# An Ontology Concept Update Method Based on Hybrid Semantic Similarity

**Peng Zhang[a], Jiahui Qi[b] and Min Wu[c]**

Center of Modern Educational Technology, University of Science and Technology of China, Hefei 230026, China

[a]zp160@mail.ustc.edu.cn, [b]jhqi@ustc.edu.cn, [c]minwu@ustc.edu.cn

**Keywords:** Ontology, Concept update, Semantic similarity, WordNet

**Abstract:** With the change of information, the existing ontology cannot satisfy users' requirements. In order to realize the extension of ontology concepts and reduce the excessive dependence on domain experts, an ontology concept update method is proposed. Based on WordNet, the calculation of IC(Information Content) is improved, and then we propose a new hybrid measurement to calculate semantic similarity. Due to the data and object properties of ontology are ignored by the semantic similarity methods, property similarity is added to make an adjustment. Through experiments and comparison, the Pearson correlation coefficient between the improved semantic similarity method and the standard datasets is higher than other related methods, and the result is closer to human subjective judgment. The ontology concept update method is used to analyze a constructed ontology of academic conference. The results show that this method can update the ontology concept, and has certa.

## 1. Introduction

The ontology allows users and computers to communicate more accurately through semantics rather than syntax. With the construction and application of various domain ontologies, there exist some restrictions on the development of ontologies. Domain ontology relies on experts to build it manually, and limited resources may lead to the loss of some entities and relations in the process of ontology construction [1]. In addition, with the constant update of information, the existing ontology also needs to be added with new concepts [2]. In order to make ontology have as many entities and properties as possible, ontology concept update is needed to improve and enrich the content of ontology. At present, the methods of adding new concepts to ontology can be divided into updating by experts manually or updating automatically using algorithms [3]. This paper proposes an ontology concept update method based on hybrid semantic similarity.

With the further study in ontology, more researchers propose to calculate concept similarity with a structured domain ontology, the semantic information in WordNet ontology is widely used in similarity measurement. At present, the similarity calculation methods based on WordNet are mainly divided into four types: distance-based, feature-based, IC-based and hybrid-based method [4, 5]. The distance-based method utilizes the shortest path between two concept nodes, and the longer the distance, the lower the similarity. The accuracy of this method is usually not very high because the distance-based method uses less semantic information. The feature-based method considers the shared properties between two concepts, such as comparing the coverage of two concepts in WordNet. The process of this method is often sophisticated and the results are not accurate enough, so it is not widely used. The IC-based method considers the semantic information contained in the concept node, and the IC value is calculated with a corpus. For two given concepts, if they share more information, they are more similar to each other. Since this method relies on a corpus, it is necessary to consider the domain to which the ontology belongs when selecting the corpus. In recent years, the hybrid-based method has been widely applied. The hybrid-based method combines the advantages of various methods, effectively utilizes the hierarchical structure between concept nodes and the information contained in concepts. This method usually obtains more accurate results.

The following is a brief introduction to this paper. The second part introduces some related work and our new IC model, then proposes a new hybrid semantic similarity method based on our IC model. The third part introduces an ontology concept update method. The fourth part shows the result of related experiments. The last part is a summary of this paper.

## 2. A Hybrid Semantic Similarity Method based on WordNet

### 2.1 Related Work

WordNet [6] is an English semantic dictionary which is widely used. It expresses all terms and concepts in the form of synonym sets [7], each synonym set has a brief definition description and a semantic relationship of the synonym set. WordNet provides a hierarchical structure for each term, so it is suitable for measuring semantic similarity. A part of WordNet 3.0 is shown in Fig. 1.

