Text Similarity Calculation Method Based on Hybrid Model of LDA and TF-IDF

Jiangyao Wang
School of Information Science and Engineering
Ocean University of China
Qingdao, China, 266000
wjy7265@stu.ouc.edu.cn

Wenhua Xu
School of Information Science and Engineering
Ocean University of China
Qingdao, China, 266000
xwh@ouc.edu.cn

Wenhao Yan
School of Information Science and Engineering
Ocean University of China
Qingdao, China, 266000
y_wenhao@163.com

Caixia Li
School of Information Science and Engineering
Ocean University of China
Qingdao, China, 266000
1877008918@qq.com

ABSTRACT

The traditional TF-IDF-based text similarity calculation model uses statistical methods to map text to the keyword vector space and convert the similarity of text into the distance between text vectors. Such methods have problems such as high computational dimensions, sparse data, and inability to take advantage of the semantic information contained in the text itself, so the results obtained are not as similar as the physical text. The text similarity model based on the topic model changes the traditional spatial similarity of keyword vectors, and can fully utilize the semantic information contained in the text itself. But this approach ignores the effect of words on text semantic representations with different weights. In the process of converting text into topic feature space, valuable information is lost. In view of the above problems, this paper proposes a text similarity hybrid model (L-THM) integrating LDA and TF-IDF for calculating text similarity. The model uses the semantic information contained in the text itself and the keyword information reflecting the text to comprehensively analyses and calculates the similarity between the texts. The experimental results show that the hybrid model can better represent the text information than the single model, and obtain a good F value in the cluster, which effectively improves the text similarity calculation effect.

CCS Concepts

• **Computing methodologies → Natural language processing**

Keywords

LDA; TF-IDF; topic model; text similarity; hybrid model.

1. INTRODUCTION

With the development of information technology, modern society has entered the era of big data. Since the amount of information is increasing at a high speed every day, it is very difficult for people to find accurate information in the face of massive information. At the same time, the surge in data and data retrieval results are unsatisfactory, and it has caused great inconvenience to all walks of life. Search engines can help people search for the information they need from massive amounts of Internet data. Similarity matching is one of the core technologies of search engines. The quality of similarity matching results directly relates to search efficiency, resource consumption, and quality of search results. Among them, text similarity is the basis of text search.

The TF-IDF-based vector space model [1] is a widely used text similarity calculation method. The model converts the text into a vector space of keywords, and calculates the similarity between the texts by calculating the distance between the vectors. For most corpora, the number of texts and words is huge, and the corresponding vector space has a high dimension. And the text of the unrelated category, the vector data is very sparse, which leads to poor accuracy of text similarity calculation. Based on the text similarity of the LDA topic model [2], which maps high-dimensional text vectors to low-dimensional semantic space-themes, effectively reducing the dimensions. However, the LDA topic model considers the potential semantic relationship between texts, ignoring the influence of words on text, and losing a lot of valuable knowledge contained in texts, which affects the calculation accuracy of text similarity.

Aiming at the shortcomings of the above methods, this paper proposes a method based on LDA and TF-IDF weighted hybrid model to calculate text similarity, considering both the potential semantics between texts and the influence of words on text. Experiments show that the method shows better results in calculating the similarity of text similarity with respect to a single LDA topic model and a single TF-IDF model.

The innovations of this paper include a new text similarity calculation method (L-T Hybrid Model) by thinking about the advantages and disadvantages of TF-IDF model and LDA model. In the TF-IDF model, in order to obtain the valid keywords of the text, the parameter γ is proposed by thinking about the TOP-K algorithm [3-4]. A linear parameter λ is proposed to represent the weight of the LDA topic model when the text similarity is weighted. In most current studies, the determination of the value

CSAI2019, December 6–8, 2019, Beijing, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7627-3/19/12…$15.00
DOI: https://doi.org/10.1145/3374587.3374590

of the parameter λ is generally determined by means of parameter adjustment. This paper proposes an empirical formula for calculating λ values, which can obtain better text similarity values. The discrimination degree Q of the similarity measure method is proposed to indicate whether a similarity measure method can distinguish whether different texts are similar.

This paper is divided into several parts, which are structured as follows: In Section 2, introduce the related work about text similarity calculation method. In Section 3, introduce relevant research and methods. In Section 4, introduce the details of proposed hybrid text similarity calculation method. In Section 5, introduce experimental setup and results.

2. RELATED WORK

In the era of big data, people are eager to get the information they need from a large amount of information. To this end, a variety of applications have been created, such as document review, text classification, and automatic question answering systems, etc. . One of the key technologies of these application scenarios is text similarity computing technology. Therefore, the processing of texts is becoming more and more important, which has attracted many scholars to engage in related research. Most scholars at home and abroad divide text similarity calculation into corpus-based (statistical) methods, string-based methods, feature-based methods, and knowledge-based methods, etc.[5]. Only statistical-based text similarity calculation methods are studied here.

