ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research, volume 205
The 2nd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2018)

Computer-aided Off-topic Composition Detection*

Qiang Qu
Department of Computer Science and Technology
Yanbian University
Yanji, China

Yahui Zhao**
Department of Computer Science and Technology
Yanbian University
Yanji, China
**Corresponding author

Rongyi Cui
Department of Computer Science and Technology
Yanbian University
Yanji, China

***Abstract*—Aiming at the problem that the lack of accurate and efficient off-topic detection algorithm for the current English composition teaching system in China, an off-topic detection algorithm based on LDA and word2vec was proposed in this paper. The algorithm used LDA to model the documents and trained the model with word2vec, with obtained semantic relation between document's topic and words, calculated the probability weighted sum of each topic and its feature words in the document. Finally, the off-topic compositions were selected by setting reasonable threshold. According to the different F-measures for the different number of topics in the document, the optimum number of topics was determined in the experiment. The experimental results show that the proposed method, with above 89% of F-measure, is more effective than traditional vector space model, and can realize the intelligent processing of off-topic compositions detection, which may be applied effectively in teaching of English composition.**

Keywords—off-topic composition detection; vector space model (VSM); latent Dirichlet allocation (LDA); semantic relations between words

I. INTRODUCTION

Composition is a narrative method to express thematic meaning by words, and the topic is the soul of a composition. If the content of an article deviates from the topic, off-topic will be caused, so that it will be unable to express the author's goal thoughts accurately and effectively, and the meaning of the composition will be lost. The core content of composition off-topic detecting is to calculate the similarity between texts [1], and at present the most classical method on text similarity is TF-IDF algorithm based on vector space model. The method constructs text vector by TF (Term Frequency) and IDF (Inverse Document Frequency), and calculates text similarity through the cosine of the angle between text vectors. This method neglects the semantic information of word itself in document, and doesn't consider the semantic similarity between words.

In view of the above shortcomings, a new method of text similarity calculation is proposed in this paper. Through modeling the document by LDA topic model, we obtained the topic of each document, the feature words of the topic and their probability distribution, and bound with the semantic relation between words got by word2vec training, and furthermore calculated the probability weighted sum of each topic of the document, by which it may be judged whether the composition deviate the topic. The proposed method can not only get more semantic information between terms, but also get the topic distribution of each document by modeling for each document in the corpus, which compensates the inadequacy that traditional vector space model method does not take into account the semantic information of the word itself.

II. LDA MODELING

A. LDA Model

LDA (Latent Dirichlet Allocation) model is a three-layer Bayesian production model [2] of "document-topic-word", proposed by Blei et al, which contains a three-layer structure of words, topics, and documents. It is a kind of unsupervised machine learning algorithm, and can be used to identify the potential topic information in large scope document set or corpus. LDA topic model can be represented by a probability graph model as shown in "Fig. 1".

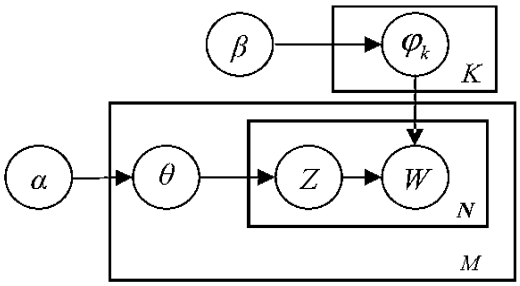

Fig. 1. LDA Directed probability graph model.

*Supported by State Language Commission of China under grant No. YB125-178 and The Education Department of Jilin Province of China under grant No. JJGZ2016-32.

LDA model is determined by the hyper parameters α and β, among which α represents the relative strength of the implicit topics in the document set, and β reflects the probability distribution of implicit topic itself. Implicit topics are expressed by Z, and θ represents document-topic probability distribution, and φk represents topic-word probability distribution. In a given document set D, it contains M documents, each of which contains N words.

B. Gibbs Sampling

The estimation on model parameter shall be carried out during constructing LDA model. Parametric inference method based on Gibbs sampling is easy to understand, simple to realize, and can extract topic from large scope document set very effectively [3]. T. Griffiths proposed to apply Gibbs sampling method to the parameter estimation of the LDA model [4]. The feature word probability distribution under every topic, and the topic probability distribution of each document are two most important parameters in LDA model.

The specific steps of Gibbs sampling algorithm are as follows [5]:

1) Initialization: The topic is initialized as a random integer ranging from 1 to T, i circulates from 1 to N, where N is the number of all the particular words in the corpus that appear in the document. This is the initial state of Markov chain.

