International Journal of Future Generation Communication and Networking
Vol.7, No.3 (2014), pp.183-194
http://dx.doi.org/10.14257/ijfgcn.2014.7.3.17

Measuring Semantic Similarity of Word Pairs Using Path and Information Content [1]

Lingling Meng[1], Runqing Huang[2] and Junzhong Gu[3]

[1]*Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China*
[2]*Shanghai Municipal People's Government, Shanghai, 200003, China*
[3]*Computer Science and Technology Department, East China Normal University, Shanghai, 200062, China*
[1]*llmeng@deit.ecnu.edu.cn,* [2]*runqinghuang@gmail.com,* [3]*jzgu@ica.stc.sh.cn*

Abstract

Measuring semantic similarity of word pairs is a popular topic for many years. It is crucial in many applications, such as information extraction, semantic annotation, question answering system and so on. It is mandatory to design accurate metric for improving the performance of the bulk of applications relying on it. The paper presents a new metric for measuring word sense similarity using path and information content. Different from previous works, the new metric not only reflects the semantic density information, but also reflects the path information. It is evaluated on the dataset provided by Rubenstein and Goodenough. Experiments demonstrate that the coefficient based on our proposed metric with human judgment is 0.8817, which is significantly outperformed than other existing methods.

Keywords: *semantic similarity, word pairs, path, information content, WordNet*

1. Introduction

Measuring semantic similarity of word pairs is a general issue in linguistics, cognitive science, and artificial intelligence. It has been successfully applied in word sense disambiguation [1], information extraction [2], semantic annotation and summarization [3-4], recommender system [5], question answering [6], and so on. Besides this, it also shows its talents in software domain [7] and bio-informatics domain [8]. Therefore a proper metric is curial for improving the performance of the bulk of applications relying on it. Many metrics have been proposed from different of view. Some metrics take the structure information into considered, and others assume that the semantic density information should be taken into account. Therefore, all the metrics can be divided into two classes: path based metrics and information content based metrics. Path based metrics assess semantic similarity by counting the number of edges separating two concepts. Information Content(IC) based metrics exploit the notion of Information Content (IC) of concepts. All the metrics are intuitive, simple and effective. However, they can't distinguish different concepts pairs. The paper proposes a new metric for measuring word sense similarity combining path and information content.

[1] The work in the paper was supported by Shanghai Industry-University Cooperation Foundation (Grant No. Shanghai CXY-2013-84) and Shanghai Scientific Development Foundation (Grant No.11530700300).

ISSN: 2233-7857 IJFGCN

We evaluate the new metric on the data set of Rubenstein and Goodenough. Experiments show that the coefficient of our proposed metric with human judgment is 0.8817, which demonstrates that the new metric significantly outperformed than related work.

The rest of the paper is as follows: we introduce WordNet, definitions of related concepts and related works in section 2. In Section 3 a new semantic similarity metric of word pair using information content and path based on WordNet is proposed. In the new metric not only the semantic density information, but also the path information has been reflected. In section 4 we compare the six chosen metrics with our new algorithm by calculating the coefficients of correlation with human judgments of semantic similarity on the dataset provided by Rubenstein and Goodenough to evaluate the performance. Section 5 illustrates conclusion and future work.

2. Related Work

2.1. WordNet

The metrics in the paper are all based on WordNet. It is necessary to give a brief introduction.