For the existing research, the corpus-based methods can be divided into three categories: one is the TF-IDF method for text similarity calculation; the other is the LDA method for text similarity research; the third is the use of neural networks. The calculation of text similarity based on TF-IDF method is first used for information retrieval. The similarity between the items to be searched and the text is matched to find out the information similar to the item to be queried. G. Salton [6] et al. improved TF-IDF methods, using the term frequency and text frequency to sort the text, ignoring the potential implicit semantics of the text. Hiemstra [7] proposed a language-driven information retrieval probability model, and obtained text weight information according to the TF-IDF method, giving the term probability meaning. Abhishek Jain [8] et al. believe that the correlation between texts does not increase linearly with the frequency of keywords, and proposes a weighted definition of the TF-IDF method to obtain the keyword language of the text. Xu [9] et al. proposed a text similarity calculation method based on singular value decomposition and semantic correlation, considering the position weight of each word and the maximum correlation semantics of the word with other words.

For the LDA topic model, the potential semantic relationship between texts is considered, and the dimensions of the text vector are also reduced. Asli [10] et al. used the LDA model in question and answer (QA), and improved the performance of the Q&A ranking system, but ignored the sparseness of shorter text. Chen [11] et al. used the similarity calculation method based on the LDA topic model to apply the relevant Weibo recommendation through the related research on Weibo content. And compared with the WordNet method, it shows that the LDA method has a good effect when dealing with long text, and the effect on short text is not very good. Rus [12] et al. proposed a measure based on LDA to define semantic similarity at the word and sentence level, which only considers the influence of words under the topic on the text. On the other hand, experiments have also shown that the similarity calculation is plagued by data sparsity. Shao[13] et al. proposed a text similarity calculation algorithm based on hidden topic models and word co-occurrence analysis. In LDA, the words under each topic are analyzed for all corpora, so that the co-occurrence of the calculated words cannot establish a one-to-one correspondence with each text. Poria[14] et al. proposed a new framework based on the LDA model, called Sentic LDA. It integrates the common-sense calculation in the LDA algorithm to obtain the word distribution of related texts, and improving the clustering precision.

In the method of neural network, with the progress of deep learning in image, speech, etc., many scholars began to use the neural network model for natural language processing. Hu[15] et al. proposed a sentence matching model based on convolutional neural network model through the study of CNN model. Kenter [16] et al. proposed short text similarity calculation based on word Embedding. The method synthesizes different dimensional word vectors obtained under different conditions and maps word similarity to text similarity. Neculoiu [17] et al. used the LSTM framework to propose a similarity measure method for variable-length character sequences, and projecting variable-length strings into a fixed-dimensional embedded space. Arora [18] et al. obtained a weighted average of all the word vectors in a sentence to obtain a sentence vector, and achieved good results in sentence similarity calculation. Lin [19] et al. proposed short text similarity calculation based on LSTM coding, and processed the relationship of word sequences through LSTM network.

In this paper, the TF-IDF method and the LDA method are deeply studied to calculate text similarity. The TF-IDF method and the LDA method are used to calculate the text similarity and focus on changing the single factor to improve the effect of text similarity calculation, without considering the factors affecting the text comprehensively. A text consists of multiple topics, each of which consists of different words, so the semantics expressed by a text are determined by the combination of the topic and the words and the underlying semantic relationships that exist between them. This paper proposes a weighted hybrid model method to calculate text similarity, which effectively shows the semantics of the text to be expressed. On the other hand, through the above scholars' research, we can find that if a text is too short and a single method cannot effectively represent text information, then the effective combination of the two methods can also solve the problem of data sparsity caused by a single method to a certain extent.

3. RESEARCH AND METHODS

3.1 LDA Topic Model

The LDA topic model is a statistical-based mathematical model, which is mainly used to train a set of potential topics containing a certain probability from the existing text set. The characteristics of the text are described by the probability distribution of the topic, with strong semantic features that reflect the type of topic the text is to express.

In the LDA model, a text is generated as follows:

(1) Drawing generation a topic distribution $\boldsymbol{\theta}_d$ of text d from a Dirichlet prior $\boldsymbol{\alpha}$;

(2) Drawing generation the vocabulary distribution $\boldsymbol{\phi}_z$ of the topic z from the Dirichlet prior $\boldsymbol{\beta}$;

(3) Drawing generation the topic $z_{d,r}$ of the rth word of the text d from the multinomial $\boldsymbol{\theta}_d$ of the topic;

(4) Drawing generation the word $w_{d,r}$ from the multinomial $\boldsymbol{\phi}_z$ of the word;

Based on the above description, the LDA probability model diagram can be drawn, as shown in Figure 1. This probability model diagram is also called "Plate Notation". The shaded circles in the figure represent observable variables, the non-shaded circles represent latent variables, the arrows represent conditional dependencies between two variables, the boxes represent repeated sampling, and the number of repetitions is in the lower right corner of the box.

