The Improved Algorithm of Semantic Similarity Based on the Multi-dictionary

Chijun Zhang[1,2,3] ,Yongjian Yang[2,3*] , Xiaoyu Guo[4], Zhanwei Du[3], Na Lin[5]
1 College of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
2 Key Laboratory of Logistics Industry Economy and Intelligent Logistics at Universities of Jilin Province, Changchun 130117, China;
3 College of Computer Science and Technology, Jilin University, Changchun 130012, China
4 College of Communication Engineering, Jilin University, Changchun 130012, China
5 Jilin Architecture and Civil Engineering Institute of computer, Changchun 130012, China
Email: cjzhang6@163.com

Abstract—**Traditional calculations of word similarity are based on a single dictionary, and ignore the coordination between dictionaries. Therefore in order to improve the reliability of similarity calculation methods based on the dictionary, this paper proposes a similarity calculation method based on multi-dictionary, namely through the combination of multiple dictionaries, to improve the reliability of the calculation method of semantic similarity, and meanwhile give the similarity calculation method and steps.**

Index Terms—**multi-dictionary, semantic similarity, reliability, improved algorithm**

I. Introduction

In Chinese, the same words in different contexts may have different semantics, namely semantic diversification which makes it difficult for the automatic processing of natural language. Thus in practical applications, in order to make quantitative distinction of the semantics, sometimes it is necessary to use a simple numerical measure of the semantic closeness between words, and word semantic similarity (hereinafter referred to as the word similarity. The semantic similarity or semantic relatedness is considered as a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning/semantic content.) is one of them and it has different meanings in different application areas[1, 2]. For example, in the field of data integration, similarity generally refers to the matching degree among texts; while in the field of information retrieval, the similarity reflects the matching degree in the semantic with the user's query, and the higher is the similarity, it indicates that the closer are the text and the user’s request.

Nowadays, the word similarity calculation has a wide range of applications in machine translation, information retrieval, information extraction, word meaning disambiguation, text clustering, ontology mapping, and other fields [3, 4]. To solve such problems, many scholars have done lots of work and put forward a number of qualitative and quantitative methods, it generally includes two types of basic methods: methods based on the world knowledge or some kind of classification and methods of context vector as well as space mode based on statistics. Literature [5] utilizes the concepts similarity to calculate the similarity of the concepts, literature [6] uses the average mutual information to calculate the similarity between words, the literature [7] adopts the statistical method of joint probability distribution for concepts instance to determine semantic similarity between concepts, the literature [8] utilizes search engines as a corpus to calculate the similarity among concepts, etc. These calculations are relatively straightforward, but rely on complete large-scale semantic dictionaries and meanwhile most traditional calculations of word similarity are based on a single dictionary, and ignore the coordination between dictionaries.

In this paper, a modified similarity calculation method based on multi-dictionary is proposed to improve the reliability of similarity calculation methods based on the dictionary, namely through the combination of multiple dictionaries, to improve the reliability of the calculation method of semantic similarity.

The remainder of this paper is organized as follows: in section 2, the related word similarity calculations based on semantic dictionary are discussed and presented; in section 3, the improved similarity calculation methods based on multi-dictionary and the experimental analysis and performance comparison are showed. And finally, in section 4 the conclusion is drawn.

II. Related Word Similarity Calculations based on Semantic Dictionary

A. The Basic Idea of the Word Similarity Calculation based on Semantic Dictionary

The word similarity calculation method based on semantic dictionary is a rationalist approach based on linguistics and artificial intelligence, which utilizes semantic dictionaries, based on hyponymy and synonymy relations between the concepts and obtains the similarity

doi:10.4304/jsw.9.2.324-328

between words by calculating the distance of the two concepts in the hierarchy structure of tree concepts. Approaches based on concept dictionaries built on the fact that two words possess certain semantic correlation, when and only when on the basis of the assumption that existing a path in the structure hierarchical network diagram among concepts.

Large-scale semantic computing resources are the basis of word similarity calculation methods based on dictionary; now commonly used dictionaries [9, 10] have WordNet, FrameNet, HowNet and Lesk, etc.

We here mainly briefly elaborate the two frequently adopted dictionaries: HowNet [11, 14] is a common sense knowledge base which treats concepts represented by the Chinese and English words as description object, and regards revealing the relationship between concepts and attributes of concept as basic content. The HowNet possesses two main concepts: “concept” and “sememe”. The "concept" is a description of word semantics; each word can be expressed as a few concepts. The “concept” is described by a “knowledge representation language”, and the “word” used is called “sememe”, the “sememe” is the smallest meaning unit of the description of “concepts” [13]. WordNet [12] is a semantic network of English vocabulary with broad coverage. The nouns, verbs, adjectives and adverbs are organized into a synonym network respectively; each synonym set represents a basic semantic concept, and is also connected by various relationships between these collections.

