Similarity Calculation Method of Chinese Short Text Based on Semantic Feature Space

Liqiang Pan, Pu Zhang, Anping Xiong
College of computer science and technology
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—In order to improve the accuracy of short text similarity calculation, this paper presents the idea that use the history of short text messages to construct semantic feature space, then use the vector in semantic feature space to represent short text and do semantic extension, and finally calculate the short text similarity of corresponding vector in the semantic feature space. This method can represent the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. The results show that the method we proposed is reasonable and effective.**

Keywords—short text; semantic feature space; similarity; semantic similarity

I. INTRODUCTION

With the wide application of short text similarity calculation method in information retrieval, question-answering system, text mining and other natural language processing fields, the research and improvement on the calculation method of short text similarity has become an important research hotspot. The research finds that there are many differences between the calculation methods of short text similarity and document similarity. As the document contains large amount of word information, most of the similarity calculation method is based on word statistical method. However, the short text contains little word information, maybe even only one word. It is not sufficient to judge the similarity between the short texts accurately only using the information of the short text itself. Therefore, in order to improve the calculation accuracy of short text similarity, we need to solve two key problems. The first problem is how to fully expressed and reflected short text information? The information includes word frequency, word meaning, etc. The second problem is how to calculate the similarity between the short texts? In order to solve these two problems, this paper presents the calculation method of Chinese short text semantic similarity based on the semantic feature space. This method represent the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. The results show that the method we proposed is reasonable and effective.

II. CONSTRUCTION METHOD OF SEMANTIC FEATURE SPACE

We take the intelligent-service system as the research background. The main short texts in the system are advisory information (namely interrogative sentences) and response short texts. In the intelligent service system, there are many users asking for advices every day, which inevitably produces massive consultation information. We can use these historical advisory information, namely short text sets to construct the semantic feature space, and then build the model by using the new consultation of the users or questioning short text in the space, finally we can calculate the similarity between the new short text and historical short text. The semantic feature space has a similar construction process with the ordinary vector space, which also consists of two main steps: feature selection and feature dimension reduction.

A. The feature selection of the semantic feature space

As the short text contains few words and may even contains only one single word, this paper only uses the feature of first level instead of phrase level because the feature of phrase level is not conducive to fully represent short text.

The initial feature set of semantic feature space FS' is constructed like this: first, segment all the historical short text data set and remove stop words (stop words have no effect on the semantic expression of the sentence); then, remove function words and remain content words according to function word table. This is because the semantic meaning of short text is mainly conveyed by content words, while function words are mostly auxiliary words of mood and not carrying much semantic information. At last, the initial feature set FS' is obtained by aggregating all the content words of short text A_i like this:

$$FS' = A_1 \cup A_2 \cup A_3 \cdots A_n \quad (1)$$

B. The feature dimension reduction based on semantic clustering

Because of the complexity and diversity of Chinese word structure, the space dimensions of the initial feature set FS' are particularly high. The direct use of the initial feature set will inevitably increase the complexity of similarity calculation. Through the experimental analysis, it is found that there are many lexical items with the same semantic meaning in the initial feature set FS'. Therefore, we use the word similarity calculation method based on "hownet"[1] to cluster the feature lexical items with higher similarity in the initial feature set. The basic idea of this clustering method is: first aggregate the feature item with higher similarity as a cluster, then choose one feature lexical item optionally as the representative, finally constitute a set of

all the representative feature lexical items. This set is the feature set of the final semantic feature space FS. Semantic clustering greatly reduces the dimension of the semantic feature space and redundant features of the initial feature set FS', so as to improve the efficiency of calculation of text similarity. The semantic clustering method is a bottom-up clustering algorithm. The pseudo code of the algorithm is as follows:

1. Initialize each feature as cluster, the whole cluster set $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$, max=0;
2. For i=1 to n do
3. For j=i+1 to n do
4. Calculate the similarity between c_i and c_j, denoted as csim(i,j);
5. If csim(i,j) > max
6. Max = csim(i,j);
7. k_1 =i; k_2 =j;
8. End if
9. End do
10. End do
11. If max > λ
12. Merge c_{k_1} and c_{k_2};
13. Update index of each cluster;
14. n = n-1;
15. Go to (2);
16. Else stop;
17. End if

