A Method for Calculating the Similarity of TF - IDF Texts for Synonyms in Biomedical Domains

Miao Hao[1, a)], Ke Fan [2, b)]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
[2] Institute for Bioinformatics, Chongqing University of Posts and Telecommunications, Guangzhou 400065, China.
[a)]78019048@qq.com
[b)]514068102@qq.com

Keyword: FT-IDF texts, Synonyms, Biomedical domains

Abstract. In the traditional text similarity calculation, most of the TF-IDF method. TF-IDF establishes the word frequency vector for the text, and calculates the cosine between the vectors as the similarity of the text. The algorithm is widely used in many search engines, information retrieval system can be seen, but in the text of the vocabulary processing is not ideal. The synonyms between professional phrases are not perceived by models, and they are used as different words to calculate similarity. In this paper, synonymous with biomedical field as an example, in the TF-IDF model embedded synonyms recognition function. Firstly, this method acquires the synonyms of the vocabulary in the biomedical field and establishes the synonyms, then identifies the synonyms in the TF-IDF model and calculates the better weight of the phrase. The experimental results show that this method can effectively improve the precision of text similarity calculation in biomedical field, and it is a more effective than the traditional TF-IDF text similarity calculation method.

INTRODUCTION

With the rapid growth of information resources, users are more concerned about how to find the information from the massive information resources. The similarity calculation of the text plays an important role in this field and has a wide range of applications in many fields: in the field of information retrieval, text similarity calculation is considered to be one of the best ways to improve the retrieval effect [1].Text similarity calculation is commonly used TF-IDF (Term Frequency & Inverse Documentation Frequency) method. It uses the frequency frequency TF (Term Frequency) and the inverse document frequency IDF (Inverse Document Frequency) as the document feature value. The text is modeled as a word frequency vector based on the Vector Space Model (VSM). And then calculate the vector cosine as the similarity between the text [2].

Now the arrival of large data age, making a large number of text data from the more accurate access to information has become an important issue of scientific research. We must make full use of the text before the different terms of the relationship, not just the frequency of each word appears. One of the obvious relationships is the synonyms.

So, this paper presents a method for calculating the similarity of TF-IDF texts for synonyms in biomedical domains, and experiments on the proteomic experimental metadata set. The experimental results show that the method can improve the accuracy of recognition. This method uses the text of the biological field to construct the text library, and combines the synonyms in the vocabulary to establish the synonyms based on the field of biomedical ontology[3]. First identify the synonyms in the text and treat them as the same term. and then calculate the text similarity based on TF-IDF method. This method avoids important information that is lost when using different biomedical phrases in the text to express the same meaning. Thereby improving the accuracy in the biomedical field text similarity.

TF-IDF METHOD AND ITS COMBINATION WITH SYNONYMS

Traditional TF-IDF

The TF-IDF method is the most typical of the methods of text similarity. The method is based on the following empirical observation, the text is expressed as the text of the n weighted word terms of the composition of the vector [4-6].

Term Frequency. The more the number of occurrences of a word in a text, the more relevant it is to the subject of the text. It should be noted that there are many specific words in a particular language environment that do not have this feature and should be excluded. these words named "Stop word", like "a" and "an".

Inverse Document Frequency. The more the number of occurrences of a term in the text of a set of texts, the poorer the difference, for example: in a collection containing 1000 texts, if an item A appears in 100 texts and the other entry B appears only in 10 texts, the term B has a better ability to distinguish it than A.

According to the TF-IDF concept, the TF-IDF method calculates the weight value of the feature t_i as follows[7-9]:

$$\begin{aligned} w_i &= TFIDF(w_i) = tf_{w_i} \times idf_{w_i} \\ &= tf_{w_i} \times \log(\frac{N}{n} + 0.01) \end{aligned} \tag{1}$$

tf_i represents the frequency at which feature t_i appears in text D. idf_i represents the inverse document frequency of t_i. K represents the total number of texts in the text library. n indicates the number of text that contains the feature t_i in the text library. Finally, add 0.01 to prevent the calculation of the return value of 0.

