A New Method for Calculating Word Sense Similarity in WordNet[1]

Lingling Meng[1] and Junzhong Gu[2]

[1]Computer Science and Technology Department, Department of Educational Information Technology
[2]Computer Science and Technology Department

East China Normal University, Shanghai, 200062, China
llmeng@deit.ecnu.edu.cn, jzgu@ica.stc.sh.cn

Abstract

Semantic similarity between word senses is hot topic in many applications of computational linguistics and artificial intelligence, such as word sense disambiguation, information extraction, semantic annotation and ontology learning. Many methods for calculating word sense similarity have been proposed. In recent years the methods based on WordNet have shown its talents and attracted great concern. In the paper, we present a new method in WordNet for calculating word sense similarity, which is noun and is-a relation based. We evaluate our method on the data set of Rubenstein and Goodenough, which is traditional and widely used. The correlation with human judgment is o.8804 in proposed measure, which is more close to human judgments than related works. Experiments show that our new measure significantly outperformed than other existing computational methods.

Keywords: *semantic similarity, information content based, WordNet*

1. Introduction

The study of semantic similarity between words has been a part of computational linguistics and artificial intelligence for many years. It is a central issue for many applications, such as text segmentation [1], word sense disambiguation [2], information extraction [3, 4], semantic annotation and summarization [5, 8], question answering [6], recommender system [7], document clustering [9, 10, 11], information retrieval [12] and so on. Many measures have been developed in the past years. Generally all the measures can be grouped into two categories. One is making use of a large corpus to estimate semantic similarity. The other is aiming to use the relations and the hierarchy of a thesaurus, such as WordNet [13]. Recently the latter have shown its talents and attracted great concern. This paper presents a new method for calculating word sense similarity, which is WordNet based. Experiments demonstrated that our new method significantly outperformed related works.

[1] The work in the paper was supported by Shanghai Scientific Development Foundation (Grant No.11530700300)

The remainder of the paper is organized as follows: Firstly we provide the background information regarding WordNet, and then we discuss popular related works of word sense similarity measures in Section 2. In Section 3 a novel semantic similarity measure based on WordNet is proposed. How to use the dataset to analyze the new measure and compare the performance with other measures are discussed in Section 4. Conclusion and future work are probed in Section 5.

2. Semantic Similarity Measures

2.1. WordNet

The measures discussed in the paper are all based on WordNet. WordNet is the product of a research project in Princeton University which has attempted to model the lexical knowledge of a native speaker of English [13]. In WordNet Nouns, verbs, adjectives, and adverbs are connected to each other into taxonomic hierarchies by well-defined types of semantic relations. These semantic relations for nouns include Hyponym/Hypernym (is-a), Part Meronym/Part Holonym (part-of), Member Meronym/Member Holonym (member-of), Substance Meronym/Substance Holonym (substance-of) and so on. For example, a car is a wheeled vehicle (is-a), and a cell is part of organism (part-of). Hyponym/hypernym (is-a) is the most common relations. A fragment of is-a relation in WordNet is shown as Figure 1.