B. The Word Similarity Calculation Method based on the Semantic Dictionary

The similarity of words is generally calculated through the word distance, word distance is a real number among the (0,), the distance between a word and itself is 0 and the greater is the distance of two words, the lower is its similarity degree[15,16]. Yan Wei, etc [17] proposed a similarity calculation method of English words based on WordNet, by extracting synonym sets, generic information and meaning interpretation of specific words from the dictionary WordNet, after that using the vector space method to calculate the similarity degree of words, which attempts to provide a similar word set for information retrieval; and the set centers on the search term, is arranged in accordance with the size of the similarity values, and ultimately is able to return the retrieval result for users and carries on the problem extension to a certain extent. Liu Qun [18] put forward a similarity calculation method of word semantics based on HowNet after considering that HowNet is a more detailed semantic dictionary and has improved steadily, although HowNet adopted a multidimensional knowledge representation for specific vocabularies, which adds a certain degree of difficulties to word similarity calculation. In the dictionaries like WordNet etc., all similar semantic items can constitute a tree structure, and calculating the distance between semantic items is equivalent to the calculation of the corresponding node distance in the structure. While in the process of semantic similarity calculation of HowNet, there exists multi-sememe nature of word semantics and the sememe contains complex relationships, the relationship is described through specialized knowledge description semantics. Through studying the corresponding grammar and the sememe relationship between vocabularies semantics in the HowNet, thus Liu Qun, etc distinguish their roles in the similarity calculation, study the similarity calculation method, the collection and the characteristic structure of the sememe, and conduct the semantic similarity calculation on the above basis. Agirre and Rigau [19] introduce factors of depth and density of hierarchical tree when making use of WordNet to calculate the English word similarity. They believe that in the word hierarchical tree: for two pairs of nodes with the same path length, the greater is the depth (away from the tree’s root) of node pairs, the smaller is the semantic distance. Because the depth is greater, which means the classification is more detailed.

Although the above methods have made certain achievements, however, they all ignore the combination and coordination among dictionaries, and the similarity calculation method based on different dictionaries for the vocabularies in different areas has different correct rate. Therefore our paper presents a similarity calculation method based on multi-dictionary to meet the demand of Qos service qualities and improve the reliability of the similarity calculation method based on the dictionary. In short, that is through the combination of multiple dictionaries to improve the reliability of the calculation method of semantic similarity.

III. The Improved Similarity Calculation Methods and Analysis based on Multi-Dictionary

First of all, in order to have some knowledge of the multi-dictionary similarity algorithm, the examples are given as follows [20, 21]:

Constructing joint matrix of the correct rate and error rate is as follows:

$$H = \begin{bmatrix} w1r & w2r & \cdots & wir & \cdots & wnr \\ w1w & w2w & \cdots & wiw & \cdots & wnw \end{bmatrix} \tag{1}$$

Where, wir represents the correct rate of the dictionary wi, and wiw represents the error rate of the dictionary wi.

TABLE I.
The Example of Similarity Calculation Methods Based Multi-Dictionary

	w1	**w2**
r	0.7	0.8
w	0.3	0.2

Table 1 shows the example of similarity calculation methods based multi-dictionary, where *w1 and w2* represent the two dictionaries respectively, *w* represents the corresponding error rate of the similarity calculating method based on the corresponding dictionary, *r* represents the corresponding correct rate of the similarity calculation method based on the corresponding dictionary.

TABLE II.
THE CALCULUS TABLE OF THE EXAMPLE OF SIMILARITY CALCULATION METHODS BASED ON MULTI- DICTIONARY

	w2r	**w2w**
w1r	0.56	0.14
w1w	0.24	0.06

Table 2 shows the calculus table of the above example. *w1r* represents the correct rate of the similarity calculation method based on dictionary *w1* with the condition of considering two expressions have the similar relationship, *w1w* represents the error rate of the similarity calculation method based on dictionary *w1* with the condition of considering two expressions have the similar relationship, *w2r* represents the correct rate of the similarity calculation method based on dictionary *w2* with the condition of considering two expressions have the similar relationship, *w2w* represents the error rate of the similarity calculation method based on dictionary *w2* with the condition of considering two expressions have the similar relationship.

Here the probability of the similarity calculation of two words in two dictionaries are correct is 0.56; the probability of the first dictionary calculated correctly and the second dictionary calculated error is 0.14; the probability of the first dictionary calculated error and the second dictionary correct is 0.24; the probability of both two dictionaries miscalculation is 0.06.