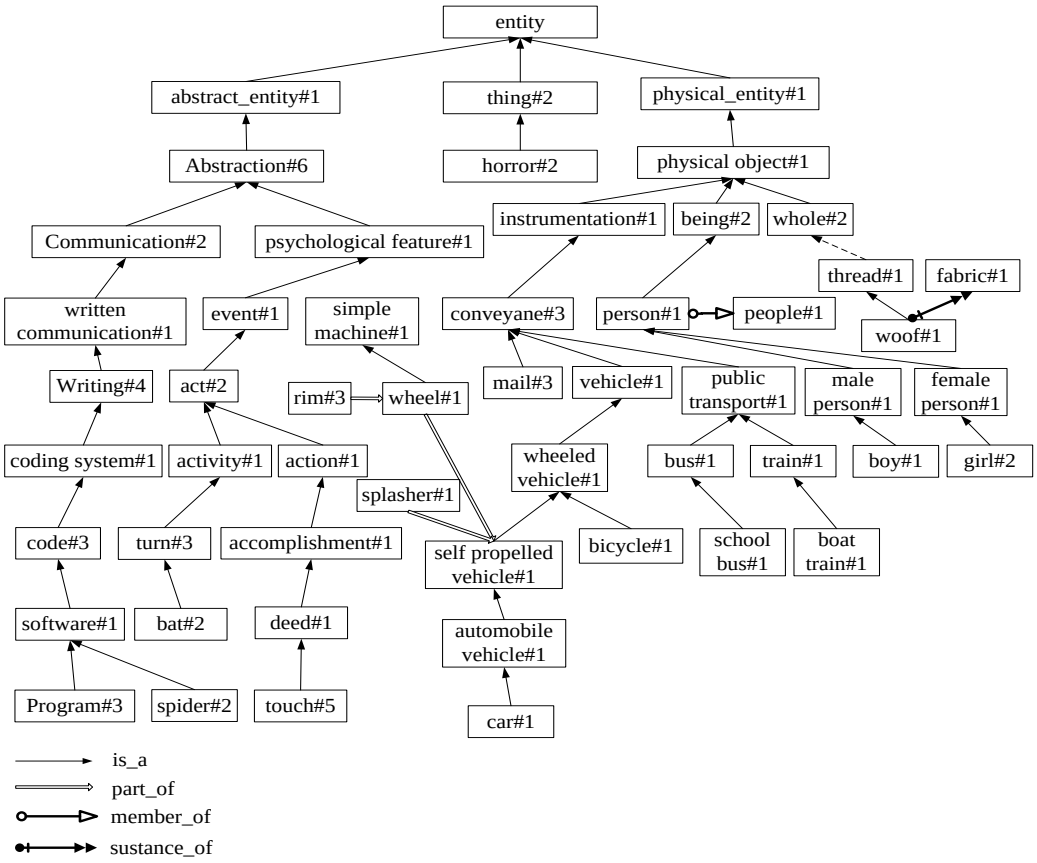

Figure 1. A Fragment is-a Hierarchy Taxonomy in WordNet

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English [9]. Now it has become a valuable resource and plays an important role in human language technology. WordNet focuses on the word meanings instead of word forms. In WordNet nouns, verbs, adjectives, and adverbs are represented by a synset, which denotes a concept or a sense of a group of terms. These synsets are organized into taxonomic hierarchies via a variety of semantic relations, which makes it a useful tool for computational linguistics and natural language processing. It is commonly argued that language semantics are mostly captured by nouns or noun phrases so that our study only focus on noun in semantic similarity calculating. In WordNet 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. Figure 1 illustrates a fragment in WordNet. In Figure 1 we can see that car is an automobile vehicle, and rim is part of wheel, and person is member of people.

Hyponym/hypernym (is-a) is the most common relations, which connects all the concepts into a hierarchy taxonomy. In the taxonomy the deeper concept is more specific and the upper concept is more abstract. For example, in Figure 1 the most abstract concept is entity. Car is more specific than automobile vehicle and automobile vehicle is more specific than wheeled vehicle.

In this paper, we are only concerned about the similarity metrics based on is-a relations of WordNet. Some metrics have been proposed in past years. Generally the typical metrics based on WordNet can be grouped into two categories: path-based metrics and information-based metrics. Next, we will introduce these metrics briefly.

2.2. Definitions

Definition of related concept in the following metrics as follows:

(1) len(c_i,c_j): the length of the shortest path from synset c_i to synset c_j in WordNet. eg. len(bus,train) is 2.

(2) lso(c_i,c_j): the most specific common subsumer of c_i and c_j. eg. lso(bus,train) is public transport.

(3) depth(c_i): the length of the path to synset c_i from the global root entity. Here depth(root) is set to 1. eg. depth (bus) is 7.

(4) deep_max: the max depth(c_i) of the taxonomy. In Figure1 deep_max is 10.

(5) hypo(c): the number of hyponyms for a given concept c. eg. hypo (sofware) is 2.

(6) node_max: the maximum number of concepts that exist in the taxonomy.

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

2.3. Semantic Similarity Metrics

2.3.1. Path Based Metrics

Path based metrics proceed from the position of each concept in the taxonomy to obtain semantic similarity and assess semantic similarity by computing geometric distance separating two concepts, such as the number of edges. It is based on the assumption that the similarity of two concepts is related with the path length between two concepts and depth of each concept in the taxonomy respectively.