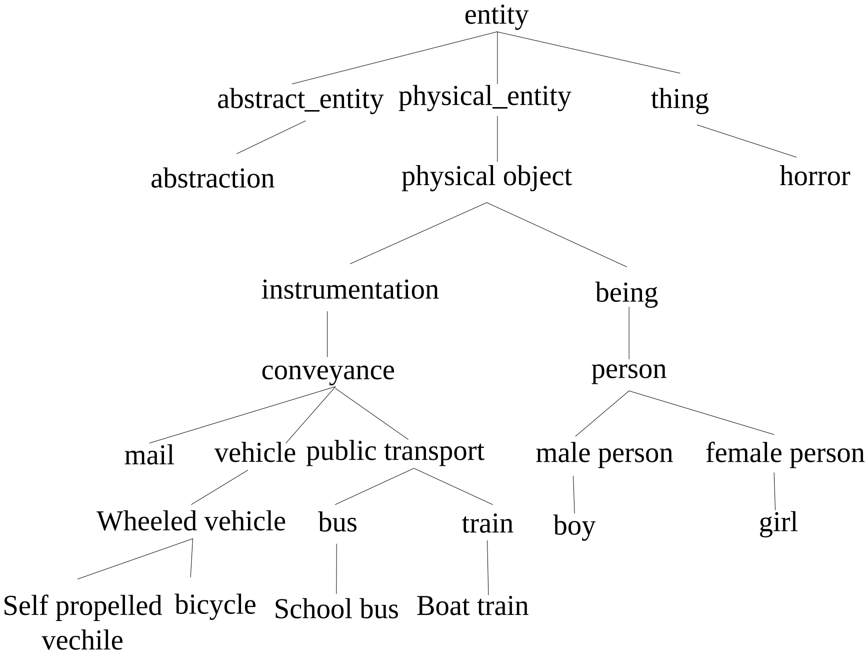

Figure 1. A Fragment of is-a Relation in WordNet

In the taxonomy the deeper concept is more specific and the upper concept is more abstract. Because language semantics are mostly captured by nouns and noun phrases, in the paper we only discuss the similarity measures based on nouns and is-a relations in WordNet. Some methods for obtaining similarity between words in WordNet have been proposed, which can be classified into two groups: edge-based measures and information-based measures. Next, we will introduce these measures briefly. Before discussion, it is necessary to define the related concepts mentioned in the following measures.

Definition 1. len(c_i,c_j): the length of the shortest path from concept c_i to concept c_j in WordNet, eg. len(boy,girl)=4.

Definition 2. lso(c_i,c_j): the most specific common subsumer of c_i and $c_{j,}$ eg. lso(boy,girl)=person.

Definition 3. depth(c_i): the length of the path to concept c_i from the global root entity. depth(root)=1, eg.depth(boy)=7. Besides this, it should be noted that:

depth(c_i) - detph(lso(c_i,c_j)) + depth(c_j) - detph(lso(c_i,c_j))
= depth(c_i) + depth(c_j) - 2*detph(lso(c_i,c_j))
= len(c_i,c_j)

Definition 4. deep_max: the max depth(c_i) of the taxonomy. In Figure 1 deep_max is equal to 8.

Definition 5. hypo(c): the number of hyponyms for a given concept c, eg. hypo(person)=4.

Definition 6. node_max: the maximum number of concepts that exist in the taxonomy of WordNet. In Figure 1 node_max is 25.

Definition 7. sim (c_i,c_j): semantic similarity between concept c_i and concept c_j.

For two compared concepts c_i and c_j in taxonomy as in Figure 1, the length of the shortest path from concept c_i to concept c_j can be determined from one of three cases:

Case1: c_i and c_j are the same concept, thus c_i, c_j and lso(c_i,c_j) are the same node. We assign the semantic length between c_i and c_j to 0, ie.len(c_i,c_j)=0.

Case2: c_i and c_j are not the same node, but c_i is the parent of c_j. thus lso(c_i,c_j) is c_i. We assign the semantic length between c_i and c_j to 1, ie.len(c_i,c_j)=1..

Case3: Neither c_i and c_j are the same concept nor c_i is the parent of c_j, we count the actual path length between c_i and c_j, therefore 1<len(c_i,c_j)<= 2*deep_max.

Based on the above definitions and cases, we discussed the following measures.

2.2. Path-based Measures

Path based measures take the path length linking the concepts and the position of the concepts into considerate. One of path-based measures is Wu and Palmer's. In a paper on translating English verbs into Mandarin Chinese, Wu and Palmer introduced a scaled measure between a pair of concepts c_i and c_j in a hierarchy as [14]:

$$sim_{WP}(c_i,c_j)=\frac{2*depth(lso(c_i,c_j))}{len(c_i,c_j)+2*depth(lso((c_i,c_j))} \tag{1}$$

The similarity between two concepts (c_i, c_j) is the function of their distance and the specific common subsumer(lso(c_i,c_j)). If c_i and c_j are the same concept, len(c_i,c_j)=0 and sim_{WP} (c_i,c_j) = 1; if c_i and c_j are the different concept, 0<sim_{WP} (c_i,c_j) < 1. Thus, the values of sim_{WP} (c_i,c_j) are in (0, 1].

