Semantic Similarity Computation Based on Multi-feature Combination using HowNet

Peiying Zhang
China University of Petroleum (East China)/ College of computer and communication engineering, QingDao, China
Email: 25640521@qq.com

Zhanshan Zhang
Tongji University / International School, Shanghai, China
Email: {zhanshan}@tongji.edu.cn

Weishan Zhang and Chunlei Wu
China University of Petroleum (East China)/ College of computer and communication engineering, QingDao, China
Email: {zhangws, wuchunlei}@upc.edu.cn

***Abstract*—Semantic similarity between words is becoming a generic problems for many applications of computational linguistics and artificial intelligence. The difficulty lies in how to develop a computational method that is capable of generating satisfactory results close to how humans perceive. This paper proposes a semantic similarity approach that is based on multi-feature combination. One of the benchmarks is Miller and Charles' list of 30 noun pairs which had been manually designated similarity measurements. We correlate our experiments with those computed by several other methods. Experiments on Chinese word pairs show that our approach is close to human similarity judgments.**

***Index Terms*—semantic similarity, semantic distance, similarity computing, HowNet**

I. Introduction

The research of semantic similarity between words has been an essential part of natural language processing and information retrieval for many years. Semantic similarity measurement plays an important role in natural language processing (NLP) and information retrieval (IR). It can be used to improve accuracy in an information retrieval task, to perform word-sense disambiguation (WSD), to discover mapping between ontology entities, to compute sentence similarity, and to estimate semantic orientation, etc.

A number of semantic similarity methods have been proposed in the previous decade. R. Rada [1] was motivated by the properties of spreading activation and conceptual distance, proposed a distance metric on the power set of nodes in a semantic net. P. Resnik [2] used information content to evaluate semantic synonymy. G. Hirst and D. St-onge [3] used lexical chains as representation of context for the detection and correction of malapropisms. Jay J Jiang and David W Conrath [4] proposed a semantic similarity method based on corpus statistics and lexical taxonomy. C. Leacock and M.Chodorow [5] proposed a method combing local context and WordNet similarity for word sense identification. Different similarity methods have proven to be useful in some specific applications of computational intelligence.

As a common sense, semantic similarity between words is influenced by contexts in which the words are used. For instance, if the context is "the outside covering of living objects", the skin and bark are more similar than skin and hair; however, the opposite is true if the context is body parts. We assume that the semantic similarity is symmetric, as experimental results investigating the effects of asymmetry suggest that the average difference in ratings for a word pair is less than 5 percent [6]. We do not consider the context sensitive effects to make our method not limited to some specific applications of computational intelligence.

Since the information content is calculated from the corpus, this similarity measurement can be adapted to a particular application provided that the corpus approximates that application area well. Different methods use different information sources and, thus, result in different levels of performance. The commonly used information sources in previous studies are shortest path length between compared words, information content; depth is the taxonomy hierarchy, and semantic density of compared words.

Some problems with these similarity methods are: information sources are directly used as a metric of similarity, a method uses a particular semantic metric without considering other factors. A hybrid algorithm is proposed to calculate semantic similarity based on multi-features combination using HowNet. We carried out some experiments on a benchmark set of word pairs and took human similarity judgments as baselines to evaluate our approach.

The rest of paper is organized as follows. First some related work is discussed in section 2, and then we introduce the HowNet and semantic similarity computation method based on multi-features combination in section 3. We evaluate the method in section 4 and conclude this paper in section 5.

doi:10.4304/jsw.9.9.2461-2466

II. Related Works

Until now, several measurements for computing similarity between words have been proposed in the past decades. These approaches can be roughly divided into three main categories, including the distance-based measurement, the information-based measurement, and the hybrid measurement which combines the first two measurements [14].

A. Distance-based Measurements

The main idea of distance-based measurement is to select the shortest path among all the possible paths between concepts. This measurement assumes that the shorter the distance, the more similar the concepts are. Distance-based measurements usually utilize some thesaurus taxonomy to compute the similarity.