In a paper on "translating English verbs into Mandarin Chinese", Wu and Palmer (W&P) presented a scaled metric for measuring the similarity between a pair of concepts c_1 and c_2. It is defined by how closely they are related in the hierarchy taxonomy. Formally, the metric is as follows [10]:

$$sim_{W\&P}(c_1, c_2) = \frac{2 * depth(lso(c_1, c_2))}{len(c_1, c_2) + 2 * depth(lso((c_1, c_2))} \quad (1)$$

It is noticed that the similarity between two concepts (c_1, c_2) is inversely proportional to length (c_1, c_2) and proportional to depth (lso(c_1, c_2)). The values are range from 0 to 1.

Leakcock and Chodorow (L&C) took the maximum depth of taxonomy into account and proposed the following metric [11]:

$$sim_{L\&C}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 * deep\_\max} \quad (2)$$

It assumes that the similarity between two concepts was the function of the shortest path from c_1 to c_2 and depth of the taxonomy. For a specific version of WordNet, deep_max is a fixed value. Therefore in the taxonomy, the smaller shortest path that two concepts have, the more similar they are. In practice, if two words have the same sense, c_1 and c_2 are the same node in the taxonomy. Then len(c_1,c_2) is 0, so we may add 1 to both len(c_1,c_2) and 2*deep_max to avoid log(0). The values of $sim_{L\&C}$ (c_1, c_2) are range from 0 to log (2*deep_max+1).

Li *et al.*, [12] uses multiple information sources to calculate the semantic similarity of concepts and proposes a metric based on the assumption that information sources are infinite to some extent while humans compare word similarity with a finite interval between completely similar and nothing similar. Intuitively the transformation between an infinite interval to a finite one is non-linear [13], which is expressed by:

$$sim_{Li}(c_1, c_2) = e^{-\alpha * len(c_1, c_2)} \frac{e^{\beta * depth(lso(c_1, c_2))} - e^{-\beta * depth(lso(c_1, c_2))}}{e^{\beta * depth(lso(c_1, c_2))} + e^{-\beta * depth(lso(c_1, c_2))}} \quad (3)$$

It is noticed that Li's metric combines the shortest path and the depth of concepts in a non-linear function. Where α (α>0) and β (β>0) are parameters and used to adjust the contribution of shortest path length (ie. length (c_1, c_2)) and depth (ie. depth (lso (c_1, c_2)) respectively, which need to be adapted manually for good performance. In our experiment the same as in literature [8]'s, α is set to 0.2 and β is set to 0.6. It is noted that sim_{Li} (c_1, c_2) will increasing with respect to depth (lso (c_1, c_2)) and decreasing with len (c_1, c_2). The values of sim_{Li} (c_1, c_2) are range from 0 to 1.

2.3.2. Information Content Based Metrics

The notion of information content of the concept is directly related to the frequency of the term in a given document collection. The frequencies of terms in the taxonomy are estimated using noun frequencies in some large collection of texts [14]. The idea behind semantic

similarity information content metrics is that each concept includes much information in WordNet. It assumes that the similarity of two concepts is related to information they share in common. The more common information two concepts share, the more similar the concepts are.

In 1995 Resnik first proposed information content based similarity metric [15]. It assumed that for a concept c,

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

Where p(c) is the probability of encountering and instance of concept c.

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_1, c_2, the similarity is indicated by a highly specific concept that subsumes them both in the taxonomy

$$sim_{\mathrm{Re}snik}(c_1, c_2) = -\log p(lso(c_1, c_2)) = IC(lso(c_1, c_2) \tag{7}$$

It is noticed that $sim_{Resnik}(c_1, c_2)$ is depended by concept pairs' most specific subsumer in the taxonomy. Lin took the IC of compared concepts into account respectively and proposed another metric for similarity metric [16]. This metric uses both the amount of information needed to state the commonality between the two concepts and the information needed to fully describe these concepts.

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(lso(c_1, c_2))}{IC\ (c_1) + IC\ (c_2)} \tag{8}$$

We can see that:

(1) A term compared with itself will always score 1.

(2) The similarity values are range from 0 to 1.