Each of the terms in the text set is subjected to the above analysis to obtain the TF-IDF value for each term in each text. And then use these TF-IDF values to establish a vector model for each text, and to calculate the similarity between texts by calculating the cosine similarity between vectors.

Combination of Biomedical Synonyms and TF - IDF Methods

There are a wide variety of professional synonyms in textual data in various fields today, as is the case in biomedical fields. The traditional TF-IDF method is based on word frequency to perform similarity calculation. When two biomedical texts with the different synonyms are used in the same content, the similarity calculated is far from the actual situation.

In this paper, we propose a method for calculating the similarity of TF-IDF texts for synonyms in biomedical domains. The method first identifies and combines the characteristics of the synonyms of biomedical synonyms. And then use the cosine method to measure the similarity of the two texts. When a text feature terms t_i and t_j are each other biomedical applications synonyms. Then the TF-IDF weights of the two feature terms are calculated and calculated in the following formula (3):

$$\begin{aligned} w_i &= w_j = (TF_i + TF_j) \cdot IDF(w_{n'}) \\ &= (tf_i + tf_j) \times idf_{n'} \\ &= (tf_i + tf_j) \times \log(\frac{K}{n'} + 0.01) \end{aligned} \tag{3}$$

$n' = n_i + n_j$ represents the sum of the number of texts containing the synonym A and B. Based on the above ideas, Figure 1 shows the data flow diagram of the text similarity calculation program proposed in this paper, which combines synonyms and TF-IDF methods in biomedical field.

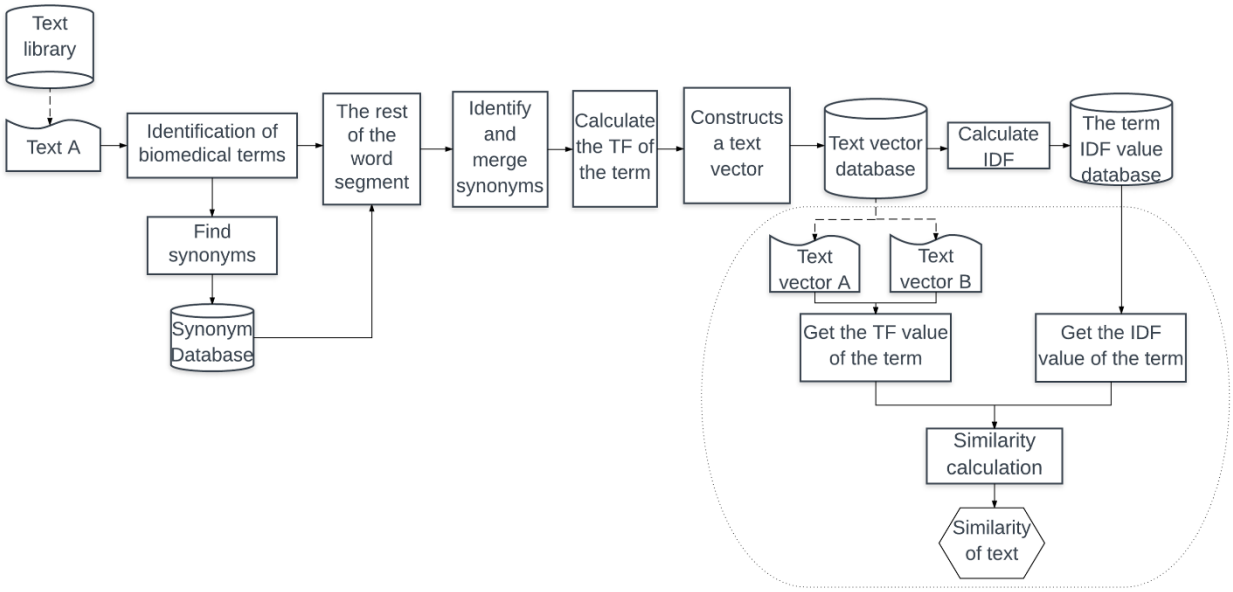

FIGURE 1. Data flow diagram of A Method for Calculating the Similarity of TF - IDF Texts for Synonyms in Biomedical Domains