In 1989, Rada et al [1] uses the minimum length of path connecting two concepts containing the words as a metric for measuring the semantic similarity of words. Their work forms the basis of distance-based similarity approaches.

Liu Qun [15] uses the shortest path between two primitive nodes to calculate the similarity, the formula is:

$$Sim(p_1,p_2)=\frac{\alpha}{dis\tan ce(p_1,p_2)+\alpha}$$

Where p1 and p2 denotes two primitives, distance(p1, p2) denotes the path length in the primitive hierarchy trees, α is a adjust parameter.

In 1994, Wu & Palmer measurement [7] calculates semantic similarity through considering the depths of two synsets, along with the depth the least common subsume (LCS) of two concepts in the WordNet taxonomy, denoted by:

$$sim_{WP}(c_1,c_2)=\frac{2\times depth(lcs(c_1,c_2))}{depth(c_1,)+depth(c_2)}$$

In 1998, Leacock & Chodorow measurement [5] (shortly LC measurement) calculates semantic similarity through considering the length of the shortest path that connects two concepts and the maximum depth of the WordNet taxonomy, denoted by:

$$sim_{LC}(c_1,c_2)=-\log\frac{len(c_1,c_2)}{2\times\max\limits_{c\in WordNet} depth(c)}$$

In summary, the distance-based measurement obviously requires a lot of information on detailed structure of lexical database. Therefore it is difficult to apply or directly utilize it on a generic lexical dataset which originally is not designed for similarity computation.

B. Information-based Measurements

Information-based measurements make use of external corpora which may overcome the unreliability of path distances and taxonomy. In fact, human judgment measurement is mainly based on people's experiences, obtained from different words used in different situations. Therefore, information-based measurements may give accurate results as long as the corpora are correct.

In 1995, Resnik measurement [2] was the first to bring together lexical database and corpora, and calculates similarity by considering the information content (IC) of the LCS of two concepts, denoted by:

$$sim_R(c_1,c_2)=-\log p(lcs(c_1,c_2))$$
$$=IC(lcs(c_1,c_2))$$

If the LCS of two concepts is identical, the similarity is same. That is the deficiency of the measurement.

In 1997, in order to address the deficiency of Resnik's measurement, Jiang & Conrath measurement [4] (shortly JC measurement) calculates similarity by considering not only the IC of the LCS of two concepts, but also every concept's IC, denoted by:

$$dist_{JC}(c_1,c_2)=IC(c_1)+IC(c_2)-2\times IC(lcs(c_1,c_2))$$

Where the semantic similarity is represented by semantic distance, and they have inverse relationships, so the bigger the distance of concepts, the less similar the concepts are.

In 1998, Lin measurement [8] (shortly L measurement) uses the same elements as Jiang & Conrath, but in a different way, denoted by:

$$sim_L(c_1,c_2)=\frac{2\times IC(lcs(c_1,c_2))}{IC(c_1)+IC(C_2)}$$

The similarity value is the ratio of the information shared in common to the total amount of information possessed by two concepts.

In summary, information-based measurements require less structural information of the lexical database, but require large corpora. This limits that it can only be used when the corpus is large and adequate.

C. Hybrid Measurements

T. Hong-Minh and D. Simth [9] proposed a hybrid approach for measuring semantic similarity which is derived from the information-based measurement by adding depth factor and link strength factor. Zhou [10] proposed a hybrid measurement which combines the path length and IC value in 2008. Although the existing hybrid measurements make use of the merits of distance-based measurement and information-based measurement, but their accuracy is not very close to what human perceives. The reason needs further investigation.

III. HowNet Based Computation

The semantic similarity computation mainly depends on some thesaurus taxonomy. For English, WordNet [7] is the most popular and valuable resource. For Chinese, there is a knowledge base called HowNet [8]. Differences and arguable advantage of HowNet over WordNet are described by Dong [12] as follows. (1) WordNet is human oriented; (2) WordNet is word based but HowNet is concept based; (3) The atomic unit of WordNet is synset but that of WordNet is sememe; (4) WordNet is defined with natural language but HowNet is represented by structural markup language. These characteristics make HowNet friendly for computer systems including word similarity measuring.