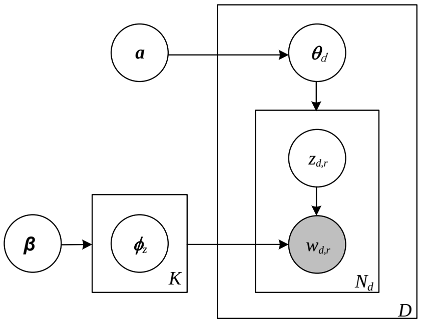

Figure 1. Model diagram representation of LDA.

A glossary of notations used in the LDA model is given in Table 1.

Table 1. Notation used in LDA model

Symbol	Description
D	total number of documents
K	number of topics
$\boldsymbol{\alpha}$	represents a priori parameter of the text topic distribution
$\boldsymbol{\beta}$	represents a priori parameter of word distribution on a topic
N_d	number of word tokens in document d
$z_{d,r}$	the topic of the r word in document d
$w_{d,r}$	the rth word in the document d
$\boldsymbol{\theta}_d$	multinomial distribution of topics on document d
$\boldsymbol{\phi}_z$	multinomial distribution of words on topic z

In the process of LDA, there are two parameters to be estimated, namely the parameter $\boldsymbol{\theta}$ of the text-topic multinomial distribution, the parameter $\boldsymbol{\phi}$ of the topic-word multinomial distribution, and there are many current estimation methods for the parameters, such as the variation Bayes Reasoning, Expectation Propagation algorithm[20] and Collapsed Gibbs sampling[21], etc., because the parameter inference method of Gibbs sampling is easy to understand and easy to implement, so the parameter estimation of LDA model mainly uses Gibbs sampling algorithm. Gibbs sampling is a type of Markov chain Monte Carlo method (MCMC)[22], which is calculated as follows:

$$P(z_r = k \mid \boldsymbol{z}_{-r}, \boldsymbol{w}) = \frac{n_{k,-r}^{(t)} + \beta_t}{\sum_{t=1}^{T}(n_{k,-r}^{(t)} + \beta_t)} \cdot \frac{n_{d,-r}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{d,-r}^{(k)} + \alpha_k)}$$

where $n_{k,-r}^{(t)}$ denotes the number of words t under the kth topic, not including the current token instance r; $n_{d,-r}^{(k)}$ denotes the number of the kth topic number of the dth document, not including the current instance r; T is the number of all words in the text set, and α_k is the Dirichlet a priori of the topic k, and β_t is the Dirichlet prior of the term t.

After obtaining the topic label for each word, the required parameter estimation formula is as follows:

$$\theta_{d,k} = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_d^{(k)} + \alpha_k)} \qquad \phi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)}$$

where $\theta_{d,k}$ represents the probability of the topic k in the document d, $n_d^{(k)}$ represents the number of words in the document d given the topic k, $\phi_{k,t}$ represents the probability of the word t in the topic k, and $n_k^{(t)}$ represents the number of times the word t is given to the topic k.

3.2 TF-IDF Model

The TF-IDF algorithm [23-24] is a commonly used technique in the extraction of text feature words based on statistical methods. It mainly evaluates the importance of a word to text and text sets by word frequency. It is mainly composed of two parts: word frequency and inverse text word frequency [25].

In a document, the term frequency (TF) is the frequency at which a word appears in the text, and the result is usually normalized to prevent it from being biased toward longer text.

$$TF_{d,t} = \frac{n_{d,t}}{N_d}$$

where $n_{d,t}$ represents the number of occurrences of the word t in the text d;

Inverse Document Frequency (IDF) indicates the importance of a word in a text set. The formula is as follows:

$$IDF_t = \frac{D}{1 + |\{d_i \mid t \in d_i\}|}$$

where d_i represents the ith text and $|\{d_i \mid t \in d_i\}|$ represents the number of texts containing the word t. To avoid the denominator not being 0, use $1 + |\{d_i \mid t \in d_i\}|$.