Contrary to the above similarity metrics, Jiang proposed a metric from a different point of view. He calculated semantic distance to obtain semantic similarity [17]. Semantic similarity is the opposite of the distance.

$$dis_{Jiang}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2IC(lso(c_1, c_2)) \tag{9}$$

It is noted that the IC value of each concept is an important dimension in assessing the similarity of two concepts or two words and provides an estimation of its abstract or specialty. Generally speaking, there are two methods to obtain IC. One is Corpora-dependent IC metric. Corpora-dependent IC metric obtains IC through statistical analysis of corpora. The other is Corpora-independent IC metric. Recent years the latter has drawn great concern. One commonly used IC model was proposed by Nuno. The model use WordNet as a statistical resource to compute the probability of occurrence of concepts. It is based on the assumption that the taxonomic structure of WordNet is organized in a meaningful and structured way, where concepts with many hyponyms convey less information than concepts that are leaves. As of this, the more hyponyms a concept has the less information it expresses. Likewise, concepts that are leaf nodes are the most informative in the taxonomy. In other words, the Information Content of a WordNet concept is commented as a function of the population of its hyponyms[13]. The model is defined as [18]:

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

According to formula (10), two concepts with the same number of hyponyms will have the same IC values.

3. A New Semantic Similarity Metric Based on WordNet

In this section, let's take Figure 1 as example and discuss the metrics stated above.

Firstly, let's discuss path-based metrics.

It is noticed that, len (mail, bicycle) and len (wheeled vehicle, bus) are both equal to 4. For a specific version of WordNet, deep_max is a fixed value. Therefore, the two pairs will have the same similarity value with L&C's metric. Another fact must be noted that, both lso (mail, bicycle) and lso (wheeled vehicle, bus) are conveyance. This fact make the two pairs will have the same similarity value with Wu&Palmer's metric and Li's metric, too.

Next, let's analyze information content based metrics.

If two pairs have the same most specific subsumer, they will have same similarity values with Resnik's metric. For example, in Figure 1, sim(bus, train) is equal to sim(bus, boat train).

Lin's metric and Jiang's metrics have taken the IC of compared concepts into account respectively. If the summation of IC of compared two pairs with the same lowest subsumer is equal, they will have the same similarity values. For example, in Fig.1 the IC value of all the leaves is equal to 1 according to formula (10), which makes the similarity values of pairs (mail, car) and pairs (school bus, car) are equal.

Based on stated above, it is noted that the similarity metric could not distinguish different concepts pairs effectively. There is still room for improvement.

Here a new metric is presented, which takes account not only path length, but also local density information. It combines information content and paths of concepts, formally:

$$sim_{new}(c_1, c_2) = \left(\frac{2 * IC(lso))}{IC\ (c_1) + IC\ (c_2)}\right)^{\left(\frac{1-e^{-k*len(c_1,c_2)}}{e^{-k*len(c_1,c_2)}}\right)} \tag{11}$$

Where k is a parameter, and 0<=k<=1, which can be adapted manually to make the metric to get the best performance.

We can see that,

(1) $Sim_{new}(c_1,c_2)$ is inversely proportional to $len(c_1,c_2)$. If $len(c_1,c_2)$ is 0, $sim_{new}(c_1,c_2)$ get the maximum value of 1.

(2)Because $0<= 2 * IC(lso)) /(IC\ (c_1) + IC\ (c_2)) <=1$, as $len(c_1,c_2)$ increases to ∞, $sim_{new}(c_1,c_2)$ is close to 0.

(3) Therefore the values of $sim_{new}(c_1,c_2)$ are range from 0 to 1.

(4) If two pairs have the same most specific subsumer and the sum of IC are the same too, but their length path are not equal, they will have different similarity values.

In next section, we will analyze our new metric from different perspectives.

4. Evaluation

In this section, we evaluated the results by correlating our similarity values with that of human judgments.

4.1. Data set and Words Similarity Calculating Method

In the experiment, the dataset provided by Rubenstein and Goodenough (1965) [19] was adopted. In R&G's study, 51 undergraduate subjects were asked for rate 65 pairs of words, which ranged from "highly synonymous" to "semantically unrelated". Subjects were asked to rate them on the scale of 0.0 to 4.0. In formula (11), the k value is very important, which decides the performance of the metric. We compute different correlation coefficients between proposed metric and human judgments corresponding to different k values. Because either or both of the words have more than one sense in WordNet, in the result, we took the most similarity pair of sense.

$$sim(Word_1, Word_2) = \max_{(i,j)}[sim(c_{1i}, c_{2j})]$$

Experiments show that when k is 0.08, the correlation gets the maximum value 0.8817.

