International Journal of Hybrid Information Technology
Vol.8, No.10 (2015), pp.417-426
http://dx.doi.org/10.14257/ijhit.2015.8.10.38

A Word Similarity Algorithm with Sememe Probability Density Ratio Based on HowNet

Rui Zheng, Huan Zhao* and Xixiang Zhang

School of Information Science and Engineering,
Hunan University, Changsha, 410082, China
hzhao@hnu.edu.cn

Abstract

The study on word similarity computation plays an important role in natural language processing (NLP). Recently the algorithm based on HowNet is widely used and proves to work well in Chinese word similarity computation. However, the relationship between the number of brother nodes and the fineness of the hierarchy is not considered. This paper investigates the ratio of two words on the brother nodes' number called sememe probability density and proposes an improved algorithm based on HowNet. The results indicate that the correlation measure of the algorithm presented by this paper is 75.4%, and it is much better than the major state-of-the-art method (68.1%).

Keywords: *word similarity, HowNet, sememe probability density*

1. Introduction

The research of word similarity is the foundation of natural language processing (NLP) and information retrieval(IR). It aims to measure complex semantic similarity between words and has been applied in text classification, question-answering, example-based machine translation, *etc*. [1-3].

Recently, the method based on ontology is the mainstream approach. It utilizes the semantic dictionary, calculates concepts' semantic distances in a tree-structured hierarchy and then obtains the word similarity. Many researchers have carried on a large number of studies and made some achievements [4-6]. On the basis of WordNet, Rada [7] presented an approach according to the concepts' shortest path constituted by hyponymy between concept nodes. Yang&Powers [8] considered about the association path in WordNet including whole/part, upper/lower, synonym/antonym and proposed the weighted similarity computation model based on semantic relations. Resnik [9] took the advantage of the theory that the more information the concepts share the greater the similarity is and proposed an approach based on information content. It relied on hyponymy in the semantic dictionary and probability model, converted the word similarity computation to solve the maximum public information.

For Chinese, Liu [10] adopted the idea that the overall similarity equals to the weighted sum of each part's similarity and presented a Chinese word similarity approach based on HowNet. Ge [11] took the tree's depth into account and weighted relation sememe and relation symbol. F. . Hu [12] introduced the fact of the least common node and merged the first and second sememes to make better.

This paper adopts the idea that the more brother nodes a sememe has, the finer a hierarchy is, then introduces the sememe probability density of two words and proposes an improved algorithm based on HowNet. The rest of this paper is organized as follows. We firstly review related works in Section 2. Then the details of our method are presented in Section 3. We evaluate the method in Section 4 and make a conclusion in Section 5.

ISSN: 1738-9968 IJHIT

2. Related Work

From the perspective of different semantic similarity theory, word similarity computation method based on ontology can be mainly divided into two: distance-based and information-based.

A distance-based approach is directly to compute the similarity in the semantic dictionary. In a tree-structured hierarchy, any two nodes have a unique path, the length of which is viewed as the two nodes' semantic distance. Shorter the length is, greater the similarity is.

In the general case, the distance-based approach uses the formula below to compute the word similarity:

$$Sim(w_1, w_2) = f(dis(w_1, w_2))$$

Where w_1 and w_2 are two words, $dis(\blacksquare)$ denotes the semantic distance between two words, $f(\blacksquare)$ is a function that transforms the semantic distance to word similarity, $Sim(w_1, w_2)$ defines the word similarity between w_1 and w_2.

A information-based approach depends on the tree-structured hierarchy and probability model. It transforms concept's semantic similarity computation to concept information. The more information two concepts share, the greater the similarity is. In a tree-structured hierarchy, the largest amount of information is the similarity between two concepts.

More generally, the information-based approach uses the formula below to compute the word similarity:

$$Sim(w_1, w_2) = f(-\log p(w_1, w_2))$$

Where w_1 and w_2 are two words, $p(\blacksquare)$ denotes a monotone function which associates with w_1 and w_2, $f(\blacksquare)$ is a function that transforms the information to word similarity, $Sim(w_1, w_2)$ defines the word similarity between w_1 and w_2.

3. Algorithm

3.1 HowNet

HowNet is an online knowledge-base which reveals the relationship among concepts, and the relationship among attributes of concepts [10].Concept and sememe are two important concepts in HowNet. Concepts compose of HowNet knowledge base and each concept is made up by a set of finite sememes.

Concept is the description of lexical semantics. Every word can be expressed as a few concepts. Each concept description is semantic expressions which use the knowledge description language. The format is as follows in Figure 1 (take 'knit' for example).

