fourth column result is better coincidence with people's normal understanding of words by comparing the third and fourth columns. This is mainly because in the calculation of the third column only the hypernym-hyponym relationship is considered. The result is rough because other relationship is ignored. Otherwise, the forth column takes the depth of every sememe into consideration. Make full use of the "HowNet" and the performance improvement is significant.

(b) Comparison of the two column 3 and 5, column 5 have varying degree increase. As the deeper definition of HowNet, the increase becomes more obvious, whereas obscure. Such as the "Pink" and "Crimson":

Pink DEF = a Value, color, red.

Crimson DEF = attribute, color, red, & physical.

In people's intuitive understanding, "pink" and "Crimson" are representing words of color, They represent a kind of property. However, because of the difference of the first basic sememe, these two words are assigned into different categories in HowNet. If we still give a higher weight to the first basic sememe, a smaller similarity will be get which do not correspond to the intuitive understanding of people. In column 5, we take the sememe number into account to reduce the weight of the first basic sememe and enhance the overall similarity of senses. Also, the result is more in line with people's intuition.

(c) The similarity of "man" and "father" are 1 in all three methods. This is mainly because the definitions of these two words in HowNet are the same, with the improvement of HowNet, this situation will improve.

## CONCLUSION

The semantic similarity calculation on HowNet has been widely used. This paper introduces the membership concept of fuzzy mathematics based on the Liu Qun's study. And this paper proposed the variable coefficient of senses similarity. We use the proportion of different types sememe number as weight to calculate senses similarity. This method reduces the impact of the first basic sememe and enhances the impact of other sememes. We can get the conclusion from the experimental data that the improved algorithm improves the objective of semantic similarity calculation to some extent. The results of this algorithm correspond with people's intuition. As the development of HowNet, the classification of sememe becomes more reasonable and detailed. The effect of this algorithm will be more obvious.

In a subsequent study, (a) we can be combined with questionnaires to get a more objective institutions; (b) compare different algorithms, from a practical point of view to explain the improvement of introduction of variable coefficient. (c) apply this method in the relevant research.

## REFERENCES


Christopher, D., & Manning, P. R. (2006). *An approach for measuring semantie simllarlty between words using multipleinformation source retrieval.* Cambridge University Press, Cambridge, England.

Ge, B. (2010). The study of word semantic similarity calculation method based on HowNet. *Application Research of Computer,* (9)

Gong, S. J. (2011). A personalized recommendation algorithm on integration of item semantic similarity and item rating similarity. *Journal of Computers,* (6), 1047-1054.

Han, L. X., Sun, G. H., & Chen, L. X. (2006). An approach to determining semantic similarity. *Advances in Engineering Software, 37,* 129–132

Janowicz, K., Raubal, M., Schwering, A., & Kuhn, W. (2008). Semantic similarity measurement and geospatial applications. *Transactions in GIS, 12*(6), 65-659.

Jeong, B. W., Lee, D. W., Cho, H. B., & Jaewook Lee, J. (2008). A novel method for measuring semantic similarity for XML schema matching. *Expert Systems With Applications, 34,* 1651-1658

Jiang, M., Xiao, S. B., & Wang, H. W., et al. (2008). An improved word semantic similarity based on HowNet. *Chinese Information Technology, 2*2(3), 84-89.

Lei, L. Q., & Gu, X. F. (2010).Words similarity algorithm based on HowNet. *Chinese Information Technology, 24*(6), 31-36.

Li, L., Xue, F., & Ren, Z. S. (2009). An improved word similarity calculation method based on HowNet. *Computer Applications, 29*(1), 217-220

Li, F., & Li F. (2007). Chinese word semantic similarity calculation based on HowNet. *Chinese Information Technology, 2*(3), 99-105.

Liu, Q., & Li, S. J. (2002). *Papers: Word similarity computing based on how-net* (pp.59-76). The third Chinese lexical semantics seminar proceedings. Taipei: [s. n. ].

Maguitman, A. G., Menczer, F., Erdinc, F., Roinestad, H., & Vespignani, A. (2006). Algorithmic computation and approximation of semantic similarity. *World Wide Web, 9,* 431-456.

Pan F. (2010). Matching strategy based on multi-mode fuzzy decision. China: Shandong University, School of Computer Science and Technology.

Wang, X. L. (2011). An improved words similarity algorithm based on Hownet. *Computer Applications,* (11).

Yang, L. B., Gao, Y. Y., & Ling, W. L. (2011). *Fuzzy mathematical theory and application*. Guangzhou: South China University of Technology Press.

Zadeh, L. A. (1968). Fuzzy algorithms. *Information and Control, 12,* 94-102.

 