Figure 1 illustrates the data flow of two biomedical text similarity calculations. First of all we extract one of the texts A from the text of the text library, and identify the biomedical synonyms in the text, then divide the remaining words into regular words and get the remaining lexical items. Get the set of text feature terms denoted by $t_1, t_2 \cdots t_n$. Then combine the synonyms to identify, merge the synonyms to get all the features that appear in the text $t_1, t_2 \cdots t_m$. And calculate the TF of the feature, construct the text vector and store the result in the text vector database. When all the text in the text library has been processed as described above. We get a complete text vector library. The IDF value of each feature is calculated based on the value of the feature TF and stored in the database. At last the dashed arrow in the figure shows the text vector A and the text vector B, which are taken out from the text vector library constructed in the previous text, and takes out all the terms and their TF values. Then, combine the data in the IDF value database. Finally, the similarity of two texts is calculated according to the cosine similarity.

DATA COLLECTION AND PROCESSING

Acquisition and Preprocessing of Text Sets

First, get all the experimental data in text PRIDE database (Get address: http://www.ebi.ac.uk: 80 / pride / ws / archive / project / list? Show = 100 & page = 0 & order = desc). Second, using JAVA scripting language, according to experimental proteomics ID download metadata text, composition data source corpus all file downloads(One example: http://www.ebi.ac.uk: 80 / pride / ws / archive / project / PXD336).

The text obtained in the PRIDE database is JSON-formatted text, and its contents are formatted and stored in the associated tags. After studying and analyzing the content of each tag, we decided to use "title", "projectDescription", "sampleProcessingProtocol", "dataProcessingProtocol" the contents of the four labels as a description of the experiment. Then, the contents of these four tags are intercepted by JAVA programming and stored in the corresponding ID text to construct the corpus for experiment.

Construction of Synonyms in Biomedical

In this paper, we choose BioPortal [10](HTTP://bioportal.bio ontology.org) as a tool for synonym recognition. BioPortal is the world's most comprehensive biomedical ontology database. It provides services for class, annotation, annotation resources discovery and biomedical ontology identification, and also supports the provision of biomedical ontology services to users through the

Web service API. This paper uses the Ontology Recommender provided by BioPortal for professional vocabulary recognition, identifying a total of 53163 biomedical vocabularies from 7453 texts. And we using the Annotator provided by BioPortal for the synonym relation query, extracts synonyms from the returned results, and stores the vocabulary and its synonyms in the local database for a total of 31259 words. The steps to build a synonyms database using BioPortal are shown in Figure 2.

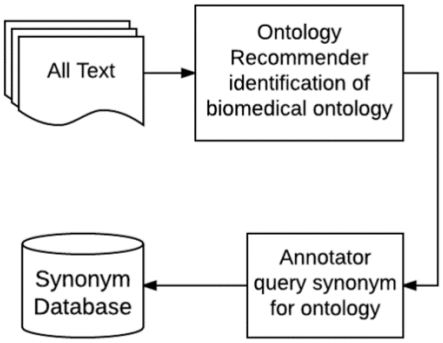

FIGURE 2. Construction of biomedical thesaurus repository

EXPERIMENT

In this paper, two experiments PXD336 and PXD998 are randomly selected to verify the effectiveness of the algorithm. For convenience, this paper constructs two replacement texts: Replace the word "human" in the original text with its biomedical synonym. Named PXD336 (substitution) and PXD998 (substitution). Table 1 shows the frequency (TF) of the word "human" and its synonyms in the four literatures.

TABLE 1. Table 1 The "protein" of the experimental text and the word frequency of its synonyms

Term / Text ID	"human"	"human being"	"Modern"
PXD336	55	0	0
PXD998	16	0	0
PXD336（substitution）	0	0	55
PXD998（substitution）	0	16	0

TABLE 2. Comparison of Two Similarity Methods

No	Text 1	Text 2	TF-IDF algorithm	This paper algorithm
1	PXD336	PXD998	0.311	0.483
2	PXD336	PXD998（substitution）	0.109	0.491
3	PXD336（substitution）	PXD998	0.193	0.505
4	PXD336（substitution）	PXD998（substitution）	0.050	0.504