Since some terms have multiple definitions, they may appear in different concept nodes. If $w_1$ and $w_2$ are two words, $synset(w_1)$ and $synset(w_2)$ are the synonym sets to which the two concepts belong, we define the semantic similarity as:

$$Sim(w_1, w_2) = \max_{c_1 \in synset(w_1),\ c_2 \in synset(w_2)} \{Sim(c_1, c_2)\}$$

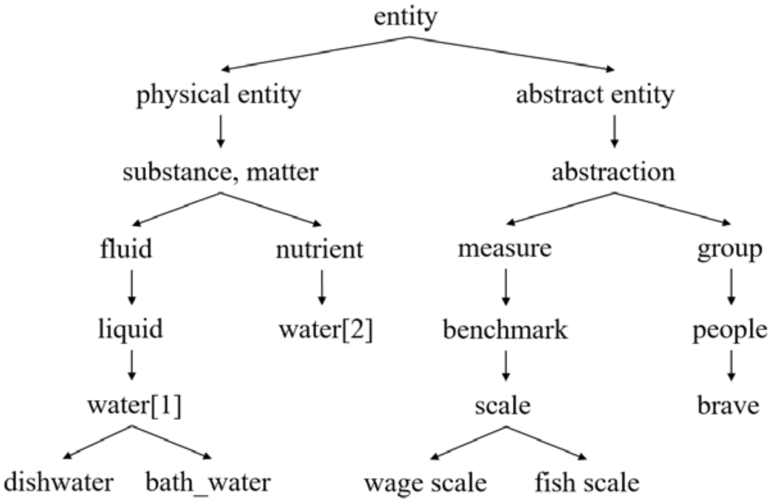

Figure 1. A part of WordNet 3.0

Here, we define some notions used below:

$Sim(c_i, c_j)$: The similarity between concept $c_i$ and $c_j$.

$length(c_i, c_j)$: The length of the shortest distance from concept $c_i$ to $c_j$ in WordNet, e.g. $length(scale, group) = 4$.

$Depth(c_i)$: The length of the distance from concept $c_i$ to the root $entity$, e.g. $Depth(fluid) = 3$.

$Depth_{\max}$: The maximum depth of all the concept nodes in WordNet.

$LCA(c_1, c_2)$: The lowest common ancestor of $c_i$ and $c_j$, e.g. $LCA(measure, group) = abstraction$.

$IC(c_i)$: The information content of synset $c_i$.

$P(c_i)$: The probability of concept $c_i$ appearing in WordNet.

$Distance(c_i, c_j)$: The IC distance between concept $c_i$ and $c_j$.

$hypo(c)$: The number of hyponyms of concept $c_i$, e.g. $hypo(benchmark) = 4$

$NodeMax$: The number of all the concept nodes in WordNet.

Leacock and Chodorow [8] considered the distance between the two concepts and the depth of the concept node in the semantic dictionary:

$$Sim(c_1,c_2) = -\log(\frac{length(c_1,c_2)}{2\times Depth_{\max}})$$

However, this method has the same result when the two concepts have the same distance. Therefore, Wu and Palmer [9] considered using the depth of their LCA to solve this problem:

$$Sim(c_1,c_2) = \frac{2\times Depth(LCA(c_1,c_2))}{2\times Depth(LCA(c_1,c_2)) + length(c_1,c_2)}$$

Resnik's [10] IC-based approach is expressed that calculating the IC value of their LCA between the two concept nodes, and then the IC value can represent the similarity:

$$Sim(c_1,c_2) = IC(LCA(c_1,c_2))$$

$$IC(c) = -\log(P(c))$$

Lin [11] improved Resnik's approach that if there are two different concept pairs with the same lowest common ancestor, these two pairs will have the same IC value, so it is necessary to consider the IC value of the concept itself:

$$Sim(c_1,c_2) = \frac{2\times IC(LCA(c_1,c_2))}{IC(c_1)+IC(c_2)}$$

Jiang and Conrath [12] proposed a semantic distance metric using IC value:

$$Distance(c_1,c_2){=}IC(c_1)+IC(c_2)-2\times IC(LCA(c_1,c_2))$$

$$Sim(c_1,c_2) = 1/\,Distance(c_1,c_2)$$

Nuno's method [13] proposed that if a concept has more hyponyms, it can have less information content, so the method's IC value is expressed by:

