Chinese Text Similarity Algorithm Based on Part-of-Speech Tagging and Word Vector Model

Zhixin Ma, Mengguang Li*
College of Information Science and Engineering, Lanzhou University, Lanzhou, China.

* Corresponding author. Tel.: +86 15516073921; email: mango58lmg@gmail.com
Manuscript submitted February 29, 2019; accepted April 10, 2019.
doi: 10.17706/jcp.14.4.311-317

Abstract: Based on the unique structure of Chinese text data in the field of building materials, this paper proposes a Chinese text similarity calculation method based on part-of-speech tagging and word vector model. By analyzing the special structure of Chines text data in the building materials field, the method realizes specific part-of-speech tagging by using the machine, which saves a lot of manpower and time consumption required for manual labeling. Then, combined with the word vector model, the text similarity calculation is realized by these steps: Chinese word segmentation, syntactic analysis, and similar matching of the annotated text. This paper comprehensively compares the word level similarity calculation method based on the vector space model. Through analysis and experimental comparison, the algorithm in this paper obtains an average F value of 72.0%, which is improved by 20.67% compared with the method based on vector space model, and has achieved better test results.

Key words: Chinese text similarity, word vector model, part-of-speech tagging, vector space model.

1. Introduction

Nowadays, the amount of information is growing so fast that efficiently extract useful and relevant information from massive data has become the key to many applications, such as automatic question answering system, text mining, information extraction [1], etc. And the text similarity calculation technology is the key to achieve the above objectives. At present, the commonly used text similarity calculation methods can be roughly divided into the following types: a string-based method, a corpus-based method, a world knowledge-based method, and a semantic and syntactic analysis method [2]. The string-based method cannot synthesize sentence semantics and statement information, which will affect the computational efficiency. The corpus-based method requires collecting and collating sufficient corpus data to train the corresponding model. The world knowledge-based method needs to process large data sizes and generally require complex data pre-processing. The semantic and syntactic analysis method needs linguistic analysis of Chinese text, requiring computers to have higher processing performance.

This paper analyzes the characteristics of building materials data and uses combination of part-of-speech tagging [3] and word vector model [4] to calculate Chinese text similarity. According to the syntactic features and data characteristics of the text, using the machine to uniformly mark the data with specific part-of-speech tagging, saving a lot of manpower and time investment required for manual labeling. Then use the collected data corpus to train the word vector model and combine the tagged part-of-speech to calculate the text similarity. Through comparison experiments and results analysis, 72.0% accuracy was obtained using the method based on part-of-speech tagging and word vector model, while the method

based on vector space model obtained the accuracy of 51.33%. Experiments have shown that for a domain-specific Chinese text similarity calculation, a combination of corpus-based and syntactic-based analysis can indeed achieve better results.

The structure of this paper is as follows. The second chapter introduces some related work of the experiment. The third chapter introduces in detail the text similarity calculation method based on part-of-speech tagging and word vector model. The fourth chapter details the process and results of the experiment, then analyzes and discusses the results in depth. The fifth chapter is a summary of the full text and a further work introduction.

2. Related Work

2.1. Chinese Word Segmentation

Since it is difficult for a computer to directly process Chinese text, it is necessary to first divide the text into a certain amount of words according to certain norms and expression habits. The main word segmentation methods now mainly include dictionary-based methods and word-based methods [5].

In view of the relatively short-term characteristics of the building materials field, the data segmentation of this experiment is implemented by the improved two-way maximum matching algorithm [6]. This is a dictionary-based word segmentation method. On the basis of the original corpus, some special vocabulary in the field of building materials is added, and some special symbols and units are limited by regular expressions. From the results of the word segmentation, it has achieved better results than the ordinary word segmentation method.

2.2. Part-of-Speech Tagging

In most natural language applications, part-of-speech [7] is a key part, and the accuracy of it directly affects the results of subsequent data analysis. At present, the commonly used part-of-speech tagging methods are mainly divided into the following categories: rule-based methods, transform-based methods, and statistical-based methods [8]. The rule-based approach requires a lot of manpower and time to compile semantics and rules, and the scalability is also poor. The transformation-based approach first uses the annotated corpus to obtain initial annotations, trains to learn new rules, and then updates the annotation data with new rules. The disadvantage is that a related corpus is required and the rule training time is often longer. At present, the most commonly used methods are statistical-based annotation methods, including part-of-speech tagging using Hidden Markov models, maximum entropy models, decision tree models, and neural network models [9].

Since the building materials data processed by the special word segmentation in this experiment are short texts with fixed grammar rules, a rule-based labeling method is adopted. In addition, by analyzing the grammatical features of the data, it is found that the data needs to be processed is basically composed of modifiers and central words. Therefore, we combine the tools of regular expressions to realize automatic labeling of text data through the machine, which saves a lot of manpower and time consumption. From the results, the labeling results also achieve sufficient accuracy.