The value of λ is range from 0 and 1. The calulate algorithm csim(i,j) is shown as follows:

1. Initialize csim(i,j)=0;
2. For each feature fw_k in cluster c_i do
3. For each feature fw_l in cluster c_j do
4. csim(i,j) = csim(i,j) +sim(fw_k , fw_l);
5. End do
6. End do
7. csim(i,j) = csim(i,j) / ($|c_i|\times|c_j|$);

The csim(fw_k, fw_l) is a sematic similarity method based on "hownet". This thesis carries on the detailed introduction to the semantic similarity calculation method[1]. The $|c_i|$ and $|c_j|$ is the feature number of C_i and C_j.

The initial feature set becomes the feature set FS after semantic clustering, which is used to construct the semantic feature space. Each feature lexical term in the semantic feature space expresses specific semantic meaning and subject. The construction of the semantic feature space needs a lot of corpus training and aggregation calculation, but as long as the first training corpus is enough, the semantic feature space can be used directly later.

III. THE SIMILARITY CALCULATION METHOD OF CHINESE SHORT TEXT BASED ON SEMANTIC FEATURE SPACE

For an arbitrary short text C, after recognition of center word and word frequency statistics, we can map it to the semantic feature space mentioned in the previous paper and build the model of the text, then calculate the similarity between short texts in the semantic space. The specific methods are as follows:

First, in order to obtain the part of speech tagging word set T, for the short text C do segmentation using automatic segmentation system, pos tagging and remove stop word ased on the stoplist. Then statistic word frequency of tagging word set T,we can use the word frequency initialization vector V_c express short text:

$$V_C=(tf_1,tf_2\cdots tf_i\cdots tf_m) \tag{2}$$

In(2), m represents the number of words have distinct speech tagging of word set. tf_i is word i 's Frequency intagging word set T. Because of the short text word have less information, common words's and the central word's word frequency values are often in the same or is 1. In order to express the importance of the center word which reflects the importance meaning of the short text, Center word's word frequency is need to heavier its weights. First, using Tian Weidong's[2] center word recognition method to recognize center word set Z. Then, for each center word's word frequency multiply a weighting factor η. So obtain a new vector representation V_c of short text:

$$V_q=(tf_1w_1,tf_1w_2\cdots tf_iw_i\cdots tf_mw_m) \tag{3}$$

In(3), w_i represents the weight of the word, its value is η (indicating central word) or 1 (indicating non-central word).

Then, all the features word of semantic feature spacethe FS and all the words of intagging word set T construct a similarity matrix which is also known as text mapping matrix.

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{4}$$

The n represents the word number of semantic feature vector space FS. a_{ij} denotes the semantic similarity between the i word in intagging word set T and the j word in semantic feature spacethe FS. pan's method[1] to calculate the similarity of the two words.

After the text mapping matrix is constructed, short text semantic mapping vector V_C' can be obtained in semantic space. The method is short text word frequency vector V_C is multiplied by the mapping matrix A.

$$V_q'=V_q\times A$$

$$=(tf_1w_1,tf_2w_2,\cdots,tf_iw_i,\cdots,tf_mw_w)\times\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$= (\sum_{i=1}^{m} tf_i w_i a_{i1}, \sum_{i=1}^{m} tf_i w_i a_{i2}, \cdots, \sum_{i=1}^{m} tf_i w_i a_{im}) \quad (5)$$

Using the mapping matrix A, we put a m dimensional vector V_C converted into a n dimensional vector $V_C^{'}$,The m less than n. By this method, we can use more features represent the semantic information of short text, so as to improve the accuracy of calculating the similarity between the short text.

Therefore, for any two short text C_1 and C_2, modeling the short text by the above methods. The two short text are mapped to the semantic feature space, so as to obtain vector V_{C_1} and V_{C_2}.

