Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(8): 225-231

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A computation study on semantics based weighted sentence similarity

[1,2]Dongmei Li, [1]Jiajia Hou, [1]Shudong Hao, [1]Na Li and [3]Bo Zhang

[1]School of Information Science and Technology, Beijing Forestry University, Beijing, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[3] Business Analytics Department, MAXIMUS Inc Austion TX, USA

ABSTRACT

Semantic similarity is the essential part of automatic question answering system, and the computation of semantic similarity based on ontology can provide more various semantic information compared with the traditional method. In this paper, we firstly propose a semantic similarity algorithm used in the sentence similarity computation, combining relations of semantics, hierarchical structure and depth. Next, based on conventional sentence similarity algorithm, we develop a weighted sentence similarity algorithm. Finally, a question answering system is implemented with the two proposed algorithms. The experimental results show that our proposed method can achieve higher accuracy than others.

Key words: ontology, semantic similarity, sentence similarity, orchard pests and diseases.

INTRODUCTION

Sentence similarity plays an important role in many areas, such as text mining, information retrieval and question answering system [1]. The research on the sentence similarity algorithm includes two main types: the computation between the questions and answers [2], and the computation between the questions and questions [3]. In this paper, we research the latter.

There are two algorithms in the computation of sentence similarity: one is based on VSM (Vector Space Model) [4,5], the other is based on semantic similarity [6,7]. The algorithm based on VSM fails to consider the similarity of the structure as a whole without any analysis of structure. The semantics based algorithm processes the sentence with complete syntax and semantic analysis, which is a kind of deep structure analysis method.

For semantic similarity computation, there is often a difference between the result of the conventional method and the real meaning of the semantic. The algorithm based on ontology, however, can eliminate that difference [8]. At present, the computation of semantic similarity based on ontology includes the semantic tree method [9-11] and the directed graph method [12,13]. Wang Jin [10] introduces a method that combines semantic relations, hierarchical structures and inheritable relations, but he do not consider the influence from the depth on the concept. Li R. [11] propose a comprehensive approach. However, the number of the sub-concept is not considered in the computation.

In this paper, we firstly propose a new algorithm of semantic similarity, and apply it into the sentence similarity computation. In this algorithm, we consider the depth and the influence of the sub-concepts, and model the domain of orchard pests and diseases based on ontology and the related current knowledge about orchard pests and diseases. Besides, considering the relation between the depth and the information, we secondly propose a weighted sentence similarity algorithm based on VSM. Finally, we construct an automatic question answering system with two algorithms. Users provide questions in natural language to this system; it can match several similar questions by computing the sentence similarity, and rank all the sentence in descending order according to similarity; then users will receive the related answer. The experimental results show that this new method, comparing to others, reaches a higher accuracy of question matching.

EXPERIMENTAL SECTION

COMPUTATION OF SEMANTIC SIMILARITY

We use tree structure to describe the logical relationships between the concepts in ontology. This kind of conceptual semantic tree provides the basis of semantic in search algorithms [14,15]. In order to increase the accuracy in search, considering semantic relations, hierachical structure and so forth, Wang Jin [10] processes the relations of different types of nodes such as grandchildren nodes, cousin nodes in the conceptual tree. We improve that method by describing and quantizing the similarities of concepts.

The realted definitions will be listed as follows:

Definition 1: Closet root concept

In the conceptual tree of ontology, if concept R is the common ancestor of A and B, and it is thefarthest node of thosewho satisfy the condition, then R is the closet root concept of A and B, denoted as R(A, B) or R.

Definition 2: Same-branch concepts

If concept A is concept B's ancestor in the conceptual tree of ontology, then A and B are the same-branch concepts, denoted as S(A, B). According to the Definition 1, R(A, B) = A, and the distance between A and B d(A, B) = dep(B) – dep(A), where dep(A) is the depth of A in the hierachical strucutre.