2.3. Word Bag Model

The basic idea of the word bag model is to recognize the distribution hypothesis and think that the semantics of words are similar to the context. Therefore, the word bag model represents a document as a combination of words, which mainly includes different word bag models such as vector space model [10], probabilistic latent semantic analysis, and potential Dirichlet distribution. In terms of semantic degree, the method based on vector space model is the weakest, and Dirichlet distribution is the strongest.

In order to achieve the comparative analysis of the experiment, a similarity calculation is performed on

the sane data using a word bag model based on vector space. After comparison, it is found that the similarity method combined with part-of-speech tagging and word vector model has obvious optimization and improvement compared with vector space model method.

3. Text Similarity Calculation Based on Part-of-Speech Tagging and Word Vector Model

This paper uses a corpus-based and syntactic-based approach to calculate Chinese text similarity. Firstly, according to the uniqueness of building material data, we use the part-of-speech tagging method based on syntactic analysis. Then combine the corpus-based word vector model to calculate the text similarity. Of course, the corpus here expends a large number of corresponding professional vocabularies.

Word2Vec [11] is a Google open source tool to make words into vectors and is a simplification of the neural network language model. Here we use Word2Vec and the optimized training model to express words into corresponding vector forms, and calculate the similarity of words by calculating the cosine of the angle of the vector through the constructed word vector space.

In the corpus after semantic annotation, the annotated sentences generally contain multiple parts of speech. Through observation, it is found that the sentence patterns of the texts in the building materials field are very similar, most of them are centered on a central word, and the rest are modified components for the central word. By using this feature, the general law of the sentence pattern can be determined, and then the machine can be automatically labeled with the regular expression tool. Finally, the labeled sentence is uniformly transformed into the sentence of the modified phrase plus the central word. The sentence similarity calculation method based on part-of-speech tagging is to convert the similarity between sentences into the similarity between two central words and two modified phrases.

When calculating the similarity of text, the central word is the core and the modified phrase is used as the reference. So for an annotated text T, use H to represent its central word, and $\mathrm{M(T)} = \{\mathrm{m}_1, \mathrm{m}_2 \ldots \mathrm{m}_n\}$ to represent its modified phrase, that is, all the labeled corpus can be expressed in a two-tuple form: $\big(\mathrm{H}, \mathrm{M(T)}\big)$.

Through the β value, we can adjust the weight of the central word similarity in the whole calculation. Relatively, we also hope that the similarity of the modified phrase can also restrict the central part. So in this algorithm, we define the similarity calculation formulas of the label texts T_1 and T_2 as follows:

$$Sim(T_1, T_2) = \mathrm{Sim}(H_1, H_2) \times [\beta + (1 - \beta) \times Sim(M(T_1), M(T_2))] \tag{1}$$

The H_1 and H_2 are the central words of the sentence T_1, T_2, and $M(T_1)$ and $M(T_2)$ are the modifiers of the sentences T_1 and T_2, respectively. The similarity of the central word $Sim(H_1, H_2)$ is calculated using the Word2Word model mention above. B is the weight of the central word similarity in the calculation, where $\beta = 0.55$.

For the similarity of the modified phrase, if the two sentences both have only one modifier, then similar to the calculation of the central word similarity, the similarity can be estimated directly by the cosine of the angle of the word vector. However, if there are multiple modifiers in two sentences and the number is not always the same, then the similarity between the elements of the phrase should be weighted comprehensively. Here, we consider $M(T_1)$ and $M(T_2)$ as two sets of words, assuming that I, J word elements are respectively included, assuming the $\mathrm{S_{ij}}$ is the similarity between the *i*-th word in $M(T_1)$ and the *j*-th word in $M(T_2)$, then the similarity matrix can be obtained as follows:

$$\begin{bmatrix} S_{11} & \cdots & S_{1J} \\ \vdots & \ddots & \vdots \\ S_{I1} & \cdots & S_{IJ} \end{bmatrix}$$

Then the similarity between the word collection $M(T_1)$ and $M(T_2)$ is as follow:

$$Sim\left(M(T_1), M(T_2)\right) = \frac{1}{2} \times \left(\frac{\sum_{(ij)} S_{ij}}{I} + \frac{\sum_{(ij)} S_{ij}}{J}\right) \times \frac{I}{J} \ , (I < J) \tag{2}$$

In the above formula, if $I > J$, the corresponding *I/J* should be transformed into *J/I*, so as to ensure that the similarity of the two texts with little difference in the number of words have a relatively high similarity.

In general, this experiment compares and analyzes the characteristics of data in the field of building materials, and carries out a special part-of-speech tagging of the data, that is, the semantic role labeling of the central words and modifiers for all text data. Then, the similarity between the sentences is calculated by the algorithm mentioned above, wherein the similarity calculation between the specific two words utilizes the cosine similarity between the word vectors calculated by the word vector model.