The sememe is the smallest unit to describe a concept. In HowNet, there are 1621 sememes. These sememes are organized a sememe tree-structured hierarchy, which is the base of word similarity computation. Part of the hierarchy of entity sememe tree is shown in Figure 2.

The original semantic structure of HowNet makes word similarity computation possible .Now there are a lot of algorithms presented to compute word silimarity [11-12]. The algorithm's basic process is showed as Figure 3. Through further analysis and comparison of these algorithms, we finds that they exist some problems such as tree hierarchical model or relationships between the concepts. This paper adopts the idea that the more brother nodes a sememe has, the finer a hierarchy is, then introduces the sememe probability density of two words and proposes an improved algorithm based on HowNet. Compared to other algorithms, it makes an obvious improvement.


```
NO. = 015492
W_C = Chinese word string
G_C = V
E_C = example of Chinese word (optional)
W_E = knit
G_E = V
E_E = example of English word (optional)
DEF = weave
```

Figure 1. Format of Knowledge Description Language

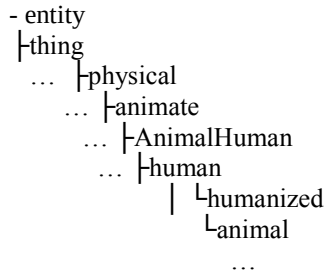

Figure 2. Hierarchy of the Entity Sememe Tree

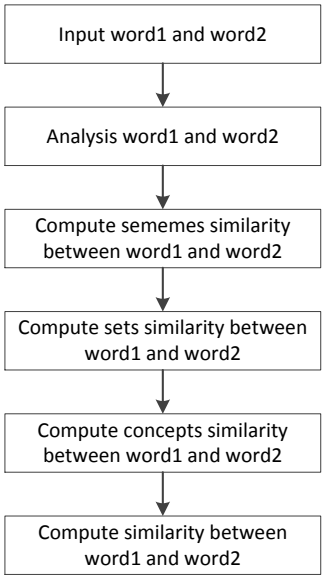

Figure 3. Basic Process of the Word Similarity Algorithm

3.2 Similarity between Sememes

The sememe is the smallest unit to describe a concept and also the basic of word similarity computation. Word similarity computation lies in the sememe similarity computation. When the sememe similarity computation gets a better accuracy, so will the word similarity computation be. Otherwise, the word similarity computation becomes worse.

Liu [10] took the hyponymy into consideration and proposed a formula to compute similarity between sememes as follow:

$$\mathrm{Sim}(p_1, p_2)=\frac{\alpha}{\alpha+\mathrm{Dis}\ (p_1, p_2)}$$

Where p_1 and p_2 respectively denote sememe1 and sememe2, $\mathrm{Dis}(p_1, p_2)$ defines the distance between p_1 and p_2, α is a fixed adjustment parameter, it is the distance between p_1 and p_2 when the similarity equals 0.5, $\mathrm{Sim}(p_1, p_2)$ is the similarity between p_1 and p_2.

However, formula2 neglects the large amount of semantic information and structural feature in HowNet. At the same time, the fixed parameter cannot achieve a good calculation result. On this basis, Hu [13] replaces the fixed parameter by least common node and the formula is shown below:

$$\mathrm{Sim}(p_1, p_2) = \frac{\mathrm{LCN}}{\mathrm{LCN} + \mathrm{Dis}(p_1, p_2)}$$

Where p_1 and p_2 respectively denote sememe1 and sememe2, $\mathrm{Dis}(p_1, p_2)$ defines the distance between p1 and p2, LCN is the least common node of p_1 and p_2, $\mathrm{Sim}(p_1, p_2)$ is the similarity between p_1 and p_2.

This paper adopts the idea that the more brother nodes a sememe has, the finer a hierarchy is, then introduces the sememe probability density of two words and proposes an improved algorithm based on HowNet, the formula is as follows:

$$\mathrm{Sim}(p_1, p_2) = \frac{\mathrm{LCN}}{\mathrm{LCN} + \rho * \mathrm{Dis}(p_1, p_2)}$$

$$\rho = \begin{cases} \frac{\lambda_1}{\lambda_2}, \lambda_1 < \lambda_2 \\ \frac{\lambda_2}{\lambda_1}, \lambda_1 \geq \lambda_2 \end{cases}$$

where p_1 and p_2 respectively denote sememe1 and sememe2, $\mathrm{Dis}(p_1, p_2)$ defines the distance between p_1 and p_2, LCN is the least common node of p_1 and p_2, λ_1 and λ_2 respectively denote the brother nodes of λ_1 and λ_2, p is the sememe probability density, defined as the ratio of λ_1 and λ_2, $\mathrm{Sim}(p_1, p_2)$ is the similarity between p_1 and p_2.