Definition 3: Different-branch concepts

If A is not the ancestor of B and B is not the ancestor of A, then A and B are different-branch concepts, denoted as D(A, B). Then the distance between A and B is d(A, B) = d(A, R) + d(B, R).

Definition 4: Semantic-related concept

Concept C is the semantic-related concept of A and B, if and only if C satisfies the conditions as follows: if A and B are same-branch concepts, C is in the subtree with root A and is not in the subtree with root B; if A and B are different-branch concepts, C is in the subtree with root R but not in the subtree with root A or B. Such concept is denoted as SR(C).

Definition 5: Sub-concept number

The sub-concept number of C is the concept number of subtree with root C, denoted as son(C), where son(R) = SR(R) + son(B) if S(A, B) and R(A, B) = A; or son(R) = SR(R) + son(A) + son(B) if D(A, B) and R(A, B) = R.According to the five definitions above, the conceptual similarity between concept A and concept B can be expressed in formula (1):

$$sim(A,B) = \begin{cases} \left(1 - \frac{\beta}{dep(R(A,B)+1)}\right) \times \frac{\gamma}{d(A,B)} \times \frac{son(B)}{son(A)}, & when\ d(A,B) \neq 0, S(A,B) \\ \left(1 - \frac{\beta}{dep(R(A,B)+1)}\right) \times \frac{\gamma}{d(A,B)} \times \frac{son(A)+son(B)}{son(R)}, & when\ d(A,B) \neq 0, D(A,B) \\ 1, & when\ d(A,B) = 0 \end{cases} \tag{1}$$

where dep(R(A, B)) is the depth of the closet root concept of A and B; d(A, B) is the non-negative distance between A and B; son(A) is the number of nodes with root A in the conceptual tree of ontology; parameter β and γ are the weights, adjusting dep(R(A, B)) and d(A, B). Note that β and γ are in the interval of (0, 1] and sim(A, B) is in the interval of [0, 1]. We combine the depth of closet root concepts, distance and the number of sub-concept to calculate the similarity.

In the conceptual semantic tree, the deeper a concept lies, the more semantic information. Therefore, depth has impact on the calculation of similarity. The formula (1) nonetheless is limited because it does not reflect such impact directly.

Consider two situations below in the Figure1.
(1) For same-branch concept S(A, B1), R(A, B1) = A;
(2) For different-branch concept D(B1, B2), R(B1, B2) = A;

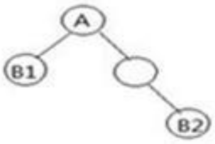

Figure 1: Two situations.

In such two cases, their same values $1-\frac{\beta}{dep(R(A,B)+1)}$ lead to the meaningless calculation. In order to make the difference between these two cases clearer, we add depth factor in the formula (1) as a kind of relative relation of depth between the two concepts. We perform the depth factor on the sub-concept number, combine the depth of closet root concept, and enlarge the impact of depth on the computation of similarity. Thus, the accuracy will be increased.

Definition 6. Depth factor
Depth factor is the relative ratio between the depths of A and B, denoted as α(A, B), as formula (2) shows:

$$\alpha(A\ ,\ B)=\begin{cases}\frac{dep(A)}{dep(A)+dep(B)}, & dep(A)\le dep(B)\\ 1-\frac{dep(A)}{dep(A)+dep(B)}, & dep(A)>dep(B)\end{cases} \tag{2}$$

Here, the semantic similarity of same concepts is 1. Besides, integrating the semantic relations of concepts, hierachical structure and inheritable relations, we develop the formula (1). Thus, the semantic similarity between concept A and B will be defined as the formula (3):

$$sim(A,B)=\begin{cases}\frac{\beta}{d(A,B)}+\gamma\frac{dep(R)+1}{dep(R)+\alpha(A,B)\cdot\left(son(A)-son(B)\right)+(1-\alpha(A,B))\cdot son(B)}, & when\ d(A,B)\neq 0, S(A,B)\\ \frac{\beta}{d(A,B)}+\gamma\frac{dep(R)+1}{dep(R)+\alpha(A,B)\cdot son(A)+(1-\alpha(A,B))\cdot son(B)}, & when\ d(A,B)\neq 0, D(A,B)\\ 1, & when\ d(A,B)=0\end{cases} \tag{3}$$

where β and γ are adjustment factors that adjust the weights of d(A, B), dep(R) and α(A, B), and they are in the interval of (0, 1].