Finally, using vector cosine value represents two short text similarity $Sim(C_1, C_2)$:

$$Sim(C_1, C_2) = \cos(V_{C_1}^{'}, V_{C_2}^{'}) = \frac{V_{C_1}^{'} \cdot V_{C_2}^{'}}{|V_{C_1}^{'}| \cdot |V_{C_2}^{'}|} \quad (6)$$

IV. The Design and Analysis of the Experiment

In this section, we mainly set parameters and validate the similarity calculation method of Chinese short text based on semantic feature space described above through several experiments. Four steps will be introduced in this section. They are experimental data, experimental evaluation method, experiment setup and tool and experimental results and analysis.

A. Experimental data

The experimental data are purchased through data [3]. High quality question and answer corpus data set of Q & A community in 2013 (including 3000000 question and answer, database format, XML) are recorded as the original data set. As this paper only judges the similarity between short texts of problem, we only need short texts of problem. Through the analysis of the original data set, we write the preprocess programs, and extract 10 categories of problems according to the classification label of XML format (2000 short texts of problem for each category, a total of 20000) in order to form experimental data set D.

B. Experimental evaluation method

- Evaluate the similarity calculation method using $F-Value$

Take similarity calculation results between the texts as the similarity measure of K-Means clustering algorithm, then evaluate the effectiveness of similarity method. Metric data $F-Value$ is a balance index[4] of combination precision ration and recall ratio in information retrieval. The metric data $F-Value$ allows us to test whether the short text is correctly classified into the corresponding categories after clustering and the text of expected category is included in the same category.

Set the number of short texts of category i is n_i, the number of short texts of cluster j is n_j, n_{ij} represents the number of short texts which belongs to the cluster j and the category i, and then the precision ratio $p(i, j)$ of cluster j, the recall ratio $R(i, j)$ of category i can be respectively defined as:

$$P(i, j) = \frac{n_{ij}}{n_j}, R(i, j) = \frac{n_{ij}}{n_i} \quad (7)$$

The corresponding $F-Value$ $F(i, j)$ is defined as

$$F(i, j) = \frac{2 \times P(i, j) \times R(i, j)}{P(i, j) + R(i, j)} \quad (8)$$

Thus,we can get the global $F-Value$ F:

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)) \quad (9)$$

The n represents the number of short text in the entire data set. The same as ordinary clustering algorithm, the larger $F-Value$, the better the effect of clustering can be inferred from the similarity algorithm

- Evaluate the similarity calculation method using P@n

In the practical application of intelligent service system or automatic question-answering system, similarity calculation method mainly calculates by comparing every short text of historical problems and then responses by outputting the most similar answer of short text of historical problems. Therefore, we can use the P@n (Precision at n) as our experimental evaluation standard. P@n represents the probability or proportion of the occurrence of the correct result (historical short text and pending short text is similar indeed) in the top n results. For example: P@4=0.5 means that there are 2 short texts similar indeed to the pending short text in the first 4 similar short texts after the similar calculation of pending short text and historical short text which is in descending order according to the similarity of the historical short texts.

C. Experiment setup and tool

K-Means clustering algorithm uses open source tool lingPipe to realize.

D. Experimental results and analysis

In the construction of semantic feature space, semantic feature space dimension has a close relationship with parameter λ. Figure 1 shows the process that dimension varies with the parameter λ (including 3 experiments). In each experiment, all the content words in data set D construct the semantic feature space initially, then set the similarity threshold value λ, make dimensionality reduction of semantic feature space and compute corresponding spatial dimensions. Because of complexity and diversity of Chinese words structure, the original space dimension is particularly high which can achieve several thousand dimensions. After screening of content words, words with no semantic meaning can be removed and the space dimension can be reduced to about 6200. We use the similarity calculation method to cluster the features of semantic similarity in semantic feature space, so as to further reduce the space dimension. When λ is small, more semantic features cluster in one category, thus the number of categories will be less. As the number of space dimension and clustering is the same, the space dimension is lower. When λ increases

gradually, the situation is the opposite.