3.3 Similarity between Sets

A concept is composed by a series of sememes. Similarity between sets is the basic of similarity of two concepts. The steps are as follows:

Step1: compute the similarity of any two elements in two sets;

Step2: select the maximum, establish corresponding relations of the two elements;

Step3: delete the similarity which has been established corresponding relations;

Step4: loop step2 and step3, until delete all similarity;

Step5: the rest which do not establish corresponding relations is associated with null value, then two elements in the sets are established a one-to-one relationship. We define the average of similarity as similarity between sets.

3.4 Similarity between Concepts

As the concepts in HowNet is described by knowledge dictionary descript language, similarity between concepts is equivalent to compute concept expression. The concept expression is generalized by 4 parts:

The first independent sememe expression: its value is a first independent sememe, formula5 is used to compute similarity of this part, denoted by $\mathrm{Sim}_1(c_1, c_2)$.

Other independent sememe expression: all independent sememe except the first independent sememe, its value is a set of independent sememe, the algorithm to compute similarity between sets is used, denoted by $\mathrm{Sim}_2(c_1, c_2)$.

The relational sememe expression: correspond to expression of relational sememe, its value is a feature structure, for each feature, there is a relational sememe which may be a first independent sememe or a specific word. Formula 5 is used to compute similarity of this part, denoted by $\mathrm{Sim}_3(c_1, c_2)$.

The symbol sememe expression: correspond to expression of symbol sememe; its value is a feature structure, for each feature, there is a symbol sememe which may be a set of first independent sememes or specific words. The algorithm to compute similarity between sets is used, denoted by $\mathrm{Sim}_4(c_1, c_2)$.

The similarity between concepts is weighted by 4 parts:

$$Sim(c_1, c_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(c_1, c_2)$$

Where $\beta_i (1 \le i \le 4)$ is the weight factors, in consequence that the first independent sememe reflects the main feature of the concept, β_i is usually greater than 0.5,and $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$.

3.5 Similarity between words

Two words in HowNet as $w1$ and $w2$, we assume that $w1$ has n concpets : $c_{11}, c_{12} \ldots c_{1n}$, $w2$ has m concpets: $c_{21}, c_{22} \ldots c_{2n}$.The similarity between two words $w1$ and $w2$ is defined as the maximum similarity in all two concepts, the formula is showed below:

$$\mathrm{Sim}(w1, w2) = \max_{i=1..n,, j=1..m} Sim(c_{1i}, c_{2j})$$

Where $Sim(c_{1i}, c_{2j})$ denotes similarity between concepts from different words, it can be computed by formula7, then we can achieve the similarity between words.

4. Evaluation

4.1 Data Set and Setting

Miller-Charles [13] presented a test set of English words which has been widely used to evaluate English word similarity. Zhao [14] translated it into Chinese to re-evaluate it and made a Chinese version (Chinese M&C). He [15] worked on it and made a re-evaluation to get a better. Result this paper use Chinese M&C to make a evaluation.

To ensure the objectivity of the experimental evaluation, we take the correlations between the similarity computed by algorithms and the similarity by hand-marked sequences as the experiment evaluation standard. The Pearson is used to compute the correlations:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) S_x S_y}$$

where $\bar{x}$ and $\bar{y}$ respectively denote the average of samples x and samples y, x_i and y_i respectively denote the values of samples x and samples y, n is the total number of samples, S_x and S_y are the standard deviations of samples x and samples y, r is the correlations of samples x and samples y.

Some parameters are defined as follows in this paper: the similarity between a sememe and a specific word is a constant ($\gamma = 0.2$);the similarity between non-null and null is a constant($\delta = 0.2$);the distance between two sememes which are not in the same sememe tree is 20;the weighted parameters in concept similarity are: $\beta_1 = 0.5$, $\beta_2 = 0.2$, $\beta_3 = 0.17$, $\beta_4 = 0.13$

4.2 Experimental Results

The major state-of-the-art methods proposed by Liu [10] and Hu [12] are used to compare the experiment results, as showed in Table 1. In this Table, the first column is ordinal number of the word pair, the second column is the M&C English pair, the third column and the forth column are the results of Liu [10] and Hu [12], the last column is the result of this paper.