TABLE III.
THE RESULTS OF THE SIMILARITY CALCULATION METHOD BASED ON MULTI-DICTIONARY

	w1 || w2
r	0.94
w	0.06

Carrying on the "or" operation of the similarity calculation method based on dictionary *w1* and similarity calculation method based on dictionary *w2* can get the results shown in Table 3. What can be found is that through the “or” operation of multiple dictionaries, can improve the correct rate to 0.94, and with respect to the original 0.8 and 0.7 in Table 1, all have greatly been improved.

The above example can be extended to n dictionaries. Assuming that, known by the principle of inclusion-exclusion:

$$(A1\cup A2\cup\cdots\cup Am)=\sum_{i=1}^{m}Ai-\sum_{\substack{i\le j;\\ i=1}}^{m}(Ai\cap Aj)+\cdots$$
$$+(-1)^{m}(A1\cap A2\cdots\cap Am) \quad (2)$$

$$wr=w1r\cup w2r\cup\cdots\cup wnr \quad (3)$$

$$wr_{\max}=1-\prod_{i=1}^{n}wiw \quad (4)$$

Where, $wr_{\max}$ is the upper limit of the theoretical correct rate and the corresponding information entropy is:

$$IC(wr_{\max})=-\log p(1-\sum_{i=1}^{n}wiw) \quad (5)$$

The mutual cooperation of multiple dictionaries, to some extent, improves the correct rate of similarity algorithms, and meanwhile the above correct rate is under the assumption of independence of each dictionary, and when there are contacts between the dictionaries, the correct rate will be less than $wr_{\max}$

In summary, in order to improve the reliability of the calculation method of semantic similarity, we propose the following similarity algorithm based on multi-dictionary in our article:

The similarity calculation method here plans to adopt the latest algorithm released by Korean scholars:

$$sim_{I}(a,b)=\frac{2\times|DL(a,b)|}{2\times\Vert\vec{a}|+|\vec{b}|)|}+\frac{2\times|IL(a,b)|}{3\times\Vert\vec{a}|+|\vec{b}|)|} \quad (6)$$

And where, $|\vec{a}|$ and $|\vec{b}|$ represent external links of the vocabulary (word), DL represents the direct link between the vocabularies, IL represents the indirect link of vocabularies.

The main steps of this modified semantic similarity algorithm are explained below:

Step1: Initialize the multiple dictionaries;

Step2: Utilize a specific similarity calculation method based on the dictionary and in accordance with a certain order to calculate the similarity of the corresponding words;

Step3: If the similarity of a dictionary is higher than the threshold, then end the calculation, turn to Step4; Otherwise, continue the calculation and turn to Step2.

Step4: Finally, output the result, the similarity takes the highest values of the similarity calculation method of the corresponding dictionary.

IV. CONCLUSIONS

The calculation of word similarity plays an important role in semantic retrieval, machine translation and many other fields. Although the in-depth studies of many scholars have achieved fruitful results, but due to the complexity of Chinese vocabularies representation and the strong subjective of vocabulary semantic concepts and other factors, up to now word similarity calculation is still the content of the in-depth study of computational linguistics. And the traditional word similarity calculations are based on a single dictionary, ignore the combination and cooperation among dictionaries, thus this paper focuses on a modified similarity calculation method based on multi-dictionary through the combination of multiple dictionaries to improve the reliability of the calculation method of semantic similarity and finally the implementation steps of the similarity calculation algorithm are given. Through the above analysis and verification, the feasibility and effectiveness of the proposed algorithm can be achieved and our future work will be as follows:

The mentioned algorithm above is not very comprehensive, and based on this consideration, we will focus on the idea of comprehensive semantic similarity in our future work, First of all we intend to calculate the semantic similarity between two concepts with the combination of the depth and density of the ontology concept tree; then calculate the attributes similarity between two concepts; and finally the two similarities are weighted by impact factor.

ACKNOWLEDGMENT

The authors are grateful to the support of Ministry of Education Humanities and Social Sciences Planning Project no. 13YJAZH130, Natural Science Foundation of Jilin Province no. 20130101074JC and 20130101072JC, Doctoral Scientific Fund Project of the Ministry of Education of China no. 20120061110044, Jilin province science and technology development plan item under Grant no. 20130101072JC, Jilin province science and technology development plan item no. 20120303.

REFERENCES

[1] D. LIN. An information-theoretic definition of similarity [M]. // in proceedings of the 15th international Conference on Machine Learning, Madison, Wisconsin, 1998.

[2] Yu, Y.X. , Wang, L.Y., Research of information retrieval based on web semantic similarity, Applied Mechanics and Materials, v 135-136, p 753-758, 2012, Advances in Science and Engineering II

[3] Zhang, Pei-Ying, Semantic similarity metric and its application in text classification, Applied Mechanics and Materials, v 170-173, p 3711-3714, 2012, Progress in Civil Engineering.