COMPUTATION OF SENTENCE SIMILARITY

Traditionally, Vector Space Model(VSM) was introduced by G.Salton [4]. The underlying principle is: calculate the similarity based on constructing document vectors and query vectors, using the matching function. Sentence similarity uses the question ontological vector and the candidate ontological vector to improve the accuracy of recall and precision. Inspired by the conventional VSM, according to the characteristics of ontology, we switch the question sentence and the candidate sentenceto the semantic vector consisting of ontological concepts and calculate the sentence similarity. In this paper, we consider the relations between information and depth, optimize the search algorithm of sentence similarity based on this method.

We denote a user's question as X, and the sentence in the answer corpus as Y. Thus, X can be regarded as the word sequence X_1, X_2, …, X_n, and Y as Y_1, Y_2, …, Y_m.
Question vector X = (X_1, X_2, …, X_n);
Candidate vector Y = (Y_1, Y_2, …, Y_m).

Conventional sentence similarity: The semantic similarity of all words in sentenceA and B can be calculated in matrix M_{xy}.
1) Construction of M_{xy}:

$$M_{xy}=\begin{bmatrix}sim(x_{1,}y_1) & \cdots & sim(x_{1,}y_m)\\ \vdots & \ddots & \vdots\\ sim(x_{n,}y_1) & \cdots & sim(x_{n,}y_m)\end{bmatrix}$$
(4)

In formula (4), sim(x_i, y_j) is the conceptual similarity of X_i and Y_j. Each row in the matrix represents the conceptual similarities between a word in X and all words in Y.

2) Dimension reduction

In this part, we can calculate the semantic similarities between all the words in X and sentenceY. For each row in the matrix, we use the max(sim(x_i, y_j)) to calculate the maximum of conceptual similarity between a certain word in X and all the words in Y. Thus, the matrix will be reduced to the one-dimension. For these maximums in each row, their average is the sentence similarity between X and Y, as formula (5) shows.

$$Sim1 = \frac{1}{n} * \sum_{i-1}^{n} \left(max\left(sim(x_{i,}y_j) \right), j \in [i,n] \right) \tag{5}$$

Weighted sentence similarity: In conceptual semantic tree, different depths of concepts will cause various meanings; that is to say, the moredepth in the ontology, the more concrete information the concept represents. Let p(A, i) be the path length from root node to leaf node i that visits node A. By formula (6), we introduce the weight of concept A as:

$$\begin{cases} W(A) = dep(A)/mindep(A) \\ mindep(A) = \ dep\ (j),\ j = \underset{i\ \in\ the\ leaf\ \ node\ set}{\arg\min}\ p\left(A,\ i\right) \\ \end{cases}$$

(6)

where dep(A) is the depth of concept A in the hierarchical structure, and mindep(A) represents the depth of the leaf node which,among all leaf nodes in the hierarchical structure, has the shortest path to the root via concept A.

Then, the weighted ratio of concept A and B can be expressed in formula (7)

$$W(A\ ,\ B) = \begin{cases} \dfrac{dep(A)/mindep(A)}{\frac{dep(A)}{mindep(A)} + \frac{dep(B)}{mindep(B)}}, & dep(A) \le dep(B) \\ 1 - \dfrac{dep(A)/mindep(A)}{\frac{dep(A)}{mindep(A)} + \frac{dep(B)}{mindep(B)}}, & dep(A) > dep(B) \end{cases} \tag{7}$$

1) Construction of weighted matrix M_{xy}.