To evaluate objectively, we chose three hand-marked sequences: Miller-Charles [13] (hand-marked sequences1) 、Zhao [14] (hand-marked sequences2)、He [15] (hand-marked sequences3). We compare the results of three methods with different three hand-marked sequences in Figure 4、Figure5 and Figure6. X-axis is the ordinal number, y-axis is the similarity, lines labeled with hand-marked sequences、Liu’sMethod、Hu’sMethod、Proposed Method respectively denotes the experiment result curve of hand-marked sequences、Liu [10]、Hu [12] and this paper. The comparison of results of three methods with hand-marked sequences1 is showed in Figure 4. The comparison of results of three methods with hand-marked sequences 2 is showed in Figure 5. The comparison of results of three methods with hand-marked sequences3 is showed in Figure 6.

Table 1. Comparison of Word Similarity Computation of Three Methods

No.	English Pair	Liu's Method	Hu's Method	Proposed Method
1	glass, magician	0.121	0.298	0.022
2	chord, smile	0.074	0.038	0.019
3	lad, wizard	0.600	0.448	0.579
4	Brother ,monk	0.661	0.533	0.480
5	forest, graveyard	0.112	0.198	0.574
6	monk, slave	0.722	0.667	0.600
7	coast, forest	0.112	0.112	0.216
8	Coast, hill	0.145	0.498	0.397
9	journey, car	0.074	0.048	0.024
10	crane, implement	0.369	0.512	0.578
11	lad, brother	0.800	0.778	0.800
12	bird, cock	1.000	1.000	1.000
13	food, fruit	0.166	0.305	0.577
14	Bird, jewel	1.000	1.000	1.000
15	food, fruit	0.126	0.222	0.596
16	gem, jewel	0.600	0.667	0.600
17	tool, implement	1.000	1.000	1.000
18	car, automobile	1.000	1.000	1.000
19	boy, lad	1.000	1.000	0.600
20	magician, wizard	0.676	0.911	0.948
21	journey, voyage	1.000	0.896	0.896
22	food, rooster	0.112	0.199	0.595
23	coast, shore	1.000	1.000	1.000
24	middy, noon	1.000	1.000	1.000