$$IC(c) = 1-\frac{\log(hypo(c)+1)}{\log(NodeMax)}$$

If the numbers of hyponyms are the same, then Nuno's method cannot distinguish them. Zhou et al. [14, 15] improved Nuno's method that they use the depth of the concept itself, the IC value and the semantic similarity are expressed below with $k$ = 0.5:

$$IC(c) = (1-k)\times(\frac{\log(Depth(c))}{\log(Depth_{\max})}) + k\times(1-\frac{\log(hypo(c)+1)}{\log(NodeMax)})$$

$$Sim(c_1,c_2) = 1-k\times(\frac{\log(length(c_1,c_2)+1)}{\log(2\times Depth_{\max}-1)}) - (1-k)\times(\frac{IC(c_1)+IC(c_2)-2\times IC(LCA(c_1,c_2))}{2})$$

### 2.2 A New IC Model

Many studies show that the semantic similarity results are more accuracy by using the IC value based on WordNet. With the methods expressed above, we can see that the IC model cannot distinguish different concepts effectively. From Fig. 2, it is noticed that two concepts may have the same number of hyponyms, but the structures of hyponyms differ. Take $C_5$ and $C_6$ as an example, the hyponyms of $C_5$ and $C_6$ are both 2, but with Nuno's or Zhou's method, the IC value of $C_5$ and $C_6$ are the same.

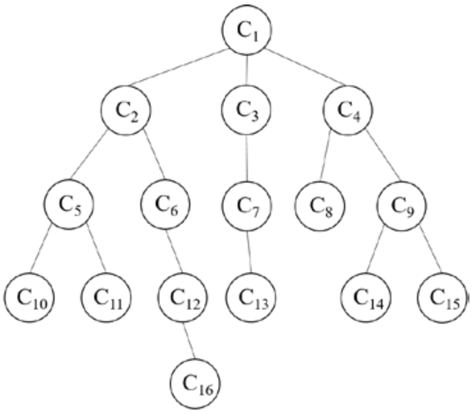

Figure 2. An abstraction of hyponyms in WordNet 3.0

So we assume that the information content is also depended on the structure of hyponyms. So, the IC value is depended on the number of hyponyms, the depth of the concept itself and the structure of hyponyms. From above, a new IC model is proposed, referring to Meng et al. [16, 17]:

$$IC_{new}(c) = (1-\frac{\log(\sum_{w\in hypo(c)}\frac{1}{Depth(w)}+1)}{\log(Depth_{\max})})\times\frac{e^{\theta\times Depth(c)}-e^{-\theta\times Depth(c)}}{e^{\theta\times Depth(c)}+e^{-\theta\times Depth(c)}} \qquad (1)$$

In formula (1), $\theta$ is a weighted factor and $\theta > 0$. When the concept is a root node, $IC_{new}(c) = 0$, the IC value ranges from 0 to 1.

### 2.3 A Hybrid Semantic Similarity Method

From above, we analyze the advantages and disadvantages of different semantic similarity methods. Distance-based methods are simple but not very accurate, they consider the hierarchy of the concepts. IC-based methods consider the semantic information of the concept but ignore the hierarchical information. Therefore, we need to use information content to make up the disadvantage of distance-based methods, and propose a new hybrid semantic similarity method:

$$Sim_{ours}(c_1,c_2) = (1+length(c_1,c_2)\times\frac{Depth_{\max}}{2\times Depth(LCA(c_1,c_2))}\times e^{-\alpha\times\frac{IC_{new}(LCA(c_1,c_2))}{IC_{new}(c_1)+IC_{new}(c_2)}})^{-1} \qquad (2)$$

In formula (2), $\alpha$ is a weighted factor and $\alpha > 0$, which can adapt during the experiment.

Our method considers both the hierarchy in WordNet and the semantic information of the concept. The method combines the distance between two concepts, the depth of their LCA, and the IC value of each concept using our new IC model. It can effectively improve the accuracy of the results.