4.2. Results Analysis

Before our analysis we first compute semantic similarity between pairs of words with the six chosen metrics listed in Section 2 and the new metric. Then the semantic similarity distributed graph with different metrics is illustrated in Figure 2. For the convenience of expression and comparison, the values are normalized in [0, 1]. The IC value is obtained according formula (10).

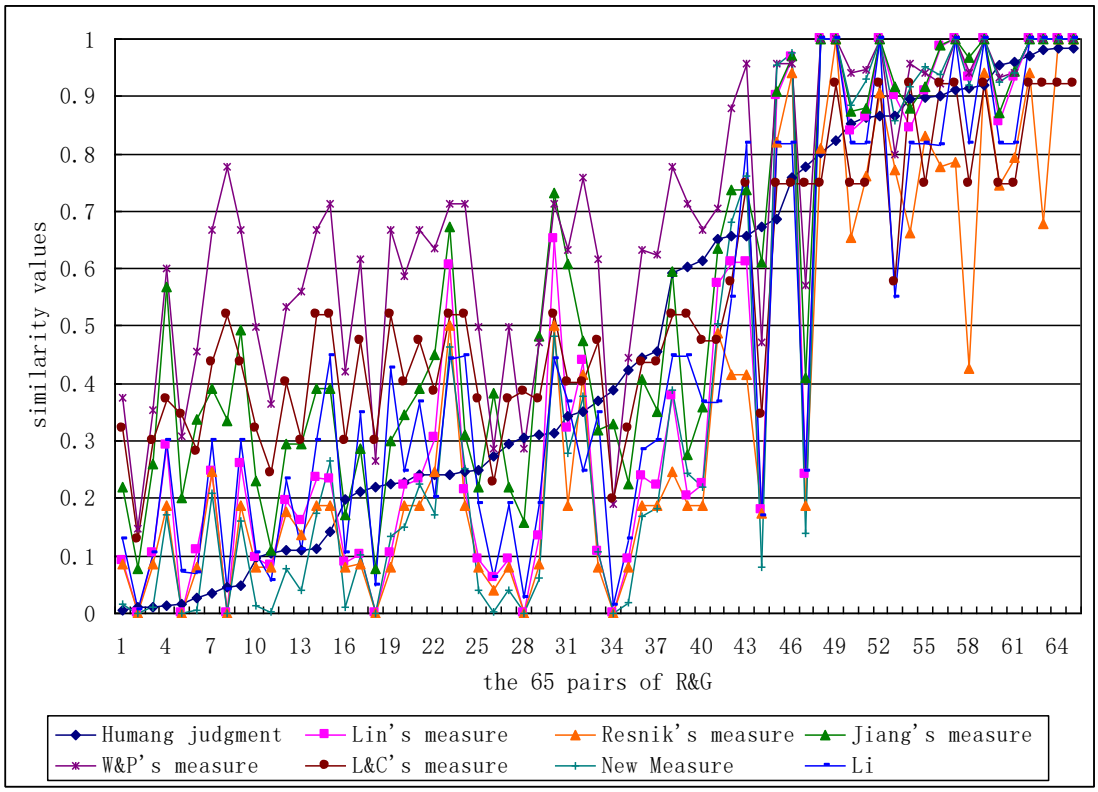

Figure 2. Semantic Similarity Distributed Graph with Different Metrics

In accordance with previous research, we compare the six chosen metrics with our new algorithm by calculating the coefficients of correlation with human judgments of semantic similarity. The results are shown in Table 1.

Table 1. Coefficients of Correlation between Human Ratings of Similarity

Semantic Similarity Metric	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 Metric	0.8817

For the convenience of comparison intuitively, the compared results of our proposed metric with other six metrics are provided in Figure 3.