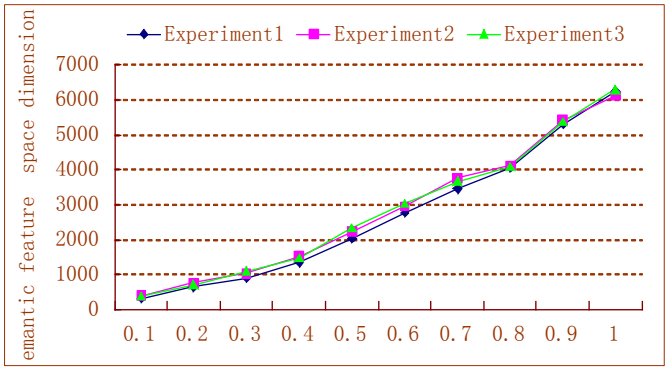

Fig. 1. semantic feature space dimension varies with the parameter λ

In order to get the best threshold value λ, we can observe the relationship between it and $F-Value$ in clustering algorithm, that is to say when $F-Value$ is the maximum (similarity effect is the best), value λ will be the best. Figure 2 is the variation diagram of threshold value λ and clustering algorithm, in which we can see that the optimal threshold value is between 0.4 and 0.6. Figure 2 only shows the results of three experiments. In every experiment, we calculate the similarity among data set D with different value λ using the similarity calculation method proposed in this paper, then take the results as the similarity measure of K-Means clustering algorithm, output corresponding $F-Value$ and form the graph. K-Means algorithm is implemented by using open source tool lingPipe. Value λ influences $F-Value$ of clustering algorithm by affecting space dimension. When the value λ is small, the space dimension is low, thus the semantic meaning of short text is not fully expressed, which leads to low $F-Value$ of clustering algorithm. When the value λ is larger, the space dimension is higher and a lot of invalid features and noises are introduced, thus the expression of the semantic meaning of the text is influenced, which results in lower $F-Value$ of clustering algorithm.

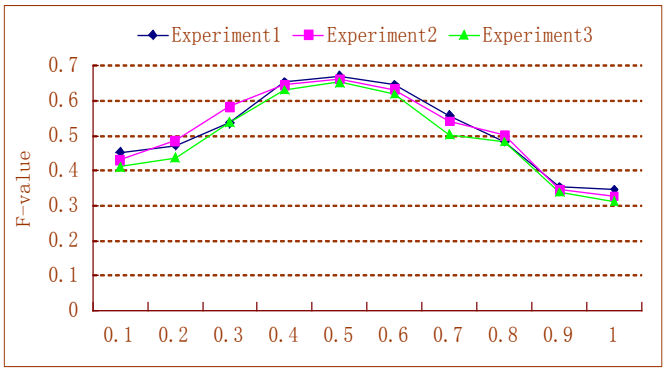

Fig. 2. $F-Value$ varies with the parameter λ

When building the model of the short text using semantic feature space, we add weight to the center word. The value of the weight will also affect the effect of similarity algorithm. Figure 3 is the relation graph of the center word weight η and $F-Value$. It shows the experimental results among many tests. In every experiment, the optimal λ is set to 0.52, the corresponding clustering algorithm $F-Value$ is output through changing the value of β the, thus the relation graph is formed. When $2 \le \eta \le 6$, $F-Value$ increases with the η, which shows that giving higher weights to the center word is helpful to improve the accuracy of the clustering, in other word, it is conducive to the similarity calculation. But with the increase of η, there is a downward trend of $F-Value$. The reason is the weight of the center word is so high that the function of other words is negligible and their semantic information is ignored. Therefore, the weight of the center word needs to be set to an appropriate value. Through repeated experimental analysis, the η should be set between 4~6.

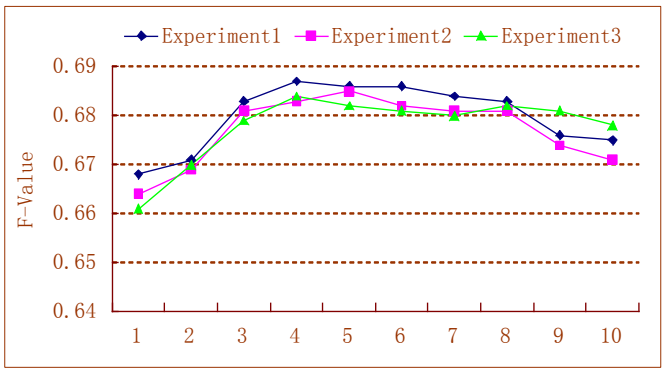

Fig. 3. $F-Value$ varies with the parameter η

We set the optimal parameters λ =0.52, η = 4 for the method in this paper. Then we conduct comparative tests with Huang Chenghui's text similarity measure method[5] which combines word semantic information and TF-IDF method and Song Wanpeng's question similarity calculation method[6] in question-answering system. Figure 4 shows that the similarity calculation method of Chinese short text based on semantic feature space can effectively improve the clustering effect, that is effectively judge the similarity between the short texts.