## 3. An Ontology Concept Update Method based on Semantic Similarity

Ontology concepts are different from the concepts in WordNet, they are usually described by properties from two types: data properties and object properties. If two concepts have more properties in common and fewer different properties, the two concepts are more similar to each other. If two concepts have completely different properties, it means there is no connection between them. The ontology concept similarity contains the semantic similarity and the property similarity. From above, we propose a semantic similarity method and here, we propose a method to calculate the property similarity. For two given concepts, A is an ontology concept, and B is a new concept to be added.

*Step 1*: Convert each property into a word set through word splitting, stopwords removing, abbreviation restoring and stemming. Because the property might be a word composed by abbreviations. For example, a property name is $numOfAttendees$:

$$numOfAttendees \xrightarrow{word\ splitting} (num, Of, attendees) \xrightarrow{stopwords\ removing} (num, attendee)$$
$$\xrightarrow{abbreviation\ restoring} (num, attendees) \xrightarrow{stemming} (number, attendee)$$

Therefore, the property sets of concept A and B can be transferred into two word sets. It is expressed by:

$$property_A = \{a_1, a_2, \ldots, a_m\},\ property_B = \{b_1, b_2, \ldots, b_n\} \tag{3}$$

In formula (3), $a_i (1 \le i \le m)$ represents a word in the property of concept A, $b_j (1 \le j \le n)$ represents a word in the property of concept B, m and n represent the number of words converted from the property of concept A and B.

*Step 2*: Constructing the similarity matrix. $sp_{ij}$ Represents the semantic similarity between the word $a_i$ and $b_j$:

$$\boldsymbol{SP} = \begin{pmatrix} sp_{11} & sp_{12} & \cdots & sp_{1n} \\ sp_{21} & sp_{22} & \cdots & sp_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ sp_{m1} & sp_{m2} & \cdots & sp_{mn} \end{pmatrix},\quad sp_{ij} = Sim_{ours}(a_i, b_j) \tag{4}$$

*Step 3*: Traverse the matrix and get the largest similarity score, then delete all the elements in the row and column where $sp_{ij}$ locates to get the matrix $\boldsymbol{SP'}$. Repeat Step 3 until the matrix is null, we can get a sequence of the largest similarity score, recorded as $d = \{d_1, d_2, \cdots, d_k\},\ k = \min(m, n)$. Step 3 actually finds the most similar pair of words every time.

*Step 4*: Calculate the property similarity between concept A and concept B, using the sequence $d$:

$$Sim_{prop}(A, B) = \frac{1}{k}\sum_{i=1}^{k} d_i \tag{5}$$

Therefore, we can get the property similarity of the ontology concepts, plus the semantic similarity of the concepts, we propose an ontology concept similarity method:

$$Sim_{onto}(A, B) = \gamma \times Sim_{ours}(A, B) + (1-\gamma) \times Sim_{prop}(A, B) \tag{6}$$

In formula (6), $\gamma$ represents a weighted factor, and $\gamma > 0$. From the experimental results, when $Sim_{ours}(A, B) = 0$, $\gamma = 0.1$; When $Sim_{prop}(A, B) = 0$, $\gamma = 0.9$; Otherwise, $\gamma = 0.5$.

Therefore, the process of ontology concept update method is proposed below:

```
Algorithm: Ontology Concept Update Method
Input: A concept $C_{new}$ to be added to the ontology, the property set of $C_{new}$ is
    $property_{new} = \{p_1, p_2, ..., p_u\}$
Output: The most similar concept $C_{output}$ in the ontology (the number of the concept is $n$) to $C_{new}$
a) Use formula (2) to calculate the semantic similarity between $C_{new}$ and every concept in the ontology, then get a sequence of concept semantic similarity: $sim\_set_{ours} = \{x_1, x_2, ..., x_n\}$;
b) Use formula (4) and (5) to calculate the property similarity between $C_{new}$ and every concept in the ontology, then get a sequence of concept property similarity: $sim\_set_{prop} = \{y_1, y_2, ..., y_n\}$;
c) Use formula (6) to calculate the ontology similarity between $C_{new}$ and every concept in the ontology, then get a sequence: $sim\_set_{onto} = \{u_1, u_2, ..., u_n\}$.
d) Get the largest similarity score from $sim\_set_{onto}$, then get the most similar concept $C_{output}$.
e) Generate a new concept node in the ontology, and then add $C_{new}$ to it.
```

Through the above operations, the ontology concept can be updated automatically.