A. The Overview of HowNet

HowNet is an online bilingual common sense ontology describing semantic relationships between concepts and semantic relationships between the attributes of concepts. It is composed of one concept lexicon file and eleven-concept feature files. HowNet has adopted 1500 sememes to describe all Chinese words, these sememes can be organized into several classes:

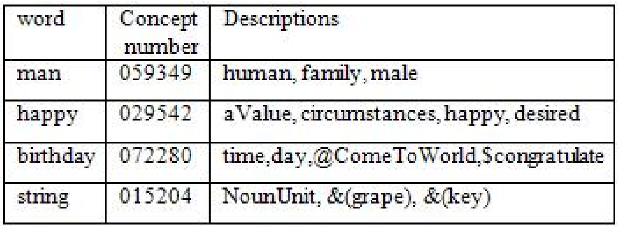
(1) Event
(2) Entity
(3) Attribute
(4) aValue
(5) Quantity
(6) qValue
(7) SecondaryFeature
(8) Syntax
(9) EventRole
(10) EventFeatures

For these sememes, we can classify these into three groups: the first group includes (1) to (7) which called "basic sememe" to describe semantic features of a single concept; the second group only includes (8) so-called "syntax sememe" to describe the syntax features of the Chinese words, mainly is the part-of-speech; the third group includes (9) to (10) so-called "relationship sememe" to describe relationships between concept pairs.

Besides sememe, HowNet uses a set of symbols to describe the semantic meanings of a concept. These symbols are divided into several categories: the first one is the "logic symbols" which is used to represent the logic relationships between semantic descriptions, including the following: , ~ ^; the second one is the "relationship symbols" which is used to represent the relationships between concepts, including the following: # % $ * + & @ ? !; the third category is the "special symbols" which can not come under the logic or relationship symbols, including: {} () [].

TABLE I.

HOWNET KNOWLEDGE DESCRIBING LANGUAGE EXCERPT



word	Concept number	Descriptions
man	059349	human, family, male
happy	029542	aValue, circumstances, happy, desired
birthday	072280	time,day,@ComeToWorld,$congratulate
string	015204	NounUnit, &(grape), &(key)

There are two ways to represent relationships between concepts: using "relationship sememe", or using "relationship symbol" of concept relationships. The former is similar to deep-layered relationships, and most of the latter ones are the inverse of the deep-layered relationships. An excerpt of HowNet description language is shown below in Table Ⅰ.

Sememes are the smallest basic units to describe concept, and there are many complicated relationships among sememes, such the following: hypernym-hyponym, synonymy, antonym, converse, attribute-host, material-product, and event-role. The relationships among sememes compose a complicated network, but not a pure tree-structure. The most important relationship in sememe relationships is hypernym-hyponym. According to the hypernym-hyponym, all the "basic sememes" compose a sememe layered system.

B. Sememe Depth Similarity

Intuitively, semantic similarity between sememes or items should be varies according to depth of sememe node. Suppose there are two sememe nodes which have the same path length. The deeper the position of two sememe nodes in the hierarchy tree is, the more similar two sememe nodes are. For instance, the length of path between "animal" and "plant" is 2; while the length of path between "mammals" and "reptile" is also 2; the former sememe pair is in the topper of hierarchy taxonomy tree. The latter sememe pair is in the deeper of hierarchy taxonomy tree. So the similarity between latter sememe pair is larger than the former sememe pair.

Definition1: Depth of tree root node is 1; others' depth of node is the parent node's depth plus 1.