$$M_{xy} = \begin{bmatrix} W_{11} sim(x_{1,}y_1) & \cdots & W_{1m} sim\left(x_{1,}y_m\right) \\ \vdots & \ddots & \vdots \\ W_{n1} sim(x_{n,}y_1) & \cdots & W_{nm} sim\left(x_{n,}y_m\right) \end{bmatrix} \tag{8}$$

2) Dimension reduction
For the maximum in each row max(W_{ij} sim(x_i, y_j)), calculate the average of these maximums:

$$Sim2 = \frac{1}{n} * \sum_{i-1}^{n} \left(max\left(W_{ij} sim(x_{i,}y_j) \right),\ j \in [i,n] \right) \tag{9}$$

Integrated sentence similarity: Combining the conventional and weighted sentence similarity algorithm, we propose an integrated sentence similarity algorithm:

$$\begin{cases} sim(X,Y) = \ \alpha Sim1 + \beta Sim2 \\ \alpha + \beta\ = 1 \end{cases} \tag{10}$$

where α and β are adjustment factors.

The algorithm is described as follows:
1) Segmentation. For the hardships of terminology of orchard pests and diseases and the construction of sentence pattern corpus, we develop the most positive director matched. We use the related concepts in ontology as a segmentation dictionary and the dictionary-based most positive director segmented algorithm to segment the question X into a sentence vector X = (x_1, x_2, …, x_n). According to the result, we judge the sentence pattern S of the question. After that, a candidate whose sentence pattern is also S will be chosen as candidate sentence, and will be segmented into the candidate sentence vector Y = (y_1, y_2, …, y_m).

2) Obtaining the similarity matrix. According to the formula (3), we calculate the semantic similarities of sentence vectors, and get the similarity matrix in the formula (4).

3) Obtaining the weighted similarity matrix. According to the formula (3) and (7), we calculate the semantic similarities of sentence vectors, and get the weighted similarity matrix in the formula (8).

4) Calculating the conventional sentence similarity Sim1 between the question and the candidate sentence, according to the formula (5).

5) Calculating the weighted sentence similarity Sim2 between question and the candidate sentence according to the formula (9).

6) Calculating the integrated sentence similarity sim between the question and the candidate sentence, according to the formula (10).

RESULTS

We design a model on semantic similarity and sentence similarity, and implement a automatic question answering system based on frequently-asked-question corpus. This system can accept questions in natrual language and return a related answer to the users.

Semantic similarity: In order to demonstrate the effectiveness, we firstly obtain a certain pairs of concepts in the domain ontology. The experimental steps are:

1) Calculate the semantic similarities based on the proposed method;

2) Calculate the semantic similarities based on [10], where only upper and lower relations and the sub-concept number will be considered, without the depth of concepts. The results are in the column “Traditional similarity”.

3) Ask 10 people to judge the similarities based on their experience. Thus we will get the subjective judgment data of these concepts, and the average will be listed in the column “Subjective judgment”.

In the computation of semantic similarity, the search algorithm should set the specific parameters based on the current situation. In the experiment, the depth of ontological conceptual tree is 6; the dep(R(A, B)) in the formula (1) and (3) is an integer in the interval of [1, 6]; the d(A, B) is an integer in the interval of [0, 10], and their weights β and γ are 0.5 respectively, related to the domain ontology.

We use three methods to measure the in ten groups, shown in the Table 1. By comparing between the two methods and the subjective judgments, we collect the results and plot the Figure 2.

Table 1. The results of semantic similarities

No	Concept comparison	Semantic similarity	Traditional similarity	Subjective judgment
1	apple diseases - apple ring rot	0.775	0.328	0.800
2	apple phytophthora rot - apple scab	0.694	0.082	0.710
3	apple diseases - pear diseases	0.400	0.417	0.450
4	insect pests - leafhopper	0.500	0.104	0.520
5	parasa consocia - anoplophora	0.398	0.078	0.430
6	forewing - forewing spot shape	0.682	0.146	0.720
7	insect color - hind wing color	0.443	0.122	0.490
8	feeler - feeler shape	0.859	0.656	0.870
9	symptom onset - blade spot size	0.505	0.104	0.510
10	fruit spot - branch spot	0.425	0.140	0.440

From the Figure 2, it is obvious that the difference between the results from our method and the standard is less than that between the conventional method and the standard.