4. Experimental Results and Analysis

4.1. Experimental Environment and Data

The machine environment of this experiment is as follow, Intel(R) Core(TM) i5-3337u CPU @ 1.80GHz, 3.8GB of RAM, 64-bit operating system, the data processing tool is python3.5.

At present, most Chinese text similarity calculation test corpora are artificially constructed data sets. The corpus used in this paper is the collected building materials data published by some building materials companies, and then the similar sentences are filtered and sorted to obtain the final experimental corpus. Before the experiment begins, it is necessary to manually construct the standard set and the test set. First, select a certain amount of sentences in the corpus that match the modifier and the central word structure as the test set. Six sentences are selected as the standard set. The standard similar sentences are the six sentences selected from the test set that are most similar to the standard set sentences. The other sentences in the test are used as noise samples. In the end, the test corpus we built contains a total of 846 sentences with participles and special semantics, most of which have more than 6 words.

4.2. Comparative Experimental Method

In order to better reflect the effect of the algorithm, another commonly used method of calculating similarity is used in the same corpus and environment for comparative experiments and quantitative analysis. Here we use the Vector Space Model (VSM) based on the word bag model to calculate the similarity for comparison experiments.

The construction of the vector space model is to first convert the text into a set of feature items without considering the position order of the feature items and the relationship between them. A text corresponds to a vector, so calculating the text similarity is equivalent to calculating the similarity between different vectors. The *TF-IDF* weight calculation method [12] is the most common weight calculation method. For example, about the feature item t_i in the text, the corresponding weight is as follows:

$$W_i = TF_i \times IDF_i \tag{3}$$

$$IDF_i = log\frac{N}{n_i + \alpha} \tag{4}$$

where *TF* is the frequency at which the feature t_i appears in the total texts, and *IDF* is the frequency of occurrence of the feature item t_i in the global text *D*. The *N* means that there are *N* pieces of text in the global text set, n_i is the total number of articles in which the feature item has appeared, and α is the coefficient of experience, usually recorded as 0.01. If a feature item appears only in a small amount of the text, it may be the central of the text, and a large *IDF* value will be obtained to increase the overall weight. A

commonly used method of calculating the similarity is to calculate the cosine of the two vectors, that is, if the eigenvectors of the two texts T_i and T_j are as follows:

$$V_i = (W_{i1}, W_{i2}, \dots, W_{in}) \tag{5}$$

$$V_j = (W_{j1}, W_{j2}, \dots, W_{jn}) \tag{6}$$

The angle between the two vectors is θ, then their similarity is as follow.:

$$S(T_i, T_j) = cos\,\theta = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{n} w_{ik}^2)(\sum_{k=1}^{n} w_{jk}^2)}} \tag{7}$$

4.3. Experimental Result Evaluation Criteria

In this experiment, we used commonly used evaluation criteria: recall rate, accuracy rate and f1-score. First, we take out the statements in the standard set in turn, and calculate the similarity values with each sentence in the test set, and sort them by decrement. Then the recall rate R_i, the accuracy P_i and the F value are as follows:

$$R_i = \frac{M_i}{N_i} \times 100\% \tag{8}$$

$$P_i = \frac{N_i}{S_i} \times 100\% \tag{9}$$

$$F = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{10}$$

The number of standard similar sentences of the *i-th* statement is N_i, which includes the number of M_i in the first N_i sentences calculated by the algorithm. The minimum number of sentences to be recalled for recalling all standard similar sentences is S_i. For the convenience of calculation, if S_i exceeds 20, the total record as 20. Correspondingly, the average recall rate, average accuracy rate, and average F value of each group can be averaged.

4.4. Experimental Results and Analysis

This experiment is aimed at the similarity algorithm research of Chinese texts in the field of building materials. Method one uses the similarity calculation method based on vector space, and method two uses the similarity calculation method based on part-of-speech tagging, which combines the word vector model to assist in calculations. For the selected standard text set, two rounds of testing are performed, and the results obtained are as follows.

Table 1. Experimental Results

Method	Round One			Round Two		
	Recall rate (%)	Accuracy rate (%)	F score (%)	Recall rate (%)	Accuracy rate (%)	F score (%)
One	61.33	46.67	52.33	57.67	45.33	50.33
Two	66.67	69.00	68.00	76.67	75.00	76.00

The above two rounds of results are averaged, and the results are shown in Fig. 1.