## 4. Evaluation

In this part, we use some standard datasets to verify the validity of our new IC model and the hybrid semantic similarity method. The experiment is based on WordNet 3.0 with the SemCor Corpus [18]. We choose MC30 [19], RG65 [20], WS353 [21], WS353-sim203 [22], SimLex999 [23], Mtruk287 [24], which are all standard datasets used to evaluate different semantic similarity methods. Every line of the datasets is represented by a triple set $(Word_1, Word_2, score)$, it means that the manual rating of the similarity between $Word_1$ and $Word_2$ is $score$. We use the Pearson correlation coefficient as a measurement to evaluate the quality of different IC models and similarity methods. If the correlation coefficient of the model or method is larger, the result is more precise.

### 4.1 Verifying the validity of the New IC Model

We use Lin's method and MC30 dataset to experiment and discover that when $\theta = 0.2$ and $\alpha = 2.5$, the correlation coefficient is the largest. The following experiment also uses the results of these weighted factors. In Table 1, Lin (corpus), Lin (Nuno), Lin (Zhou), Lin (Meng) represents the IC model based on SemCor corpus, Nuno, Zhou et al. and Meng et al. The comparison between our IC model and others is shown in Table 1.

From Table 1, we can discover that our IC model gets a larger correlation coefficient than others, the new IC model is more suitable for calculating semantic similarity based on WordNet.

Table 1. Comparison of different IC models using Lin's method on MC30

| Similarity Method | Pearson Correlation Coefficient |
|---|---|
| Lin(corpus) | 0.737 |
| Lin(Nuno) | 0.830 |
| Lin(Zhou) | 0.811 |
| Lin(Meng) | 0.832 |
| Lin(Ours) | 0.841 |

### 4.2 Experiment of the Improved Hybrid Semantic Similarity Method

We compare different similarity methods on standard datasets. The results are shown in Table 2.

Table 2. Comparison of different semantic similarity methods on standard datasets

| Similarity Method | MC30 | RG65 | WS353 | WS353-sim203 | SimLex999 | Mtruk287 |
|---|---|---|---|---|---|---|
| Wu&Palmer | 0.722 | 0.733 | 0.216 | 0.468 | 0.175 | 0.422 |
| Jiang&Conrath | 0.473 | 0.575 | 0.227 | 0.326 | 0.228 | 0.117 |
| Leacock&Chodorow | 0.784 | 0.841 | 0.315 | 0.581 | 0.281 | 0.370 |
| Lin | 0.747 | 0.737 | 0.268 | 0.502 | 0.401 | 0.399 |
| Resnik | 0.811 | 0.824 | 0.341 | 0.627 | 0.344 | 0.450 |
| Zhou | 0.812 | 0.841 | 0.343 | 0.657 | 0.402 | 0.439 |
| Meng | 0.823 | 0.854 | 0.367 | 0.649 | 0.397 | 0.514 |
| Ours | 0.850 | 0.872 | 0.394 | 0.678 | 0.448 | 0.480 |

From Table 2, we discover that our semantic similarity method receives a good effect on the small dataset MC30 and RG65. However, when the dataset is larger, the effect of the method is not very good. We infer that there exist some word pairs with unexpected results comparing to the expected score. When the dataset is getting larger, the number of these word pairs increases and influences the final results. However, our method gets a better score on different standard datasets comparing to other methods, and it improves the accuracy of semantic similarity calculation.

### 4.3 Experiment of the Ontology Concept Update Method

We do some experiments on an academic ontology called “Conference” from MapOnto [25] and expand some properties to the concept. The ontology contains 27 ontology concepts and 143 ontology properties. A part of the ontology is shown in Fig. 3. We prepare some concepts to be added which is shown in Table 3, and due to the results are quite large, we just list the most similar concept matching with the new concept. The results are shown in Table 4.