The main idea of TF-IDF [26] is that if a word appears frequently in a text and almost does not appear in other texts, it is considered that the word can distinguish other texts, that is, it has a strong power to representation the text. The calculation is as follows:

$$TF - IDF_{d,t} = \frac{n_{d,t}}{N_d} \cdot \frac{D}{1 + |\{d_i \mid t \in d_i\}|}$$

4. TEXT SIMILARITY CALCULATION

4.1 Hybrid Model

The text similarity calculation based on the LDA topic model is to obtain the text-topic probability distribution by mapping the text to the feature space of the implicit topic, and then find the similarity between the two texts. Since the text is composed of topics and the topic is composed of words, it can be said that the text is ultimately composed of words. LDA-based text similarity calculations only consider the impact of the topic. Text similarity calculation based on TF-IDF is to convert text into keyword vector space, considering only the influence of words on text, and losing the potential semantic information that text may exist. This paper combines the two, and proposes a hybrid model text similarity calculation method integrating LDA and TF-IDF—LDA-TF-IDF Hybrid Model (L-THM) based on LDA and TF-IDF, fully consider the advantages and disadvantages of the two, make up for each other, and improve the accuracy of text similarity calculation.

The specific algorithm is described as follows:

In the first step, a text-topic probability distribution is generated by the LDA model for calculating the topic similarity $S_{LDA}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)$ of each text.

$$S_{LDA}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)=\frac{\boldsymbol{\theta}_i\cdot\boldsymbol{\theta}_j^{\mathrm{T}}}{\|\boldsymbol{\theta}_i\|\cdot\|\boldsymbol{\theta}_j\|}=\frac{\sum_{k=1}^{K}\theta_{i,k}\cdot\theta_{j,k}}{\sqrt{\sum_{k=1}^{K}\theta_{i,k}}\cdot\sqrt{\sum_{k=1}^{K}\theta_{j,k}}}$$

where $\boldsymbol{\theta}_i$ is the topic probability vector of the ith text, $\boldsymbol{\theta}_j$ is the jth text topic probability vector, $\|\boldsymbol{\theta}_i\|$ is the norm of the ith text topic vector, $\|\boldsymbol{\theta}_j\|$ is the norm of the jth text topic vector, $\theta_{i,k}$ is the kth topic probability value of the ith text, $\theta_{j,k}$ is the kth topic probability value of the jth text.

In the second step, the TF-IDF model is used to find the keywords and weights of each text. In order to better represent the text, this paper proposes parameters γ (percentage) to get valid keywords for each text.

The specific process is as follows:

First, a word of text d is selected by TF-IDF, and the number of words is represented by N_d, then all sets of text words can be expressed as:

$$N=\{N_1,\cdots,N_d,\cdots,N_D\}$$

Then, select the words of the first γ of each text as valid keywords for the text. The formula for calculating the number of valid keywords M in a text is:

$$M_d=\lceil N_d\cdot\gamma\rceil$$

Then, by sorting the weights of the words, a valid keyword is obtained. A valid set of keywords for a text is W_d :

$$W_d=\{w_{11},w_{12},w_{13},\cdots,w_{1M_d}\}$$

where w_{1M_d} is the M_d th word of a text.

Finally, all the text keywords taken out form a vocabulary set V;

$$\begin{aligned}V&=\{W_1,\cdots,W_d,\cdots,W_D\}\\&=\{w_{11},\cdots,w_{1M_d},\cdots,w_{d1},\cdots,w_{dM_d},\cdots,w_{D1},\cdots,w_{DM_D}\}\end{aligned}$$

Find the vocabulary and calculate the TF-IDF value of each text in the vocabulary for calculating the similarity $S_{TF-IDF}(\boldsymbol{V}_i,\boldsymbol{V}_j)$ of the text-word.

$$S_{TF-IDF}(\boldsymbol{V}_i,\boldsymbol{V}_j)=\frac{\boldsymbol{V}_i\cdot\boldsymbol{V}_j^{\mathrm{T}}}{\|\boldsymbol{V}_i\|\cdot\|\boldsymbol{V}_j\|}=\frac{\sum_{v=1}^{|V|}V_{i,v}\cdot V_{j,v}}{\sqrt{\sum_{v=1}^{|V|}V_{i,v}}\cdot\sqrt{\sum_{v=1}^{|V|}V_{j,v}}}$$

where $\boldsymbol{V}_i$ is the vector of the ith text under the vocabulary; $\boldsymbol{V}_j$ is the vector of the jth text under the vocabulary; $\|\boldsymbol{V}_i\|$ is the norm of the vector of the ith text under the vocabulary; $\|\boldsymbol{V}_j\|$ is the jth text Vector norm under the vocabulary. $|V|$ represents the number of words in the vocabulary set V, $V_{i,v}$ is the TF-IDF value of the vth word in the vocabulary of the ith text, and $V_{j,v}$ is the TF-IDF of the vth word in the vocabulary of the jth text value.