Another classical path-based measure is Leakcock and Chodorow's [15]. The maximum depth of taxonomy had been taken into account in their method.

$$sim_{LC}(c_i, cj) = -\log \frac{len(c_i, c_j)}{2 * deep\_\max} \tag{2}$$

For a specific version of WordNet, deep_max is a fixed value, therefore the similarity between two concepts (c_i, c_j) is the function of the shortest path len(c_i,c_j) from c_i to c_j. If c_i and c_j are the same concept, len(c_i,c_j) is 0. In practice, we may add 1 to both len(c_i,c_j) and 2*deep_max to avoid log (0). Thus the values of sim_{LC}(c_i,c_j) are in (0, log(2*deep_max+1)].

Li et. al., [16] combines the shortest path and the depth of concepts in a non-linear function, expressed by:

$$sim_{Li}(c_i, c_j) = e^{-\alpha * len(c1,c2)} \frac{e^{\beta * depth(lso(c_i,c_j))} - e^{-\beta * depth(lso(c_i,c_j))}}{e^{\beta * depth(lso(c_i,c_j))} + e^{-\beta * depth(lso(c_i,c_j))}} \tag{3}$$

Where α (α>0) and β (β>0) are parameters scaling the contribution of shortest path length and depth respectively, which need to be adapted manually for good performance. In our experiment the same as in literature [16]'s, we set α = 0.2 and β=0.6. It is noted that sim_{Li}(c_i,c_j) will monotonically increasing with respect to depth(lso(c_i,c_j)) and decreasing with len(c_i,c_j). The values of sim_{Li}(c_i,c_j) are between 0 and 1.

2.3. Information Content based Measures

Information content based similarity measure usually employ the notion of information content, which can be considered as a measure quantifying the amount of information a concept expresses. It was first proposed by Resnik [17] in 1995 following information theoretic approach, after which Jiang [18], Lin [19], also proposed two other measures respectively. All of these measures rely on information content(IC) values assigned to the concepts in the taxonomy, but their usage of IC are different.

Resnik assumed that for a concept c, let *p(c)* be the probability of encountering and instance of concept c. The IC value is obtained by considering the negative log likelihood.

$$IC = -\log p(c) \tag{4}$$

Probability of a concept was estimated as follows:

$$p(c) = \frac{freq(c)}{N} \tag{5}$$

Where N is the total number of nouns, and freq(c) is the frequency of instance of concept c occurring in the taxonomy. When computing freq(c), each noun or any of its taxonomical hyponyms that occurred in the given corpora was included.

$$Freq(c) = \sum_{w \in W(c)} count(w) \tag{6}$$

Where *W(c)* is the set of words subsumed by concept c.

For two given concepts c_i, c_j, semantic similarity depended on the amount of information two concepts c_i and c_j shared in common, the more information two concepts share in common, the more similar they are. The shared information was indicated by the information of the specific common subsumers, ie. lso(c_i,c_j).

$$sim_{\mathrm{Res}nik}(c_i, c_j) = -\log p(lso(c_i, c_j)) = IC(lso(c_i, c_j)) \quad (7)$$

In Lin's measure the similarity between concept c_i and c_j depended on not only their shared information content, but also their self information content respectively. It assumed that the similarity between c_i and c_j was measured by the ratio between the amount of information needed to state the commonality of c_i and c_j and the information needed to fully describe what c_i and c_j are.

$$sim_{Lin}(c_i, c_j) = \frac{2 * IC(lso(c_i, c_j)))}{IC\ (c_i) + IC\ (c_j)} \quad (8)$$

Jiang proposed a measure from different view by calculating semantic distance to obtain semantic similarity in 1997. Semantic similarity is the opposite of the semantic distance.

$$dis_{Jiang}(c_i, c_j) = (IC(c_i) + IC(c_j)) - 2IC(lso(c_i, c_j)) \quad (9)$$

So both Jiang's measure and Lin's measure have taken the IC of compared concepts into account respectively.