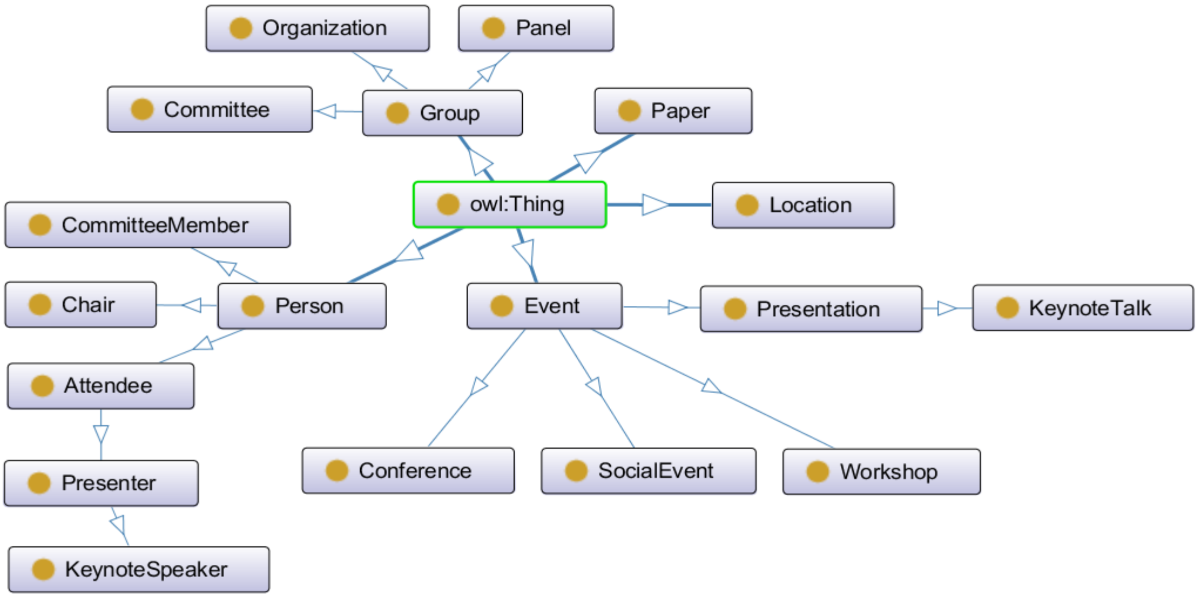

Figure 3. An ontology of academic conference

Table 3. Comparison of different semantic similarity methods on standard datasets

| Name | Data Properties | Object Properties |
|---|---|---|
| *co_chair* | {*name*, *eventTitle*, *venue*} | *null* |
| *meeting* | {*title*, *time*, *year*, *numOfAttendees*} | {*hasVenue*} |
| *report* | {*title*, *speaker*, *startTime*, *endTime*} | {*hasPresenter*} |

Table 4. The results of Ontology Concept Update Method

| $C_{new}$ | $C_{output}$ | $Sim_{ours}(C_{new}, C_{output})$ | $Sim_{prop}(C_{new}, C_{output})$ | $Sim_{onto}(C_{new}, C_{output})$ |
|---|---|---|---|---|
| *co_chair* | *chair* | 0 | 0.817 | 0.735 |
| *meeting* | *conference* | 0.906 | 0.715 | 0.811 |
| *report* | *presentation* | 0.773 | 0.421 | 0.597 |

From Table 4, we can see that, the ontology can expand properties when using our ontology concept update method, the ontology similarity can make up the lack of semantic similarity, e.g. co_chair. And our method fits the manual judgment, it can reduce the dependence on the experts.

## 5. Conclusion

A new IC model based on WordNet is proposed in this paper, then we construct a hybrid semantic similarity method considering the distance between two concepts, the depth of the concept itself, and the information content. Comparing to other different similarity methods, our IC model is more suitable for calculating the information content, and the method gets a better correlation coefficient on the standard datasets. Finally, we propose an ontology concept update method using property similarity to make up the lack of semantic similarity. From the experimental results, we verify that the ontology update method is effective for expanding new concepts and can reduce the dependence on the experts.