Definition2: Suppose two sememe node's depth is DA and DB; if two sememe node is the same, the Sim_depth(A,B)=1; if node B is root node, the Sim_depth(A,B)=1/DB; if node A is root node, the Sim_depth(A,B)=1/DA; if node B is the ancestor of node A, the Sim_depth(A,B)=(DB-1)/(DA-1); if node A is the ancestor of node B, the Sim_depth(A,B)=(DA-1)/(DB-1); if two concept nodes have no relationship between grandparent and grandchild, if DA>DB, the Sim_depth(A,B)=(DB-1)/(DA-1), if DB>DA, the Sim_depth(A,B)=(DA-1)/(DB-1).

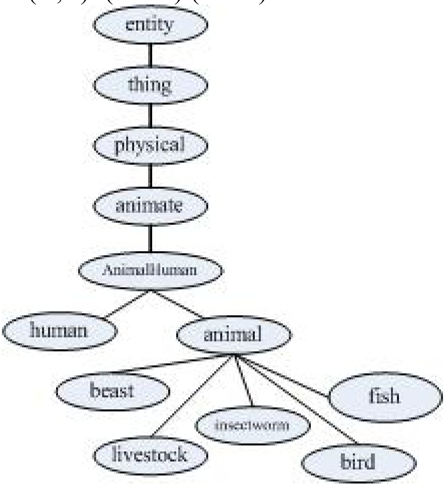


Fig. 1. Diagram 1: the structure of hierarchy sememe tree

Take the above diagram 1 for example, Sim_depth('entity', 'physical')=1/3; Sim_depth('animate', 'animal')=(4-1)/(6-1)=0.6; Sim_depth('beast', 'fish')=1.

C. Sememe Width Similarity

Semantic similarity varies according to the number of children from its parent node in the sememe trees. The higher sememe density of local area indicates parent node has more number of children nodes. Hence, the less information will be inherited from its parent nodes, also demonstrate the more dissimilarity between these two sememe nodes.

Definition 3: The width of leaf node is 1; others' width of node is the number of children from its parent node.

Definition 4: if one node is identical to another node, Sim_width(A,B)=1; if node A is the ancestor of node B or node B is the ancestor of node A, the Sim_width(A,B) is the reciprocal product of the shortest path on all node width; if two node have no relationship about ancestor, the Sim_width(A,B) is the semantic similarity of its parent node divided by the product of its parent node's width.

Diagram 1 used as an example, Sim_width('animate', 'animal') = 1/(1*2*5)=0.1.

D. Sememe Density Similarity

From another point of view, sememe similarity varies according to the density of local area which sememe in, the higher sememe density of local area is, the more detailed classification of the concept nodes are. Therefore, the greater semantic distance between two nodes is, the more similar the two sememe nodes are.

Definition 5: The density of sememe node is defined as the number of nodes in a subtree with the current node as root divided by the number of nodes in a subtree with the current node's parent as root.

For instance, the sememe density similarity between animate and animal is 6/9. The sememe density similarity between animal and beast is 1/6.

E. Sememe Overlap Similarity

From the point of view with sememe overlap similarity, the semantic similarity between two sememe nodes should have something with its information overlap degree in the sememe trees. The more sememe overlap extent is; the bigger sememe similarity is.

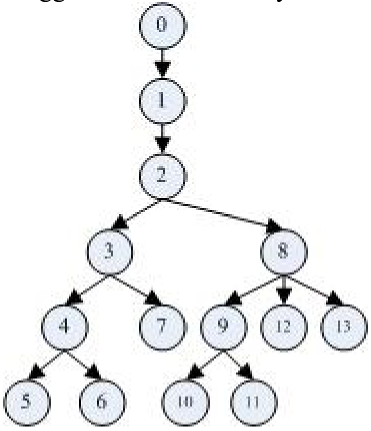


Fig. 2. Diagram 2: the local structure of hierarchy sememe tree

In diagram 2, the shortest path between node 3 and node 8 is the same as the shortest path between node 3 and node 5. The sememe similarity is the same using Liu [15] algorithm; but intuitively, the bigger similarity of the overlap degree of node is, the bigger similarity of sememe node pairs is. The sememe similarity between node 3 and node 5 is bigger than the sememe similarity between node 3 and node 8 since the sememe overlap degree between node 3 and node 5 is bigger than sememe overlap degree between node 3 and node 8.