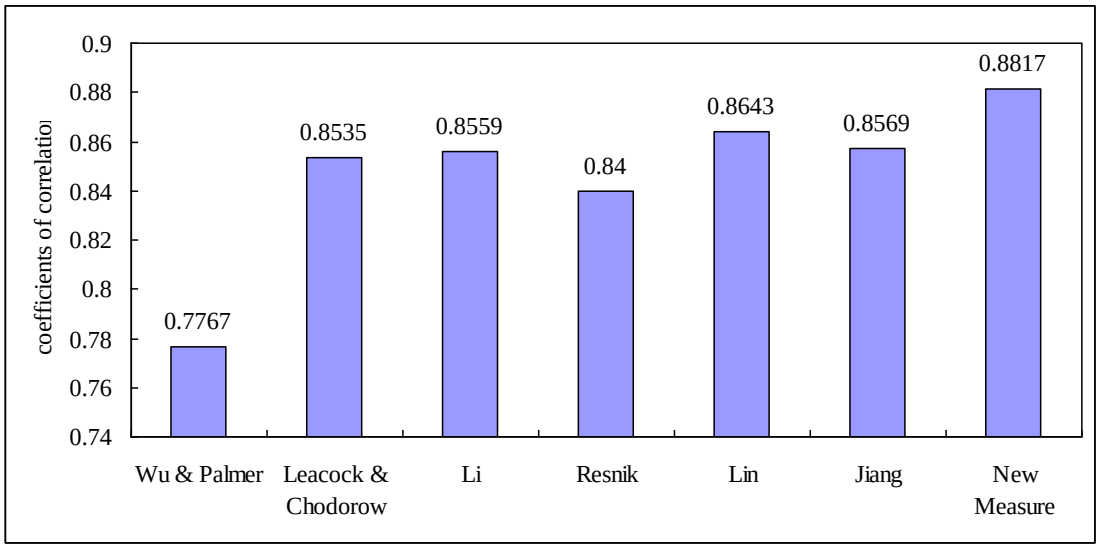

Figure 3. The Compared Results of our Proposed Metric with other Six Metrics

5. Conclusion and Future Work

This paper presents metric for measuring word sense similarity using path and information content. It combines path based metric and information content based metric. Different from previous works, in the new metric not only the semantic density information, but also the path information has been reflected. We evaluate our model on the data set of Rubenstein and Goodenough and compare the results of our proposed metric with Wu&palmer's metric, Leacock&Chodorow' metric, Li's metric, Resnik's metric, Lin's metric and Jiang's metric. The distributed graphs of 65 word pair's similarity value with different metrics are illustrated. Experiments show that the coefficient of our proposed metric with human judgment is 0.8817, which is significantly outperformed than related works. In future work, we will put the metric into query suggestion, ontology construction, documents clustering and so on for practical application.

Reference

[1] 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, Mexico City, Mexico, **(2003)** February 16-22.
[2] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst, vol. 22, no. 7, **(2009)**.
[3] 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)**.
[4] C. Y.Lin and E. Hovy, "Automatic evaluation of summaries using ngram co-occurrence statistics", Proceedings of Human Language Technology Conference, Canada, Edmonton, **(2003)** May 27-June 1.
[5] 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 J. 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)**.
[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] P. Gomes, N. Seco, F. C. Pereira, P. Paiva, P. Carreiro, J. L. Ferreira and C. Bento, "The importance of retrieval in creative design analogies", Procceedings of the International Joint Conference on Artificial Intelligence, Acapulco, Mexico, **(2003)** August 9-15.
[8] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology", Procceedings of the 8th Pacific Symposium on Biocomputing, Kauai, Hawaii, **(2003)** January 3-7.
[9] C. Fellbaum, "WordNet: An electronic lexical database. MIT Press, **(1998)**.
[10] Z. Wu and M. Palmer, "Verb semantics and lexical selection", Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, **(1994)** June 27-30.
[11] 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
[12] 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)**.
[13] Giannis Varelas Epimenidis Voutsakis Paraskevi Raftopoulou Euripides G.M. Petrakis Evangelos E. Milios , 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, (2005) October 31- November 05, Bremen, Germany.
[14] P. Resnik, "Semantic Similarity in a Taxonomy: An information-based measure and its application to problems of ambiguity and natural language", Journal of Artificial Intelligence Research, vol.11, no.0, **(1999)**.
[15] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal Quebec, Canada, **(1995)** August 20-25.
[16] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, USA, **(1998)** July 24-27.
[17] 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, Taipei, Taiwan, **(1997)** August 22-24.
[18] N. Seco, T. Veale and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet", Proceedings of the 16th European Conference on Artificial Intelligence,Valencia, Spain, **(2004)** August 22-27.
[19] H. Rubenstein and John B. Goodenough, "Contextual correlates of synonymy", Communications of the ACM, vol. 8, no. 10, **(1965)**.

Authors

Lingling Meng, is an associate professor of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Runqing Huang, has a PhD from Shanghai Jiao Tong University. He works in Shanghai Municipal People's Government, P. R. China. His present research interests include modeling strategic decisions, economic analysis, electronic government and Logistics.

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.