## References

[1] Saia R, Boratto L, Carta S. Introducing a weighted ontology to improve the graph-based semantic similarity measures [J]. INTERNATIONAL JOURNAL OF SIGNAL PROCESSING SYSTEMS, 2016, 4 (5): 375 - 381.

[2] Liu K, Mitchell K J, Chapman W W, et al. Formative evaluation of ontology learning methods for entity discovery by using existing ontologies as reference standards[J]. Methods of information in medicine, 2013, 52 (04): 308 - 316.

[3] Yun Zhou, Dong Liu. Concepts upgrading method of domain ontology based on semantic similarity [J]. Computer Engineering and Design, 2011, 32 (8): 2833 - 2835.

[4] GUO Xiaohua, PENG Qi, DENG Han, et al. Edge weight-based word similarity computation in WordNet. Computer Engineering and Applications, 2018, 54 (1): 172 - 178.

[5] GAO J B, Zhang B W, Chen X H. A WordNet-based semantic similarity measurement combining edge-counting and information content theory [J]. Engineering Applications of Artificial Intelligence, 2015, 39: 80 - 88.

[6] Miller G A. WordNet: a lexical database for English [J]. Communications of the ACM, 1995, 38 (11): 39 - 41.

[7] Zhu X, Li F, Chen H, et al. A Novel WordNet-based Approach for Measuring Semantic Similarity⋆ [J]. Journal of Information & Computational Science, 2015, 12 (13): 4919 - 4927.

[8] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification [J]. WordNet: An electronic lexical database, 1998, 49 (2): 265 - 283.

[9] Wu Z, Palmer M. Verb's semantics and lexical selection[C]//Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994: 133 - 138.

[10] Resnik P. Using information content to evaluate semantic similarity in a taxonomy [J]. arXiv preprint cmp-lg/9511007, 1995.

[11] Lin D. An information-theoretic definition of similarity[C]//Icml. 1998, 98(1998): 296 - 304.

[12] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy [J]. arXiv preprint cmp-lg/9709008, 1997.

[13] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet [C]//ECAI. 2004, 16: 1089.

[14] Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet[C]//Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. IEEE, 2008, 3: 85 - 89.

[15] Zhou Z, Wang Y, Gu J. New model of semantic similarity measuring in wordnet [C]//Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on. IEEE, 2008, 1: 256 - 261.

[16] Meng L, Gu J, Zhou Z. A new model of information content based on concept’s topology for measuring semantic similarity in WordNet [J]. International Journal of Grid and Distributed Computing, 2012, 5 (3): 81 - 94.

[17] Meng L, Huang R, Gu J. An effective algorithm for semantic similarity metric of word pairs [J]. Issues, 2016 (468).

[18] Information on: http://web.eecs.umich.edu/~mihalcea/downloads.html

[19] Miller G A, Charles W G. Contextual correlates of semantic similarity [J]. Language and cognitive processes, 1991, 6 (1): 1 - 28.

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

[21] Finkelstein L, Gabrilovich E, and Matias Y, et al. Placing search in context: The concept revisited [J]. ACM Transactions on information systems, 2002, 20 (1): 116 - 131.

[22] Agirre E, Alfonseca E, Hall K, et al. A study on similarity and relatedness using distributional and wordnet-based approaches[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 19 - 27.

[23] Hill F, Reichart R, and Korhonen A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation [J]. Computational Linguistics, 2015, 41 (4): 665 - 695.

[24] Radinsky K, Agichtein E, Gabrilovich E, et al. A word at a time: computing word relatedness using temporal semantic analysis[C]//Proceedings of the 20th international conference on World Wide Web. ACM, 2011: 337 - 346.

[25] An Y, Borgida A, Mylopoulos J. Inferring complex semantic mappings between relational tables and ontologies from simple correspondences [C]//OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2005: 1152 - 1169.