Definition 6: The shortest path between node A and root node is denoted LA, the shortest path between node B and root node is denoted LB, the shortest path of node A to root node and the shortest path of node B to root node overlap path denoted by C; the sememe overlap similarity is calculated as follows:

$$Sim\_overlap(A,B) = \frac{2*C}{(LA+LB)}$$

Based on this algorithm, the sememe overlap similarity between node 3 and node 5 is: (2*4)/(4+6)=0.8; the sememe similarity between node 3 and node 8 is: (2*3)/(4+4)=0.75.

F. Sememe Similarity Based on Multi-feature Combination

From the aforementioned analysis, only considering a single factor is not comprehensive. Taking all these factors into consideration, we propose a sememe similarity computation method based on multi-features combination, the formula is:

$$Sim(A,B) = \lambda_1 \times Sim\_depth(A,B) + \lambda_2 \times Sim\_width(A,B)$$
$$+ \lambda_3 \times Sim\_density(A,B) + \lambda_4 \times Sim\_overlap(A,B)$$

Where $\lambda_i (1 \le i \le 4)$ is and adjustable parameters and satisfy the equation: $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ and $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.25$.

G. Proposed Semantic Similarity Algorithm

Semantic similarity means the similar degree between two words and can be represented as a real number in domain [0, 1]. Two words with identical DEF (Definition) will get the similarity of 1.0. Liu et al proposed an approach for calculating Chinese words semantic similarities based on HowNet is as follows [11].

For two words W_1 and W_2, if there are n "sense" (concept) $S_{11}, S_{12}, \ldots, S_{1n}$, included in W_1, and there are m "sense" (concept) $S_{21}, S_{22}, \ldots, S_{2m}$ included in W_2. The similarity between W_1 and W_2 is defined using maximum similarity of every "sense" included in W_1 and W_2 respectively, as in Formula (1):

$$Sim(W_1, W_2) = \max_{i=1..n, j=1..m} Sim(S_{1i}, S_{2j}) \qquad (1)$$

"Sense" is represented by "sememe" in HowNet, then, the "sense" similarity can be deduced from the "sememe" similarity. The similarity calculation is obtained with the approach proposed in literature [15], similarity between semantic expressions of two concepts is defined as in Formula (2):

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(S_1, S_2) \qquad (2)$$

Where $\beta_i (1 \le i \le 4)$ is and adjustable parameters and satisfy the equation: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$.

We modify the strategy proposed by Liu [15] which assumes that the similarity between two sememes only depends on the distance between them. The sememe similarity computation is adopted section F method.

IV. Experiments and Results

A. Dataset and Measurement

Rubenstein and Goodenough [12] established synonymy judgments for 65 pairs of nouns. They invited 51 people to assign every pairs a score between 0.0 and 4.0 to indicate semantic similarity. Miller and Charles [13] followed this idea and restricted themselves to 30 pairs of nouns selected from Rubenstein and Goodenough's list, divided equally amongst words with high, intermediate and low similarity.

To evaluate our algorithm on the efficiency of measuring semantic similarity between Chinese words, we translated R.G.-65 dataset into Chinese manually.

B. Experimental Results

We use the human measurements of RG's experiment and MC's experiment as the baseline. We choose some nouns from the MC's list of 30 nouns since some English words cannot be translated into Chinese properly. We compute the correlation coefficient between the human judgments and the measures achieved by our approach. Table Ⅱ shows the semantic similarity based on sememe depth similarity, sememe width similarity, sememe density similarity and sememe overlap similarity, compared with the Miller and Charles' semantic similarity value and give the correlation value.

TABLE II.