The third step is a weighted hybrid calculation. The text similarity obtained by the LDA topic model is weighted with the text similarity obtained by the TF-IDF model, and the final text similarity s_{ij} is calculated.

$$s_{ij}=\lambda\frac{\sum_{k=1}^{K}\theta_{i,k}\cdot\theta_{j,k}}{\sqrt{\sum_{k=1}^{K}\theta_{i,k}}\cdot\sqrt{\sum_{k=1}^{K}\theta_{j,k}}}+(1-\lambda)\frac{\sum_{v=1}^{|V|}V_{i,v}\cdot V_{j,v}}{\sqrt{\sum_{v=1}^{|V|}V_{i,v}}\cdot\sqrt{\sum_{v=1}^{|V|}V_{j,v}}}$$

Where s_{ij} is the similarity between the ith text and the jth text, and $\lambda\in(0,1)$ is the weight parameter of the model.

Cosine similarity calculation is used in the calculation.

4.2 Model Weight Parameter

The hybrid model parameter λ represents the weight of the LDA topic model when the text similarity is weighted hybrid. In many studies, most of the experimental methods obtain appropriate λ values through adjusting parameters. The parameter adjustment process is complicated and time consuming, and the range of values is difficult to control. In order to obtain the λ value, this paper proposes an empirical formula for calculating the λ value. Not only can it reduce complex operations, but it can also improve the accuracy of similarity calculations. First, the discrimination index Q of the similarity measure is proposed. The matrix $\boldsymbol{S}$ is used to represent the similarity values between the two texts. For a similarity measure, if the texts are similar, the calculated similarity measure should be larger; if the two texts are not similar, the calculated similarity measure should be smaller, then the similarity measurement method has a high degree of discrimination. For example, If the first text is similar to the second text, then the value of s_{12} is high; otherwise, the value of s_{12} is low.

$$\boldsymbol{S}=\begin{pmatrix}s_{11}&\cdots&s_{1j}&\cdots&s_{1d}\\\vdots&\ddots&\vdots&\cdots&\vdots\\s_{i1}&\cdots&s_{ij}&\cdots&s_{id}\\\vdots&\vdots&\vdots&\ddots&\vdots\\s_{d1}&\cdots&s_{dj}&\cdots&s_{dd}\end{pmatrix}$$

In the experiment, the degree of discrimination is reflected by calculating the standard deviation of the similarity value of the measurement method. Then, for the calculation of the standard deviation σ_d of the similarity value between a text and other text:

$$\mu_d=\frac{1}{D-1}(\sum_{j=1}^{D}s_{dj}-1)$$

$$\sigma_d = \sqrt{\frac{1}{D-1}(\sum_{j=1}^{D}(s_{dj}-\mu_d)^2-(1-\mu_d)^2)}$$

where μ_d is the mean of the similarity values of the dth text and other texts. In the experiment, because each text has a similarity to itself of 1, in order not to affect the volatility of the standard deviation, the similarity between each text and itself is removed, and the number of texts becomes D-1.

Then, make the discrimination enough to measure a similarity calculation method as a whole, and the degree of discrimination is measured by the mean of the standard deviation of the similarity values between each text and other texts. The degree of discrimination Q of the text similarity calculation method is expressed as:

$$Q = \frac{\sum_{d=1}^{D}\sigma_d}{D}$$

It can be seen that the higher the value of Q, the more distinguishing the similarity method is whether the two texts are similar. Therefore, in the text similarity weighted mixed calculation, the proportion of the method is larger. The formula for calculating λ is as follows:

$$\lambda = \frac{Q_{LDA}}{Q_{LDA}+Q_{TF-IDF}}$$

where Q_{LDA} is the degree of discrimination of the text similarity calculation based on the LDA topic model, and Q_{TF-IDF} is the degree of discrimination of the text similarity calculation based on the TF-IDF model.

5. EXPERIMENTAL

5.1 Data and Evaluation Indicators

This paper calculates the similarity of Chinese and English corpora. The Chinese experimental data uses the Fudan text classification datasets and the English experimental data uses 20Newsgroups datasets. Table 2 shows the characteristics of each data set in the experiment.