It should be noted that, the IC value of each concept is crucial in Information content-based similarity measure. Each of the measures in Section 2.3 attempts to exploit the information contained at best to evaluate the similarity between the pairs of concepts. There are two methods to obtain IC. One is estimated from a corpus, and the other is using WordNet as a statistical resource for computing the probability of occurrence of concepts. In this paper, we adopt the latter one. One commonly used measure to obtain the information content(IC) of a given concept was proposed by Nuno. It is based on the assumption that in WordNet IC value of a concept is regarded as the function of the hyponyms it has. Concepts with more hyponyms expressed less information than the concepts with less ones. It is defined as [20]:

$$IC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(node\_\max)} \quad (10)$$

3. A New Method for Calculating Word Sense Similarity

The measures of above-mentioned are all simple, but the results are not very close to human's judgment. There is still room for improvement. We have developed a new method which shares some properties of Lin's method. Different from Lin's, exponential function is taken to smooth the resulting values. The new method is defined as:

$$sim_{new}(c_i, c_j) = e^{sim_{lin}} - 1 = e^{\left(\frac{2*IC(lso(c_i,c_j))}{IC\ (c_i)+IC\ (c_j)}\right)} - 1 \quad (11)$$

The new measure is based on Lin's method. Therefore, it is an information content based semantic similarity measure. From formula (11) we can see that it will monotonically increase with sim_{Lin}. Its curve graph is shown in Figure 2. In Lin's method, the values are in [0,1]. If sim_{Lin} $(c_i,c_j)=0$, $sim_{new}(c_i,c_j)=e^0-1=1-1=0$; if sim_{Lin} $(c_i,c_j)=1$, $sim_{new}(c_i,c_j)=e^1-1=e-1$. Thus, the values of sim_{new} are in [0, e-1].

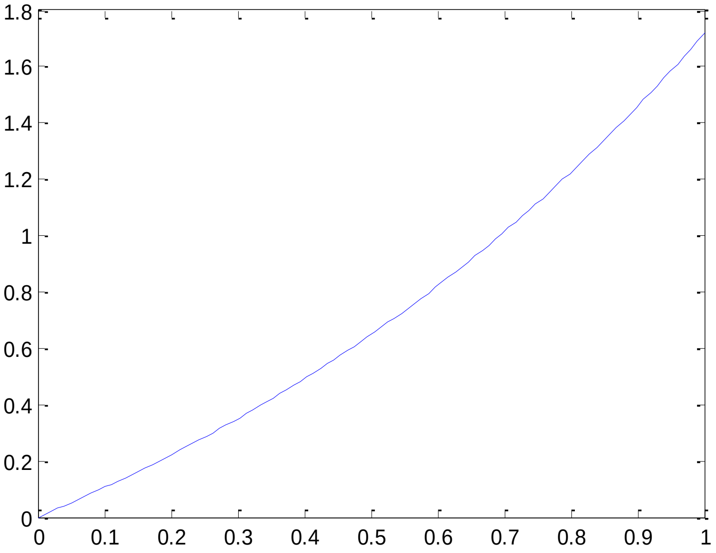

Figure 2. The Curve Graph of Formula (11)

Notes: In Figure 2, X-axis is the semantic similarity value computing with Lin's method and Y-axis is the semantic similarity value computing with formula (11).

Table 1. Comparison of Different Semantic Similarity Measures

Category	Principle	Measure	Features
Path based measure	Function of path length linking the concepts and the position of the concepts in the taxonomy	W&P	Function of their distance and the specific common subsumer(lso(c_i,c_j)).
		L&C	Function of the shortest path len(c_i,c_j) from c_i to c_j.
		Li	Non-linear function of the shortest path and depth of lso(c_i,c_j).
IC based measure	The more common information two concepts share, the more similar the concepts are.	Resnik	Function of IC value of lso(c_i,c_j).
		Lin	Function of IC value of lso(c_i,c_j) and the compared concepts respectively.
		Jiang	Function of IC value of lso(c_i,c_j) and the compared concepts respectively.
		new measure	Function of IC value of lso(c_i,c_j) and the compared concepts respectively.