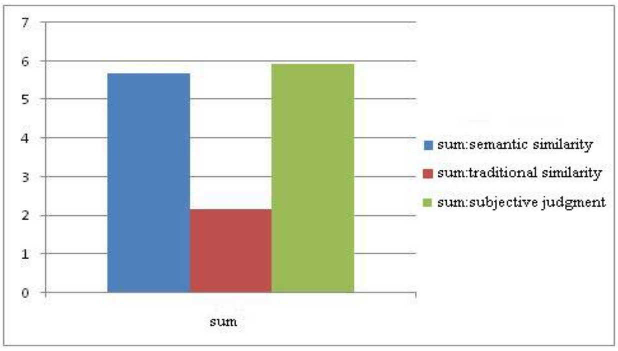

Figure 2: The results of semantic similarity

Sentence similarity: We set the parameters in different situations. In order to choose the proper α and β in the formula (10), we choose 500 candidate sentences from the corpus, including 10 types sentences such as reason pattern, symptom pattern and comparison pattern. By calculating the sentence similarities between these candidate sentences and questions, we plot the Figure 3 to show the relation between the accuracies and the α. From the Figure 3, the accuracy reaches the optimum when α= 0.65. Therefore, we set α = 0.65 and β = 0.35.

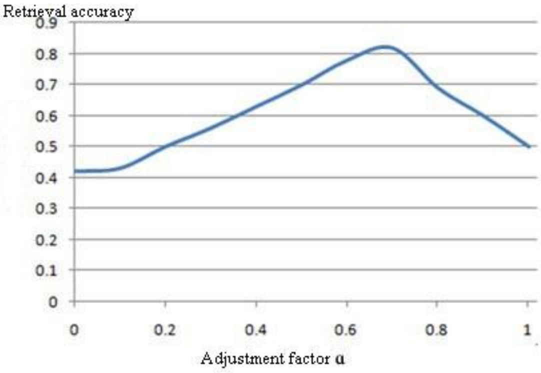

Figure 3: The relations between α and accuracy.

Firstly, we choose some candidates from the corpus, and get some questions from the users.
1) Calculate the similarity between the question and the candidate sentence, according to the conventional sentence similarity algorithm.

2) Calculate the similarity between the question and the candidate sentence, according to the proposed weighted sentence similarity algorithm.

3) Calculate the integrated sentence similarity.

4) Ask 10 people to judge the sentence similarities as the subjective judgments.

Then we use four methods to measure the similarities, and get five groups, shown in Table 2 by comparing the results of the three methods with subjective judgments, we plot the Figure 4.

Table 2. The results of sentence similarities

No.	User questions	Candidate sentence	sim1	sim2	sim	Subjective judgment
1	Why apple's blades will curl	Reasons leading to apple blades' curling	0.992	0.479	0.812	0.950
2	Which parts of the pear tree will be harmed by pear leaf spot moth	Which parts of the pear tree will be harmed when suffering pear leaf spot moth	0.970	0.473	0.796	0.920
3	What is the morphological characteristics of hawthorn spider mite	What is the morphological characteristics of peach buds	0.833	0.417	0.687	0.650
4	Will peach trees' branches be dead when suffering peach leaf disease	Will peach trees' branches be dead when suffering powdery mildew	0.873	0.452	0.726	0.700
5	How many kinds of diseases grape trees have	Species of diseases of grape trees	0.933	0.482	0.775	0.900