Table 2. Experimental data

Dataset	Number of clusters	Number of texts in a cluster	Total number	Maxim-um text	Minim-um text
Fudan text	8	300	2400	58KB	1KB
20News groups	8	200	1600	15KB	2KB

The experiment uses the classical K-means algorithm [27] for clustering, and the F metric is used to measure the accuracy of text similarity. D_a is the number of texts of category a, D_b is the number of texts of cluster b, and D_{ab} is the number of texts belonging to a in cluster b, then the accuracy rate $p(a,b)$ and the recall rate $R(a,b)$ can be defined as:

$$p(a,b) = \frac{D_{ab}}{D_b} \qquad R(a,b) = \frac{D_{ab}}{D_a}$$

The F metric is defined as:

$$F(a,b) = \frac{2 \cdot p(a,b) \cdot R(a,b)}{p(a,b)+R(a,b)}$$

The F metric for global clustering is defined as:

$$F = \sum\nolimits_a \frac{D_a}{D}\max_b(F(a,b))$$

5.2 Experimental Steps

First, preprocess the text, load the corpus and perform word segmentation, remove the stop words, and use the result as a word bag. Then the text is vectorized and the LDA model modeling training text is performed. In the modeling process, the parameters required by the Gibbs sampling algorithm are set, where $\alpha = 50/K$, $\beta = 0.01$, and the number of iterations of Gibbs sampling is 1000. For the determination of the number of topics K, the effect of the number of topics K on the global clustering F value of the LDA model is considered. The experimental results in the Chinese corpus and the English corpus are respectively shown in figure 2 and 3.

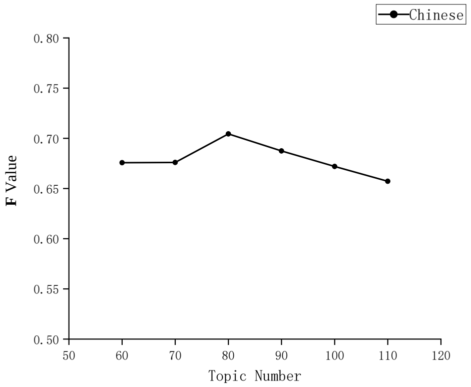

Figure 2. The number of Chinese corpus topic K and F values.

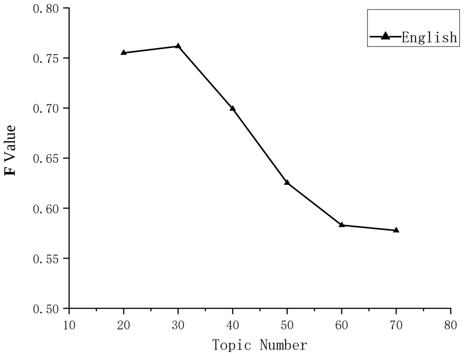

Figure 3. The number of English corpus topic K and F values.

It can be found from figure 2 that the F value is the largest when the Chinese corpus K is 80, indicating that the LDA model obtains the best effective information fit for the text set data, so the value of K is selected to be 80 in the experiment. Then the F value needs to be calculated for the English corpus. It can be concluded from figure 3 that when K is 30, the F value is the largest. Finally, a weighted hybrid similarity calculation is performed. In the process of calculating text similarity, there are two parameters γ and λ , both of which affect the value of similarity. Since γ affects the number of valid keywords for each text, which affects the standard deviation of the similarity between each text and other texts, that is, affect the size of the λ value. In order to be able to determine the value of λ , need to first determine the value of γ .

For the Chinese corpus and the English corpus, the experimental values were taken as γ -values of 0.1-0.9 and intervals of 0.1, and the F-value corresponding to TF-IDF was calculated. The results are shown in Fig. 4.

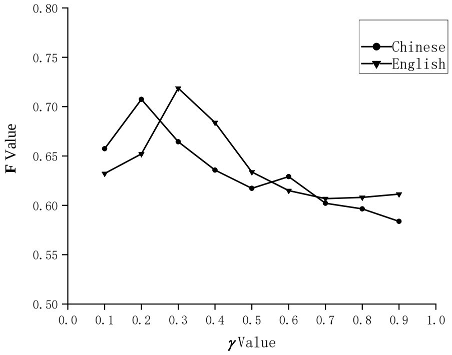

Figure 4. Chinese and English corpus γ values and F values.

It can be observed from figure 4 that, with the change of γ value, the F values of Chinese corpus and English corpus show a trend of increasing first and then decreasing. It indicates that there are too many keywords selected, including some invalid keywords, which reduces the value of text similarity; on the contrary, the selected keywords are not enough to represent text information. When the text-related noise is removed, the F value of the Chinese corpus $\gamma = 0.2$ is the largest. When the English corpus $\gamma = 0.3$, the F value is the largest. In addition, it can be seen that for the English corpus, the value of F is almost the same when γ is 0.7, 0.8, 0.9. The main reason is that the English corpus selects the text of the news class. The length of each text is limited. After multiplying the γ value, there is no significant difference in the effective keywords taken, which ultimately leads to little change in the F value.

5.3 Experimental Results

The optimal number of topics K and parameter γ are determined by experiments, and then different model experiments are performed to obtain the λ value.