Next, let's compare the features of our new measure with measures mentioned in Section 2. Table 1 presents the results.

In next section, we will analyze and evaluate our measure from many perspectives.

4. Evaluation

There is not a standard to evaluate computational measures of semantic similarity. Generally there are three kinds of methods. The first one is a theoretical examination of a computational measure for those mathematical properties thought desirable. The second one is to compare the measures by calculating the coefficients of correlation with human judgments. The third one is application-oriented, which is to compare the performance of different measures in a particular application. In this paper, we select the second evaluation measure.

4.1. Data set

To evaluate the performance of our new method, a dataset is necessary. One commonly used dataset is provided by Rubenstein and Goodenough (1965) [21]. Rubenstein and Goodenough obtained “synonymy judgment” from 51 human subjects on 65 pairs of words ranged from “highly synonymous” to “semantically unrelated”, and the subjects were asked to rate them, on the scale of 0.0 to 4.0.

4.2. Words Similarity Calculating Method

Because either or both of the words have more than one sense in WordNet, we took the most similarity pair of sense:

$$sim(w_1, w_2) = \max_{(i,j)}[sim(c_{1i}, c_{2j})] \tag{12}$$

Where c_{1i} is the sense of word1, and c_{2j} is the sense of word2. For each of seven implemented measures, we compute similarity scores for the human-rated pairs.

4.3. Results Analysis

Before our analysis, we first compute semantic similarity between pairs of words with formula (1) ~ (3), (7) ~ (9) and our new method and draw the obtained similarity values in diagrams. For the convenience of expression and comparison, we normalized the values in [0,1]. The IC value is obtained according formula (10).Theses results are shown in Figure 3.

As shown in Figure 3, in most pairs of words our method is more accurate than other six measures. In accordance with previous research we compare the six chosen measures listed in Section 2 with our new method by calculating the coefficients of correlation with human judgments of semantic similarity. The chosen algorithms and their correlation coefficient are illustrated in Table 2.