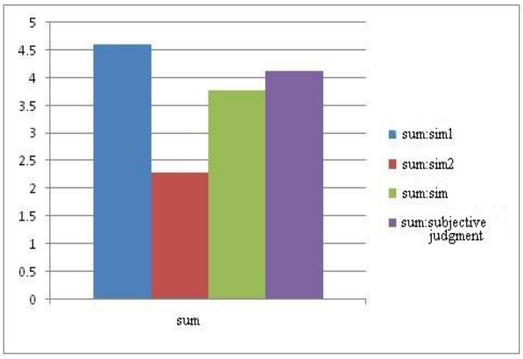

Figure 4: The results of sentence similarity

From Table 2 and Figure 4, we see that traditional method is appropriate when the question has closer or is the same to the candidate sentence, whereas the weighted algorithm is appropriate when the question is not so close to the candidate sentence. Therefore, the integrated similarity which combines these two methods has higher accuracy. The difference between the integrated similarity and the standard is less than the difference between the conventional and weighted method and the standard.

The system calculates the sentence similarities between the questions and the sentences in the corpus, matches the sentences with the closet meaning, and sorts the results by descendent similarities. The users will get the answers according to the similarities immediately, since the answers to the questions in the corpus have been saved. But the users need to use other methods,such as information retrieval, answer extraction, to get the answer if no corresponding questions in the corpus. If the answer is reasonable, users can add the question and the corresponding answer to the corpus.After receiving users' problems in a natural language description,the automated answering system returns the answers to the users in accordance with the above method.

CONCLUSION

In this paper, we propose a new semantic similarity algorithm, integrating the relations of conceptual semantics, hierachical structure and the depth, and apply the algorithm in the computation of sentence similarity. Besides, considering the depths and the meanings, we propose a weighted sentence similarity algorithm based on VSM. By combining thse two algorithm, we introduce an integrated sentence similarity algorithm. Finally, a system based on frequently-asked-question corpus of orchard pests and diseases demonstrates the effectiveness of our method. This paper does not take the property factors into consideration. Therefore, our future work focuses on the property factors in the ontology for those who have lots of property nodes.

Acknowledgment
This work was supported in part by the National College Students' Training Programs of Innovation and Entrepreneurship(No.201210022055), and the Fundamental Research Funds for the Central Universities(No.YX2011-30).

REFERENCES

[1] Zhang P. Y.. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, **2013**, 11, 1909-1915.
[2] Cui H.; Cai D. F.; Miao X. L.. *Journal of Chinese Information Processing*, **2005**, 18, 24-31.
[3] Li X. L.; Liu J. M.; Shi Z. Z.. *Journal of Computer Research and Development*, **2000**, 37, 1032-1038.
[4] Salton G.. Englewood cliffs, New Jersey: Prentice Hall Inc, **1971**.
[5] Turney P. D.; Pantel P.. Journal of Artificial Intelligence Research, **2010**, 37, 141-188.
[6] Islam A.; Inkpen D.. *ACM Transactions on Knowledge Discovery from Data* (TKDD), **2008**, 2, 1-25.
[7] Oliva J.; Serrano J. I.; Castillo M. D.; Iglesias A.. Data & Knowledge Engineering, **2011**, 70, 390-405.
[8] Naji Hasan A. H.; Gao S.; Malek A. G.; Jiang Z. L.. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, **2013**, 11, 4505-4511.
[9] Li Y.; Bandar Z. A.; McLean D.. *IEEE Transactions on Knowledge and Data Engineering*, **2003**, 55, 871-882.
[10] Wang J.; Chen E. H.; Shi D. M.; Zhang Z. Y.. PR & AI, **2006**, 19, 696-701.
[11] Li R.; Yang D.; Liu L.. *Journal of Computer Research and Development*, **2011**, Suppl: 312-317.
[12] Strube M.; Ponzetto S. P.. In: Proc.of AAAI. Boston, Massachusetts, **2006**, 1-6.
[13] Gabrilovich E.; Markovitch S.. n: IJCAI. Hyderabad, India, **2007**, 1606-1611.
[14] Sánchez D.; Batet M.; Isern D.; Valls A.. Expert Systems with Applications, **2012**, 39, 7718-7728.
[15] Vafaee F.; Rosu D.; Broackes-Carter F.. *BMC Systems Biology*, **2013**, 7, 1-17.