For the Chinese corpus, get:

$$Q_{LDA} = 0.1340 \quad Q_{TF-IDF} = 0.0157$$

Then, $\lambda = 0.8947$

For the English corpus, get:

$$Q_{LDA} = 0.2456 \quad Q_{TF-IDF} = 0.0232$$

Then, $\lambda = 0.9134$

In order to verify the rationality of the discrimination degree Q proposed in this paper, a three-dimensional map is made for the Chinese-based topic model similarity matrix value and the TF-IDF model similarity matrix value.

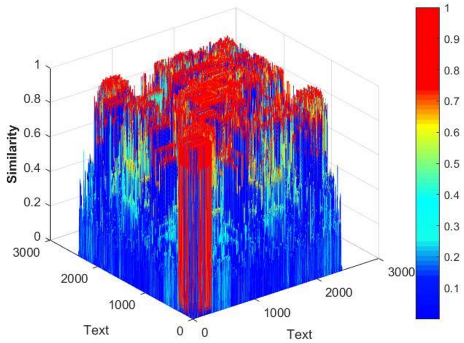

Figure 5. Chinese based on the topic model text similarity values three-dimensional map.

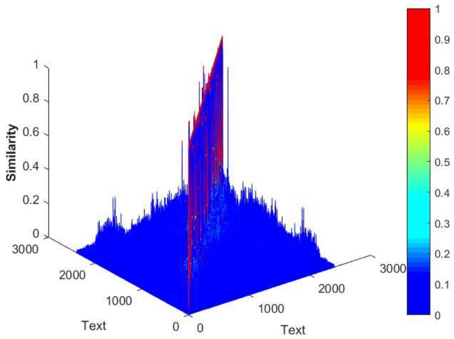

Figure 6. Chinese TF-IDF model text similarity values three-dimensional map.

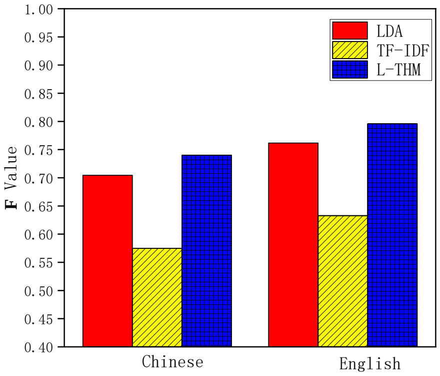

Figure 7. Clustering F-values for different models of Chinese and English texts.

In figure 5 and 6, it can be found that since the value of s_{dj} in the similarity matrix $\boldsymbol{S}$ is equal to the value of s_{jd} , the three-dimensional map is symmetrical. And the Chinese similarity value based on the topic model is in each interval, especially between 0-0.3 and 0.7-1, the number is large, and there is obvious discrimination. However, the similarity value of Chinese based on TF-IDF model is more between 0-0.3, and the other intervals are less, and the discrimination is not obvious. It can be seen that the above results are consistent with the degree of discrimination mentioned in this paper, which also verifies the correctness of the calculated λ value.

Clustered by different features, the corresponding F value is obtained. Specific experimental results are given in Table 3. Among them, TF-IDF is characterized by text keywords, LDA is text-topic distribution for feature clustering, and L-THW is clustered with text-topic and text's pre-γ keywords. In addition, Xu [9] et al. proposed the WordSim+N3 and ProMethod+N3 models and Shao[13] et al. proposed the LDA+JS+Computer and LDA+JS+Word Co-occurrence+Computer models are compared.

It can be seen intuitively from figure 7 that the cluster F value obtained by the Chinese-English corpus using the TF-IDF model is the smallest and the clustering effect is the worst. Further comparing the F values of the TF-IDF model, the LDA model and the L-THM in the Chinese and English corpora, the F value of L-THM is higher than the F value of the single TF-IDF model and the LDA model, and a better clustering effect is obtained. According to the calculation in table 3, the F value obtained by Chinese L-THM is increased by 5.07% compared with the LDA model, and the F value obtained by the English L-THM is 4.49% higher than that of the LDA model. Therefore, It can be concluded that L-THM effectively compensates for the deficiency of single model in calculating text similarity, and can reasonably represent text information and improve the accuracy of text similarity calculation.