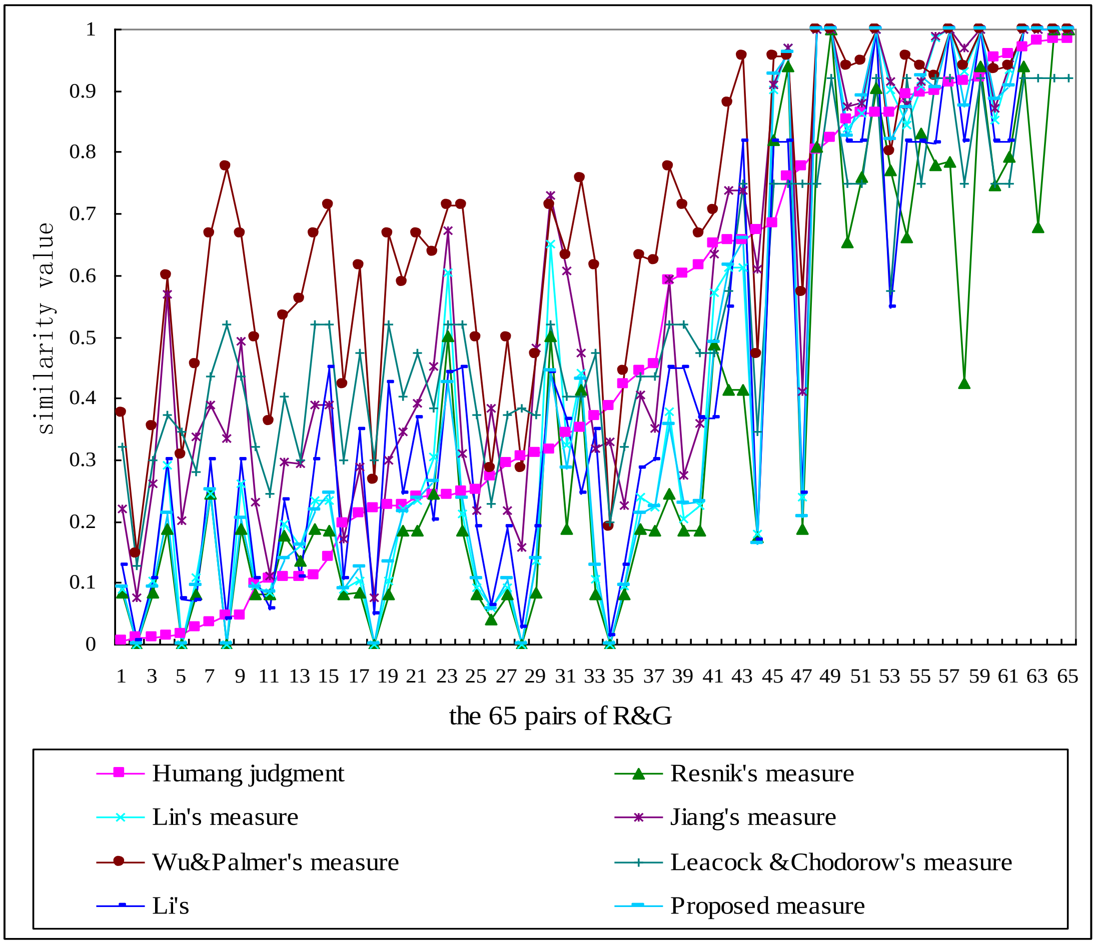

Figure 3. Compare the Similarity Obtained from the Six Measures

Table 2. Coefficients of Correlation between Human Ratings of Similarity

Similarity algorithm	Coefficients of correlation (R&G)
Wu & palmer	0.7767
Leacock & Chodorow	0.8535
Li	0.8559
Resnik	0.8400
Lin	0.8643
Jiang	-0.8569
New measure	0.8804

The compared results of our proposed measure with other six measures are presented in Figure 4.

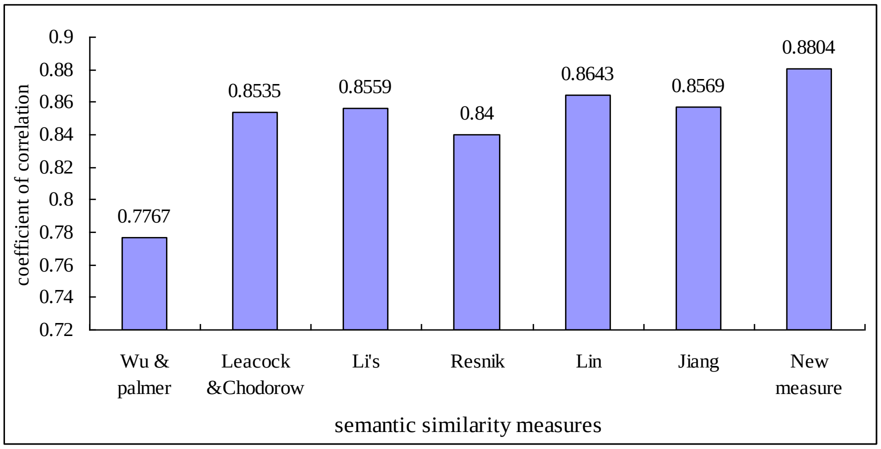

Figure 4. The Compared Results of our Proposed Measure with other Six Measures

From Table 2 and Figure 4, we can see that the coefficient of correlation between human ratings of similarity computed with our proposed method is the highest among the seven measures, which indicates the good performance of our measure.

5. Conclusion and Future Work

This paper presents a new method of word sense similarity based on WordNet. In our measure, noun and is-a relation have been concerned about. We evaluate our measure on the data set of Rubenstein and Goodenough (1965), and compare the results of our proposed measure with Wu&palmer's method, Leacock &Chodorow' method, Li's method, Resnik's method, Lin's method, Jiang's method. The distributed graphs of 65 word pair's similarity value with different methods are illustrated. Experiments show that the correlation with human judgment is 0.8804 in proposed method, which is better than other sixes. In future work, we will put the method into practice. We intend to make some attempt in ontology construction and big data analysis.