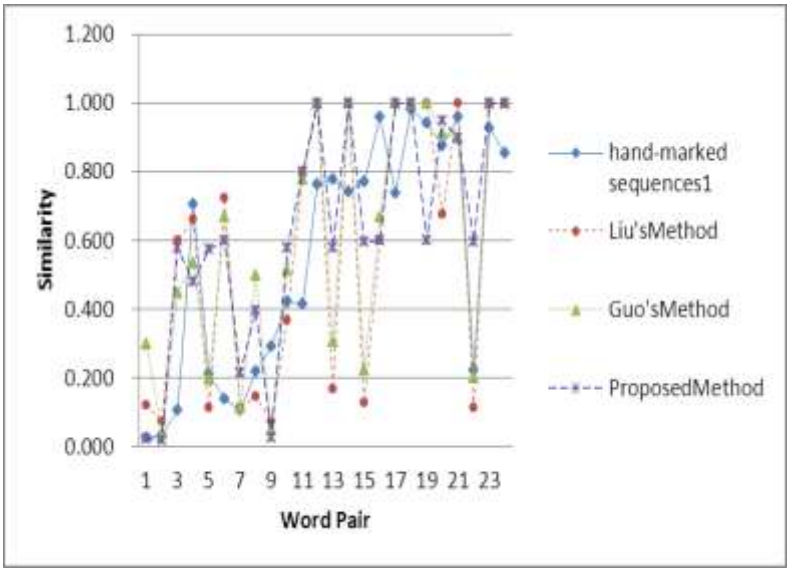

Figure 4. Comparison of Results of Three Methods with Hand-marked Sequences 1

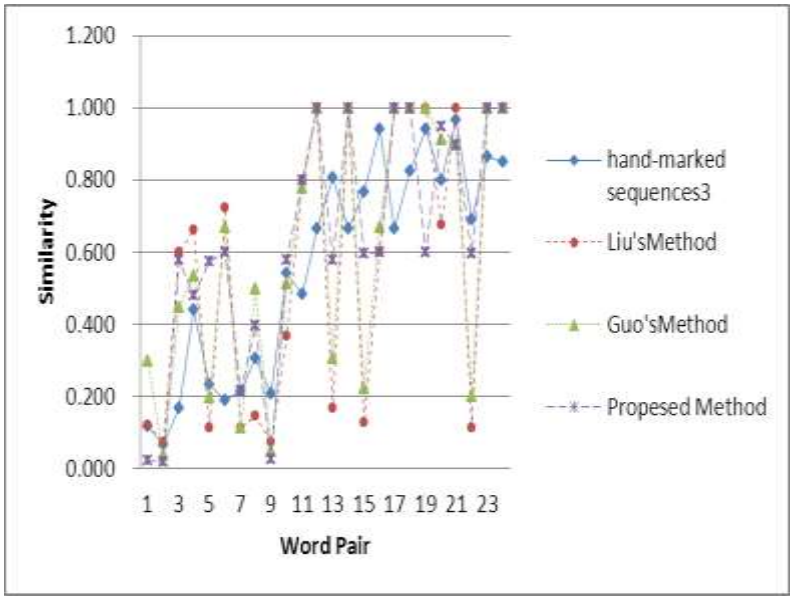

Figure 5. Comparison of Results of Three Methods with Hand-marked Sequences 2

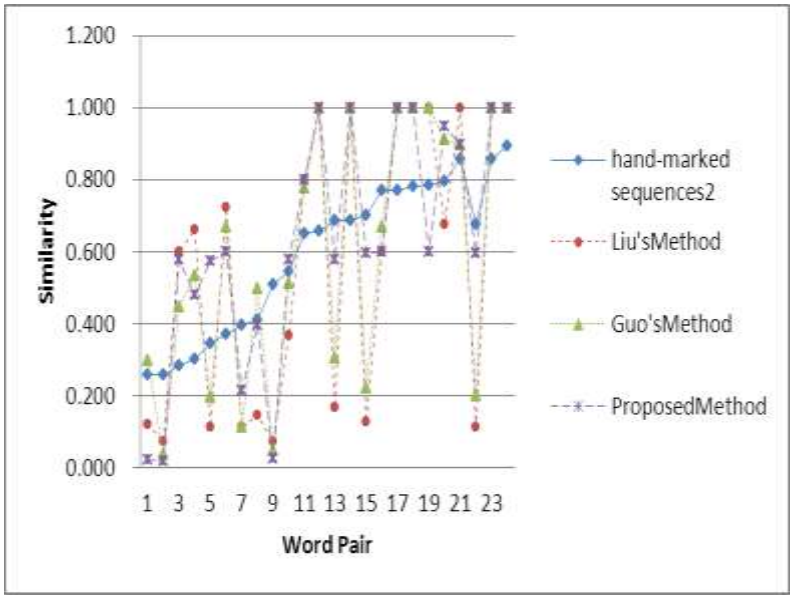

Figure 6. Comparison of Results of Three Methods with Hand-marked Sequences 3

In order to more intuitive and effective to evaluate the experiment results, the correlations are calculated between the similarity computed by the methods and the similarity by the three different hand-marked sequences, as Table 2. We take the results of Liu [10] as the baseline. From Table 2, the correlation between the similarity computed by this paper and the similarity by the hand-marked sequences1 is 0.757, which is higher than Liu [10] (0.715) and equals to Hu [13] (0.757).However, the correlations between the similarity computed by this paper and the similarity by the hand-marked sequences2 and the hand-marked sequences3 are 0.784 and 0.754, which are obviously higher than the other two. The bar graphs of correlations are showed in Figure 7. From Figure 7, the correlation between the similarity computed by this paper and the similarity by hand-marked sequences are much higher and are much closer to the hand-marked sequences to proof the method proposed by this paper to be best.

Table 1. Correlations of Word Similarity Computation and Hand-marked Sequences

Method	hand-marked sequences1	hand-marked sequences2	hand-marked sequences3
Liu'sMethod	0.715	0.635	0.609
Guo'sMethod	0.757	0.707	0.681
NewMethod	0.757	0.784	0.754