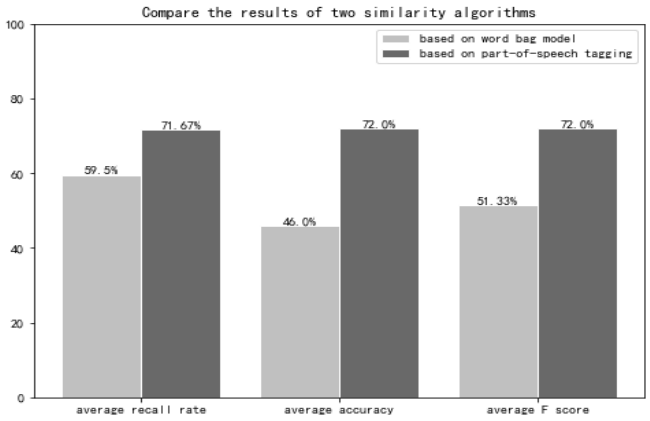

Fig. 1. Comparison of two similarity algorithms.

It can be seen from the Table 1 that the similarity calculation method based on part-of-speech tagging and word vector model is significantly improved compared with the method based on vector space model. By observing the erroneous results of the first method, it is found that for sentences with fewer words but similar semantics, a low recall rate is generated, resulting in a low final *F* value. The second method considers the semantics of the sentence, improves the weight of the central word, and integrates the semantic information of the sentence to calculate the similarity, so the effect is better.

5. Conclusions

5.1. Summary

This paper fully combines the data characteristics of building materials corpus, and proposes a method based on part-of-speech tagging and word vector model to calculate the similarity between Chinese texts. Compared with the traditional method, the method of this paper uses part-of-speech tagging. When calculating the similarity, the semantic information of the sentence is comprehensively considered, so that the final test result is more accurate.

However, the efficiency and scalability of the current algorithm need to be optimized, and further research is needed for the more complex text similarity calculation.

5.2. Further Work

In the next step, we will conduct research and analysis from the aspects of efficiency and scalability of the algorithm. What's more, the text of this experiment is not too much, so the algorithm does not run for a long time, but if you process a large amount of text data later, you may need a more efficient method. The neural network related algorithm may be a good choice. In addition, this experiment is aimed at text in a specific field, so the algorithm is not highly scalable, and it is necessary to consider more common Chinese texts in semantic and syntactic analysis to improve the scalability and portability of the algorithm.

References

[1] Anette, H. (2003). Improved automatic keyword extraction given more linguistic knowledge. *ACL*. 2003.

[2] Chen, E., & Jiang, E. (2017). A review of research on text similarity calculation methods. *Data Analysis and Knowledge Discovery*.

[3] Tian, K., Ke, Y., & Sui, Z. (2016). Chinese sentence similarity algorithm based on semantic role labeling. *Chinese Journal of Infomation*.

[4] Guo, S., & Xing, D. (2016). Research on sentence similarity calculation and related applications based on word vector. *Modern Electronic Technology*, *39(13).*

[5] Zhang, H., & Shi, S. (2014). Which performs better for new word detection, character based or Chinese word segmentation based? *IEEE*.

[6] Mai, F., Li, D., & Yue, X. (2011). Research on Chinese word segmentation technology based on two-way matching method and feature selection algorithm. *Journal of Kunming University of Science and Technology*, *36(1).*

[7] Hong, M., Zhang, K., Tang, J., & Li, J. (2006). Cineses part-of-speech tagging method based on conditional random fields (CRFs). *Journal of Computer Science*.

[8] Wei, O., Wu, J., & Sun, Y. (2000). Analysis and improvement of Chinese part-of-speech tagging method based on statistics. *Journal of Software*.

[9] Sharon G., & Thomas, L. G. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. *ACL.*

[10] Xue, S., & Niu, Y. (2016). Research on Chinese text similarity based on vector space model. *Journal of Electronic Design Engineering*, *24(10).*

[11] Scharolta, K. S. (2015). Adapting word2vec to named entity recognition. *Proceeding of the 20th Nordic Conference of Computational Linguistics*.

[12] Zhou, L., Yu, H., & Guo, C. (2015). Research on text similarity algorithm based on improved TF-IDF method. *Journal of Taishan University*, *37(3).*

Zhixin Ma was born in Gansu Province, China. He received the B.S. degree from the Lanzhou University, Lanzhou, in 1994. He received the M.S. degree from the Lanzhou University, in 1997. He received the Ph.D degree from Lanzhou University, in 2005. Currently, he is the vice president of the School of Information Science and Engineering, Lanzhou University. He is currently engaged in the research of data mining and software engineering. His main research interests are sequential pattern mining, skyline query, multi-criteria decision analysis, non-deterministic data analysis and security evaluation technology.

Mengguang Li was born in Henan Province, China, in 1994. He received the B.S. degree from the North China University of Technology, Beijing, in 2014. He is currently pursuing the M.S. degree with the Lanzhou University, Lanzhou. His research interests include natural language processing and data mining.