COMPARISON OF SEMANTIC SIMILARITY MEASUREMENTS

noun pair	Miller's	depth	width
car-automobile	3.920	0.870	0.870
boy-lad	3.760	1.000	0.750
implement-tool	2.950	0.917	0.413
brother-monk	2.820	0.861	0.722
brother-lad	1.660	0.806	0.500
car-journey	1.160	3.556	0.000
monk-oracle	1.100	0.875	0.500
cemetery-woodland	0.950	0.500	0.0001
coast-shore	0.870	0.167	0.000
forest-graveyard	0.840	0.417	0.000
correlation	1.000	0.696	0.753

noun pair	Miller's	density	overlap
car-automobile	3.920	0.870	0.870
boy-lad	3.760	0.750	0.875
implement-tool	2.950	0.398	0.841
brother-monk	2.820	0.722	0.819
brother-lad	1.660	0.500	0.653
car-journey	1.160	0.0001	0.217
monk-oracle	1.100	0.500	0.688
cemetery-woodland	0.950	0.0001	0.286
coast-shore	0.870	0.000	0.100
forest-graveyard	0.840	0.000	0.231
correlation	1.000	0.857	0.852

We compare the proposed approach with some classic semantic similarity methods, and the result is show in Table Ⅲ:

TABLE III.

COMPARISON OF SEMANTIC SIMILARITY MEASUREMENTS

noun pair	WordNet Edges	Hirst St.Onge	Propsed Method
car-automobile	30.000	16.000	0.847
boy-lad	29.000	5.000	0.837
implement-tool	29.000	4.000	0.620
brother-monk	29.000	4.000	0.794
brother-lad	26.000	3.000	0.606
car-journey	17.000	0.000	0.134
monk-oracle	23.000	0.000	0.636
cemetery-woodland	21.000	0.000	0.176
coast-shore	26.000	2.000	0.068
forest-graveyard	21.000	0.000	0.146
correlation	0.732	0.689	0.852

noun pair	Jiang Conrath	Leacock Chodorow	Propsed Method
car-automobile	1.000	3.466	0.847
boy-lad	0.231	2.773	0.837
implement-tool	0.546	2.773	0.620
brother-monk	0.294	2.773	0.794
brother-lad	0.071	1.856	0.606
car-journey	0.075	0.827	0.134
monk-oracle	0.058	1.386	0.636
cemetery-woodland	0.049	1.163	0.176
coast-shore	0.148	1.856	0.068
forest-graveyard	0.050	1.163	0.146
correlation	0.695	0.821	0.852

C. Discussion

The results of the experiment confirm that our approach based on multi-feature combination provides a significant improvement over the traditional method. The results from the Table 2 show that semantic similarity approach which based on density or overlap outperform the semantic similarity approach which based on depth or width. Although the semantic similarity approach which based on density has a high correlation in these test set, but the overall correlation of semantic similarity approach which based on multi-feature combination is higher.

IV. Conclusion and Future Works

Liu [15] proposed a HowNet based lexical similarity calculation approach, but this algorithm only takes the sememe shortest path into consideration. To address this problem, this paper has presented a semantic similarity computation method based on multi-features combination using HowNet, the measure is obtained from sememe hierarchy structure, quantifying semantic path between sememes, considering the sememe depth similarity, sememe width similarity, sememe density similarity and sememe overlap similarity, then combining these factors into a single measure on the notion of sememe similarity degree. Experiment results show that our algorithm is comparative with other classic similarity algorithms, with the results closet to human similarity judgments.

In the future, some other factors which affect accuracy of semantic similarity will be explored. We will conduct more experiments to find out the factors which influence the similarity value. We will concentrate on a hybrid semantic similarity computation method between word pairs which uses some thesaurus taxonomy to generate satisfactory results which are more close to what human perceives than we can achieve in this paper.