In order to verify the correctness of the algorithm. In this paper, the text given in Table 1 was

compared and calculated to obtain the experimental results in Table 2. First, we use two algorithms to calculate the similarity of the two original texts. Subsequently, we cross-calculated the similarity of the four texts , and the results are shown in Table 2 and Figure 4 shows the graphical representation of these results. It is shown that the similarity score obtained by the TF-IDF algorithm is significantly reduced in the post-replacement comparison, from 0.050 to 0.311, which indicates that the feature The substitution of the synonyms for the term "protein" significantly affects the results, and the TF-IDF algorithm loses the replaced information because it can not perceive the synonyms. But the similarity calculated by the algorithm of this paper has been kept at around 0.48, which is not much floating, which indicates that the synonyms have no significant effect on the algorithm.

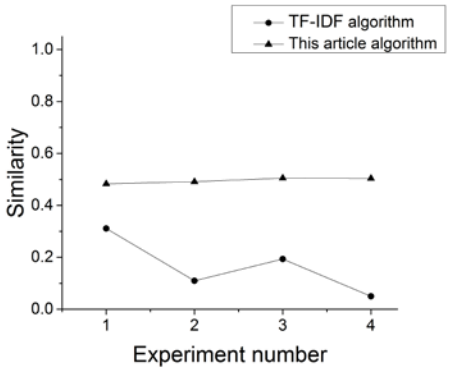

FIGURE 3. Similarities on cross comparison of the texts

CONCLUSION

The traditional TF-IDF method can not effectively identify the synonyms between professional vocabularies, and thus lose a lot of information in the literature mining work often need to calculate the similarity between the text. In order to compensate for this deficiency, this paper presents a method for calculating the similarity of TF-IDF texts for synonyms in biomedical domains. This method identifies the professional terms and synonyms in the text by calling the biomedical ontology knowledge base service, and uses the synonyms as the same phrase to calculate the feature terms A and B. Finally, the similarity of the text is calculated by the cosine value. The experimental results show that this method can make full use of the information of professional synonyms in biomedical field, and give more accurate similarity between text.

The results of this study further enhance the ability of literature mining and knowledge discovery in biomedical field, and the method can be extended to other fields to improve the accuracy of similarity calculation and the ability of text knowledge mining. The next step, we will be based on the existing results, in the text of the word segmentation algorithm for further research, the biomedical text of the word more accurate, and better improve the text similarity of the calculation results.

REFERENCES

[1]. H. Gomaa W, A. Fahmy A. A Survey of Text Similarity Approaches[J]. International Journal of Computer Applications, 2013, 68(13): 13-18.

[2]. Lu Song, Li Xiaoli,Bai Shuo.Improvement of Calculation Method of Word Weight in Document [J].Chinese Journal of Information,2000,14(6):8-13,20.

[3]. Yang Chunyuan, Li Mansheng, Zhu Yunping. The Construction, Evaluation and Application of the Ontology of Biomedicine [J]. Chinese Science: Life Sciences, 2013,43(3):223-239.

[4]. Fan Xiaochao, Zhang Chongyang, Deng Xiongwei,et al. Text Feature Weighting Method Based on Mutual Information[J]. Computer Engineering and Applications, 2015, 51(13):145-148

[5]. Wang Xiaolin, Xiao Hui, Tai Weipeng. Research on text similarity detection system based on

Hadoop platform[J]. Computer Technology and Development, 2015, 25(8):90-93.

[6]. Ou Yangning, Luo Yan.Text Similarity Calculation Based on Domain Feature Word Weighting[J]. Computer Engineering and Design, 2012, 33(11):4338-4342.

[7]. Mingyong Liu, Jiangang Yang. An improvement of TFIDF weighting in text categorization[J]. International Conference on Computer Technology and Science, 2012, 12(47): 44-47.

[8]. Paik J H. A novel TF-IDF weighting scheme for effective ranking[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013:343-352.

[9]. Doan A H. Learning to map between ontologies on the semantic web[C]// International Conference on World Wide Web. ACM, 2002:662-673.

[10]. Salvadores M, Alexander P R, Musen M A, et al. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF[J]. Semantic web, 2013, 4(3): 277-284.