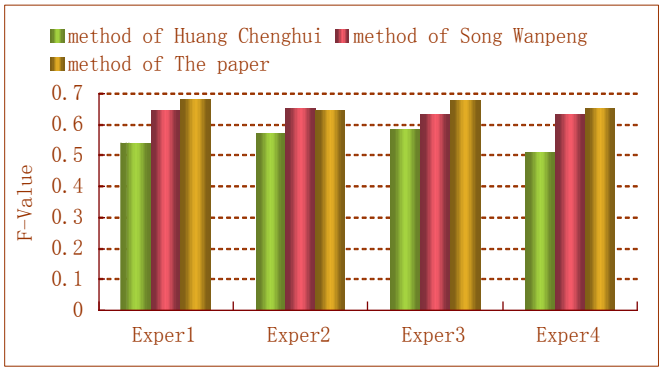

Fig. 4. Comparative experiment on Chinese short text similarity methods

To validate the effect of similarity calculation method in the actual application, we also use P@n to evaluate the similarity calculation method. P@n represents the probability or proportion of the occurrence of the correct result (historical short text and pending short text is similar indeed) in

the top n results. Table 1 presents the corresponding accuracy and also lists the accuracy of question similarity calculation method[11] in question-answering system of Song Wanpeng.

TABLE I. COMPARATIVE EXPERIMENT

method	P@1	P@2	P@3
method of	0.412	0.537	0.564
method of Song Wanpeng	0.423	0.573	0.605
method of the paper	0.483	0.581	0.627

V. CONCLUSION

The paper's method represents the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. This method is feasible and applicable.

REFERENCES

[1] Q.Liu and S.J.li, "Word's semantic similarity computation Based on the HowNet", The 3rd Chinese lexical and semantic proseminar, Taipei, China, 2002.

[2] Tian Weidong, Li Yajuan. center word recognition based on CRF and error driven .[J]. Application Research of computers.2013.

[3] http://www.datatang.com/data/44720

[4] Oliva J. Serrano J I. Del Castillo M D. et al. SyMSS: A syntax-based measure for short-text semantic similarity[J].Data&Knowledge Engineering.2011.70(4): 390-405.

[5] Huang Chenghui. Jian Yin. Hou Fang. A combination text similarity measure method of word semantic information and TF-IDF method [J].Computer science.2011.34 (5): 857-864.

[6] Song W.Liu W.Gu N.A Semantic Space for Question Similarity Calculation in User-Interactive Question Answering Systems. Journal of Computational Information Systems. 5(3): 1055-1063. June. 2009.

[7] B.Ge,F.F.Li,S.L.Guo, "Word's semantic similarity computation method based on HowNet", Application Reserarch of Computers, Vol.27, No.9, pp.3329-3333, Sep.2010.

[8] Hu, Feng Song, Guo, Yong, "An improved algorithm of word similarity computation based on HowNet", Computer Science and Automation Engineering, IEEE International Conference, Vol.3, May 2012.

[9] Z.Dong and Q.Dong,HowNet,http://www.keenage.com.

[10] He X.Liu L.Wu J. Semantic Similarity Calculation Based on Sememe Set. In:Proc of the 2010 International Conference on Artificial Intelligence and Computational Intelligence. Sanya. China: IEEE Computer Society.2010.423-428

[11] Feng song Hu. Yong Guo. An Improved Algorithm of Word Similarity Computation Based on HowNet. In: Proc of the 2th International Conference on Computer Science and Automation Engineering. Zhang jia jie. China. 2012

[12] Luo Jun. Ke Zhang and Xilin Chen . Text Similarity Computing Based on Sememe Vector Space. IEEE ICSESS 2013.