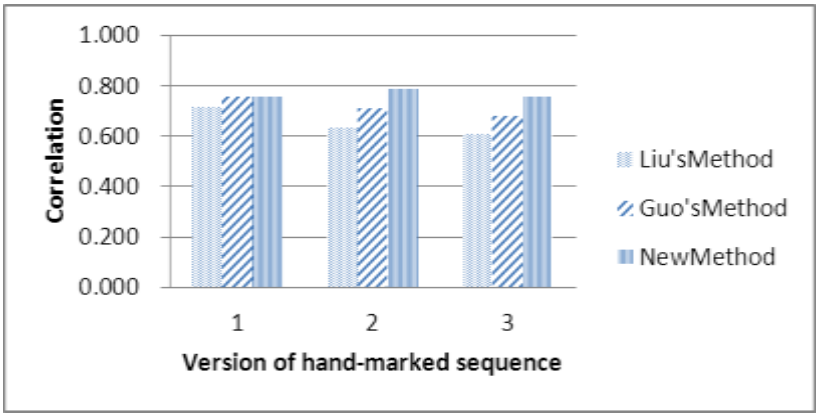

Figure 7. Correlations of Word Similarity Computation and Hand-marked Sequences

5. Conclusions

This paper studies on the original semantic structure of HowNet and algorithms of word similarity computation based on it. We investigate the ratio of two words on the brother nodes' number called sememe probability density and propose an improved algorithm based on HowNet. Finally, we take the correlation measure of the algorithm presented by this paper and compare it with Liu [10] and Hu [12]. The results indicate that the algorithm presented by this paper is much better than major state-of-the-art method. In the current algorithm, we only take advantage of the hyponymy in the semantic sememes. In future works, we will explore more sememe relationships do the influence to word similarity computation and study how to use the other sememe relationships to get a closer to the manual evaluation.

ACKNOWLEDGEMENTS

The research work was supported by National Natural Science Foundation of China under Grant No. 61173106.

References

[1] X. Chen, Y. Zhang, L. Cao and D. Li, "An Improved Feature Selection Method for Chinese Short Texts Clustering Based on HowNet", Computer Engineering and Networking: Springer, Berlin, **(2014)**.
[2] P.-Y. Zhang "A HowNet-Based Semantic Relatedness Kernel for Tex Classification", TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 11, no. 4, **(2011)**, pp. 1909-1915.
[3] F. Xianghua, L. Guo, G. Yanyan and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon", Knowledge-Based Systems, vol. 37, **(2013)**, pp. 186-195.
[4] J. Liu, M. Shi and G. Wang, "Research on automatic acquisition method of Chinese domain ontology backbone based on Hownet", International Journal of Wireless and Mobile Computing, vol. 7, no. 2, **(2014)**, pp. 147-152.
[5] F. Bond and R. Foster, "Linking and Extending an Open Multilingual Wordnet", ACL, **(2013)** August 4-9, Sofia, Bulgaria.
[6] P. Vossen, C. Soria and M. Monachini, "Wordnet- LMF: A Standard Representation for Multilingual Wordnets", LMF Lexical Markup Framework, **(2013)**, pp. 51-66.
[7] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and application of a metric on semantic nets. Systems, Man and Cybernetics", IEEE Transactions, vol. 19, no. 1, **(1989)**, pp. 17-30.
[8] D. Yang, "Powers DM, Measuring semantic similarity in the taxonomy of WordNet", Proceedings of the Twenty-eighth Australasian conference on Computer Science, vol. 38, **(2005)** January 27-30, Darlinghurst, Australian.
[9] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", Proceedings of the 14th international joint conference on Artificial intelligence, **(1995)** August 20-25, Montreal, Canada.

[10] Z. Dong and D. Qiang, "Construction of a knowledge system and its impact on Chinese research", Contemporary Linguistics, vol. 1, **(2001)**, pp. 3.
[11] B. Ge, F. F. Li, S. L. Guo, *et al.*, "Word's semantic similarity computation method based on Hownet", Jisuanji Yingyong Yanjiu, vol. 27, no. 9, **(2010)**, pp. 3329-3333.
[12] F. S. Hu and Y. Guo, "An improved algorithm of word similarity computation based on HowNet", Computer Science and Automation Engineering (CSAE), IEEE International Conference on, **(2012)** May 25-27, Zhangjiajie, China.
[13] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity", Language and cognitive processes, vol. 6, no. 1, **(1991)**, pp. 1-28.
[14] J. Zhao, H. Liu and R. Lu, "Attribute-base Computing of Word Similarity", The 11th China Conference on Machine Learning, **(2008)** August 18-21, Dalian, China.
[15] X. He, L. Liu and J. Wu, "Semantic similarity calculation based on sememe set[C]//Artificial Intelligence and Computational Intelligence (AICI)", International Conference on IEEE, **(2010)** May 3-8, Cape Town, South Africa.

Authors

Rui Zheng, (1990.3), female, Master Degree of Communication Engineering, Hunan University, her research interests include natural language processing and information retrieval. E-mail:S12101056@hnu.edu.cn.