[4] Ren, Wuling, Guo, Jinju, Word similarity algorithm based on wordnet and hownet, Applied Mechanics and Materials, v 155-156, p 375-380, 2012

[5] Li, Wenjie , Zhao, Yan; Shen, Nan, Concept similarity calculation in ontology mapping, 6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009, v 2, p 214-218, 2009

[6] WONG A K Y, RAY P, WARAN N P. Ontology Mapping for the Interoperability Problem in Network Management [J]. IEEE Journal on Selected Areas in Communication, 2005, 23(10):2058-2026.

[7] BPOWN P. Word Sense Disambiguation Using Tactical Methods[C]//29th Annual Meeting of the Association for Computational Linguistics, 18- 21 June 1991 University of California, Berkeley, California USA, Proceedings ACL 1991.

[8] DOAN A H, MADHAVAN J, DOMNGOS P. Learning to Map between Ontologies on the Semantic Web [C] // Proceedings of WWW, New York, USA: ACM Press, 2002, 662-673.

[9] Lu, Gang, Huang, Peng; He, Lijun; Cu, Changyong; Li, Xiaobo, A new semantic similarity measuring method based on web search engines, WSEAS Transactions on Computers, v 9, n 1, p 1-10, January 2010.

[10] Losif, Elias, Potamianos, Alexandros, Unsupervised semantic similarity computation using web search engines, PROCEEDINGS OF THE IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE: WI 2007, pp. 381-387, 2007.

[11] Dong, ZD; Dong, Q, HowNet - A hybrid language and knowledge resource, 2003 INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING, PROCEEDINGS, pp. 820-824, 2003

[12] Mititelu, Verginica Barbu, Wordnets: State of the art and perspectives. Case study: The Romanian wordnet, International Conference Recent Advances in Natural Language Processing, RANLP, pp. 672-677, 2011.

[13] He, Xiayan, Liu, Lei; Wu, Jinqiao,Semantic similarity calculation based on sememe set, Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010, v 1, p 423-428, 2010.

[14] Hu, Feng Song, Guo, Yong, An improved algorithm of word similarity computation based on HowNet, CSAE 2012 - Proceedings, 2012 IEEE International Conference on Computer Science and Automation Engineering, v 3, p 372-376, 2012.

[15] Liu, Lei , Zhong, Maosheng; Liu, Hui; Lu, Ruzhan, Word similarity measurement based on basic lexical set, Journal of Information and Computational Science, v 8, n 5, p 799-807, May 2011.

[16] H. A. M. N. YASSER GANJISAFFAR. A Similarity Measure for OWL-S Annotated Web Services [M]. // WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society Washington, DC, USA ©2006.:,:621-624

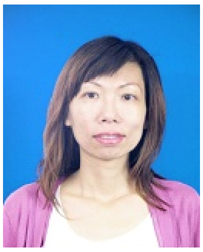
[17] Yan Wei, Xun Endong, English words Similarity calculation based on WordNet [C]. The Second session of the National Student Linguistics Symposium: 281-282-283-284-285-286-287-288.

[18] Liu Qun, Li Sujian. Word semantic similarity calculation based on HowNet [M]. Taipei: The Third Chinese Lexical Semantics Workshop, 2002.

[19] AGIRREE, RIGAUG. A Proposal for Word Sense Disambiguation Using Conceptual Distance [ED/OL]. (1995-12-11) [2009-04-20]. http://eprints kfupm. edu.sa/20289/1/20289.pdf

[20] Lei, Jingsheng , Measuring semantic similarity between words using multiple information sources, Journal of Information and Computational Science, v 7, n 2, p 601-608, February 2010

[21] Liu, Lei , Zhong, Maosheng; Lu, Ruzhan; Liu, Hui, Computing semantic similarity using Machine-Readable Dictionary and HowNet sememe, ICIC Express Letters, v 5, n 4 B, p 1391-1396, April 2011.



Chijun Zhang, born in 1972, Vice Professor and Ph.D. Member of Key Laboratory of Logistics Industry Economy and Intelligent Logistics at Universities of Jilin Province. His current interests include things of internet, wireless network and semantic theory.

Yong-jian Yang, born in 1960, Professor and Ph.D supervisor. His current interests include network communication and grid computing.

Xiao-yu Guo, M.S. candidate. His current research interests include multi-agent theory.

Zhan-wei Du, born in 1988, Ph.D. candidate. His current research interests include complex mobile communication network, delay tolerant network and social network.

Na Lin, member of Jilin Architecture and Civil Engineering Institute of computer. His current research interests include wireless network.