References

[1] H. Kozima, "S: Computing Lexical Cohesion as a Tool for Text Analysis", doctoral thesis. Computer Science and Information Math, Graduate School of Electro-Comm. Univ. of Electro-Comm., **(1994)**.

[2] S. Patwardhan, S. Banerjee and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation", Proceedings of 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, **(2003)** February 16-22; Mexico City, Mexico.

[3] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst., vol. 22, no. 7, **(2009)**.

[4] M. Stevenson, M. A. Greenwood, "A semantic approach to IE pattern induction", Proceedings of 43rd Annual Meeting on Association for Computational Linguistics, **(2005)** June 25-30; Ann Arbor, Michigan, USA.

[5] D. Sánchez, D. Isern and M. Millán, "Content annotation for the Semantic Web: an automatic web-based approach", Knowl. Inf. Syst., vol. 27, no. 3, **(2011)**.

[6] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce", Knowl.-Based Syst., vol. 21, no. 8, **(2008)**.

[7] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López- Nores, J. García-Duque, A. Fernández-Vilas, R. P. Díaz-Redondo and Jesús Bermejo-Muñoz, "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems", Knowl.-Based Syst., vol. 21, no. 4, **(2008)**.

[8] C. Y. Lin and E. Hovy, “Automatic evaluation of summaries using ngram co-occurrence statistics”, Proceedings of Human Language Technology Conference, **(2003)** May 27-June 1; Canada, Edmonton.

[9] A. Hotho, S. Staab and G. Stumme, “Wordnet improves text document clustering”, Proceedings of the Semantic Web Workshop at 26th Annual International ACM SIGIR Conference, **(2003)** July 28 - August 1; Canada,Toronto.

[10] J. Jing, L. Zhou, M. K. Ng and Z. Huang, “Ontology-based distance measure for text clustering”, Proceedings of SIAM SDM workshop on text mining, **(2006)** April 20-22; USA, Maryland, Bethesda.

[11] I. Yoo, X. Hu and I. -Y. Song, “Integration of Semantic-based Bipartite Graph Representation and Mutual Refinement Strategy for Biomedical Literature Clustering”, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2006), **(2006)** August 20-23; Philadelphia, USA.

[12] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis and E. E. Milios, “Semantic similarity methods in wordNet and their application to information retrieval on the web”, WIDM '05. ACM Press, New York, NY, **(2005)**, pp. 10-16.

[13] C. Fellbaum (editor), “WordNet: An electronic lexical database”, MIT Press, **(1998)**.

[14] Z. Wu and M. Palmer, “Verb semantics and lexical selection”, Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, **(1994)** June 27-30; Las Cruces, New Mexico.

[15] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification”, WordNet: An Electronic Lexical Database, MIT Press, **(1998)**, pp. 265-283.

[16] Y. Li, A. B. Zuhair and D. McLean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, **(2003)**.

[17] P. Resnik, “Using information content to evaluate semantic similarity”, Proceedings of the 14th International Joint Conference on Artificial Intelligence, **(1995)** August 20-25; Montréal Québec, Canada.

[18] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy”, Proceedings of International Conference on Research in Computational Linguistics, **(1997)** August 22-24; Taipei, Taiwan.

[19] D. Lin, “An information-theoretic definition of similarity”, Proceedings of the 15th International Conference on Machine Learning, **(1998)** July 24-27; Madison, Wisconsin, USA.

[20] N. Seco, T. Veale and J. Hayes, “An intrinsic information content metric for semantic similarity in WordNet”, Proceedings of the16th European Conference on Artificial Intelligence, **(2004)** August 22-27; Valencia, Spain.

[21] H. Rubenstein and J. B. Goodenough, “Contextual correlates of synonymy”, Communications of the ACM, vol. 8, no. 10, **(1965)**.

Authors

Lingling Meng

Lingling Meng is a PhD Candidate of Computer Science and Technology Department and a teacher of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Junzhong Gu

Prof. Junzhong Gu is Supervisor of PhD Candidates, full professor of East China Normal University, head of Institute of Computer Applications and director of Lab, Director of Multimedia Information Technology (MMIT). His research interests include information retrieval, knowledge engineering, context aware computing, and data mining.