Table 3. Model experiment results

Category	Chinese F value	English F value
LDA	0.7044	0.7617
TF-IDF	0.5748	0.6329
WordSim+N3	-	0.6000
ProMethod+N3	-	0.6200
LDA+JS+Computer	0.7252	-
LDA+JS+Word Co-occurrence+Computer	0.7353	-
L-THM	**0.7401**	**0.7959**

6. CONCLUSION

In this paper, a method based on LDA and TF-IDF weighted hybrid model (L-THM) is proposed to calculate text similarity, which improves the accuracy of text similarity calculation. This method considers both the potential semantics that may exist between texts and the influence of words on text semantic representations with different weights. By combining the topic information with the effective keyword information, it effectively solves the problem that the single factor cannot fully represent the similarity between the texts. Experiments show that the two have different degrees of representation on text information, and different weights can effectively express text semantics. The text similarity calculation based on TF-IDF model also has the problem of computing data sparseness. Increasing the text topic information solves the data sparse problem to some extent. This new text similarity calculation method will effectively improve the efficiency of data mining and other aspects. In further research, this method will be applied to examples, such as information retrieval, question and answer systems, and so on. In addition, the optimal setting and selection of the parameter γ is also an important issue worthy of further study.

7. ACKNOWLEDGMENTS

The work is funded by Natural Science Foundation of Shandong Province, China (No. ZR2015FM025), and by Qingdao National Laboratory of Marine Science and Technology Ao'shan Science and Technology Innovation Plan (No. 2016ASKJ07-05).

8. REFERENCES

[1] Drucker H, Shahrary B, Gibbon D C. Support vector machines: relevance feedback and information retrieval[M]. Pergamon Press, Inc. 2002.

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of Machine Learning Research Archive, 2003, 3:993-1022.

[3] Ronald F, Ravi K, Sivakumar D. Comparing Top k Lists[J]. Siam J on Discrete Mathematics, 2003, 17(1):28-36.

[4] Suh C, Tan V Y F, Zhao R. Adversarial Top-K Ranking[J]. IEEE Transactions on Information Theory, 2017, 63(4):2201-2225.

[5] H. Gomaa W, A. Fahmy A. A Survey of Text Similarity Approaches[J]. International Journal of Computer Applications, 2014, 68(13):13-18.

[6] G. Salton C B . Term weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1987, 24(5):513-523.

[7] Hiemstra D . A probabilistic justification for using tf×idf term weighting in information retrieval[J]. International Journal on Digital Libraries, 2000, 3(2):131-139.

[8] Jain A , Jain A , Chauhan N , et al. Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model[J]. International Journal of Computer Applications, 2017, 164(6):28-30.

[9] Li X, Yao C, Fan F, et al. A Text Similarity Measurement Method Based on Singular Value Decomposition and Semantic Relevance[J]. Journal of Information Processing Systems, 2017, 13(4)

[10] Celikyilmaz A , Hakkani-Tur D , Tur G . LDA based similarity modeling for question answering[C]// Proceedings of the NAACL HLT 2010 Workshop on Semantic Search. Association for Computational Linguistics, 2010.

[11] Chen X , Li L , Xu G , et al. Recommending Related Microblogs: A Comparison Between Topic and WordNet based Approaches[J]. Aaai Press, 2012.

[12] Rus V , Niraula N , Banjade R . Similarity Measures Based on Latent Dirichlet Allocation[C]// International Conference on Computational Linguistics & Intelligent Text Processing. Springer-Verlag, 2013.

[13] Shao M, Qin L. Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence[C]// International Conference on Software Engineering, Knowledge Engineering and Information Engineering. 2014.

[14] Poria S , Chaturvedi I , Bisio F , et al. Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis[C]// International Joint Conference on Neural Networks. IEEE, 2016.

[15] Hu B , Lu Z , Li H , et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[J]. 2015.

[16] Kenter T , Rijke M D . Short Text Similarity with Word Embeddings[C]// the 24th ACM International. ACM, 2015.

[17] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks[C]//Proceedings of the 1st Workshop on Representation Learning for NLP. 2016: 148-157.

[18] Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings[J]. 2016.

[19] Yao L, Pan Z, Ning H. Unlabeled Short Text Similarity With LSTM Encoder[J]. IEEE Access, 2018, 7: 3430-3437.

[20] Minka T. Expectation-Propagation for the Generative Aspect Model[C]// Conference on Uncertainty in Artificial Intelligence. 2002.

[21] Heinrich G. Parameter Estimation for Text Analysis[J]. Technical Report, 2008.

[22] Chernozhukov V, Han H. An MCMC approach to classical estimation[J]. Social Science Electronic Publishing, 2003, 115(2):293-346.

[23] Robertson, Stephen E. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Journal of Documentation ,2004,60: 503-520.

[24] Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification[J]. Expert Systems with Applications, 2011, 38(3):2758-2765.

[25] Kim D , Seo D , Cho S , et al. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec[J]. Information Sciences, 2018.

[26] Christian H, Agus M P, Suhartono D. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)[J]. 2016, 7(4):285.

[27] Xiong C, Hua Z, Lv K, et al. An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers[C]// International Conference on Cloud Computing and Big Data. IEEE Computer Society, 2016:265-268.