Acknowledgment

This work is supported by "the Fundamental Research Funds for the Central Universities" of China University of Petroleum (East China) (Grant No.13CX02031A). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] R. Rada, H. Mili, E. Bichnell, M. Blettner. Development and application of a metric on semantic nets. IEEE Trans. Systems, Man and Cybernetics, vol9, No. 1, Jan. 1989,pp.17-30.

[2] P. Resnik. 1995. Using information content to evaluate semantic synonymy. Proc 14th International Joint Conference on Artificial Intelligence, 448-453, Montreal.

[3] G. Hirst and D. St-onge. Lexical chains as representation of context for the detection and correction of malapropisms. In Fellbaum (1998), p.305-332

[4] Jay J Jiang, David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. Int. Conf. Research on Computational Linguistics (ROCLING X), 1997.

[5] C. Leacock, M.Chodorow. Combing local context and wordnet similarity for word sense identification. Proceeding of Fellbaum, 1998, pp.265-283.

[6] D.L. Medin, R.L. Goldstone, and D. Gentner, "Respects for Similarity", Psychological Rev., vol.100, no.2, pp.254-278, 1993.

[7] Wu Zhibiao and Martha Palmer. Verb semantics and lexical selection. In proceedings of the 32th annual meeting of the association for computational linguistics, Las Cruces, NM, June, 1994, pp.133-138.

[8] D. Lin. An information theoretic definition of similarity. In Proceedings of 15th international conference on machine learning. San francisco, Morgan Kaufmann Publishers Inc, 1998, pp.296-304.

[9] Tran Hong-Minh and Dan Simth. Word similarity in wordnet. Proceedings of the third international conference on high performance scientific computing, Hannoi, Vietnam, March, 2006, pp.1-10.

[10] Zili Zhou, Yanna Wang and Junzhong Gu. New model of semantic similarity in the taxonomy of wordnet. Proceedings of the 28th Australasian computer science conference, Australia, February, 2005, pp.315-322.

[11] WordNet: http://wordnet.princeton.edu

[12] INFORMATION ON HTTP://WWW.KEENAGE.COM

[13] Tongyici Cilin (Extended) [online], available: http://ir.hit.edu.cn/demo/ltp/Sharing Plan.htm, October 26, 2009.

[14] PENG Jing, YANG DongQing, TANG ShiWei and et al. A new similarity computing method based on concept similarity in Chinese text processing, *Science in China Series F: Information Sciences*, vol.51, no.9, pp.1215-1230, 2008.

[15] Liu Qun, Su Jian. "HowNet" lexical similarity calculation. *Proc of Chinese lexical semantics workshop third TECHNOLOGY*, Taipei, pp.59-76, 2002.

[16] H. Rubenstein, J.B. Goodenough. Contextual correlates of synonymy. Communications of the ACM, 1965,8(10),pp.627-633.

[17] G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*,6(1):1-28

[18] G. Varelas, E. Voutsakis, and P. Raftopoulou. Semantic similarity methods in WordNet and their application to information retrieval on the web. Proceedings of the 7th annual ACM international workshop on web information and data management Bremen, Germany, 2005, pp.10-16.

[19] TIAN Jiu-le, ZHAO Wei. Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system. Journal of Jilin University(Information Science Edition) , vol 28, No 6, 2010, pp.602-608.

[20] Yuhua Li, Zuhair A.Bandar, and David MclLean. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering. Vol 15, No 4, 2003, pp.871-882

[21] Songmei Cai, Zhao Lu. An Improved Semantic Similarity Measure for Word Pairs. 2010 International Conference on e-Education, e-Business, e-Management and e-Learning. pp. 212-216.

[22] Liuling DAI, Bin LIU et al. Measuring Semantic Similarity between Words Using HowNet. International Conference on Computer Science and Information Technology 2008. pp. 601-605.

Pei-ying Zhang received the master's degree in computer science and technology from china university of petroleum (east china), in 2006. He is a research teacher of china university of petroleum (east china). Currently, he is a lecturer at school of computer and communication engineering. His interests are in natural language processing and oil-field information development management.