2) Cyclic sampling: After iterating enough times, until Markov chain approach target distribution, Zi can estimate the value of φ and θ as follows:

$$\hat{\varphi}_k^{(t)} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t} \tag{1}$$

$$\hat{\theta}_m^{(k)} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{2}$$

Where $n_k^{(t)}$ represents the time that the k-th topic appears in t-th feature word, and $n_m^{(k)}$ represents the time that the m-th document appears in k-th topic. Through the Gibbs sampling indirectly obtained φ and θ values, their posterior distribution are Dirichlet distribution, which denoted by posterior probability P (Zi=k|Z¬i,w) and calculated as follows:

$$P(Z_i = k|Z_{-i},w) \propto \frac{n_{m,-i}^{(k)} + \alpha_t}{\sum_{k=1}^{K}(n_{m,-i}^{(k)} + \alpha_t)} * \frac{n_{k,-i}^{(t)} + \beta_k}{\sum_{t=1}^{V}(n_{k,-i}^{(t)} + \beta_k)} \tag{3}$$

Where Zi represents the topic variable corresponded by the i-th word, ¬i represents that the i-th item is not included, Z¬i represents the probability distribution of all topics Zk (k≠i), and $Z_{k,\neg i}^{(t)}$ indicates that the frequency of feature word t occurred in topic k, and $ZZ_{m,\neg i}^{(t)}$ represents the scope that document m distributes to the feature word set of topic k.

C. The Process of LDA Modeling

Prior to LDA modeling of document set, for the given document set D={d1,d2,…,dn}, the preprocessing is needed to each document dm (dm∈D), which mainly includes word segmentation, stop-word deletion, punctuation elimination and so on, and finally, saves the processed word items with space delimitation to obtain the corresponding corpus for next step of data processing.

In our work, Gibbs sampling arithmetic of MCMC method [6, 7] was adopted for parameter estimation, which can be regarded as an inverse process of document generation process. According to the diagrammatic map in Fig. 1, we can obtain the probability distribution of a document:

$$p(\omega|\alpha,\beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)\,p(\omega|z_n,\beta))d\theta \tag{4}$$

In training process, the topics of document collection and the samples of feature words are obtained through Gibbs sampling. After convergence of the algorithm, parameter estimation can be carried out with the final samples.

III. TOPIC RELEVANCY CALCULATION BASED ON LDA AND WORD2VEC

The expression of LDA model on document adopts the form of probability to extract topic and feature word corresponding to the topic, therefore there exists some uncertainty. In order to express the semantic information of word item in document more accurately, word2vec method was introduced to express the semantic information between words, and the calculation on similarity between the word item and topic feature word given by LDA modeling was carried out, which provides the final topic relevancy.

A. Word2vec

For calculating word vector, Mikolov et al put forward word2vec language model [8,9], which reduces the processing of document content to the operation of K-dimensional vector by using contextual information of a word, and expresses the similarity of document semantics with the similarity in vector space [10]. Word2vec contains two kinds of training model, CBOW and Skip-Gram, both of them contain input layer, projection layer and output layer. Among them, CBOW model predicts the word vector of the current word with context, and selects target vector of training as the sum of the context word vectors, while Skip-Gram model predicts the context through current word. Through the two models, word2vec can consider contextual information fully.

B. Topic Relevancy Calculation

At first, the semantic information of word items can be obtained by carrying out training on the document set with word2vec. After word2vec training, the semantic similarity between two words can be expressed as the cosine of the angle between two vectors. The word vector information

obtained after training is stored in the file for subsequent calculation.

Based on above information, for every term wj of every document, word2vec is used to calculate the cosine similarity between the word item and feature word wn under topic ti. The relevancy between the word item wj and the topic ti, denoted as S (wj,ti), is the probability weighted sum of the cosine similarities between wj and each feature words under the topic ti :

$$S(w_j,t_i)=\sum_{n=1}^{N}P(w_n|t_i)\times\cos(w_j,w_n) \tag{5}$$

Thus the relevancy between word item wj and document dm, denoted as S (wj, dm), is defined as the probability weighted sum of the relevancy between wj and the each topics of dm :

$$S(w_j,d_m)=\sum_{i=1}^{K}P(t_i|d_m)\times S(w_j,t_i) \tag{6}$$

At last, we accumulate the S (wj,dm) of each word item in document to get the total relevancy of the m-th document Sm:

$$S_m=\sum_{j=1}^{J}S(w_j,d_m) \tag{7}$$

IV. Off-topic Detection Algorithm

The specific steps of off-topic detection algorithm are designed as follows:

- Step 1: Preprocessing. Segment the words in the English document according to space delimiter, and transfer the capital letters into lowercase. Remove stop-words and punctuation, extract the stem of each word.
- Step 2: Establishing document-term matrix for the document set preprocessed.
- Step 3: LDA modeling. Carryout LDA modeling for each document in above document-term matrix to obtain the value of θm, the topic probability distribution of the m documents, and the value of φk, the probability distribution of feature word under the k-th topic, and sort the probability value in descending order, to obtain the topics of documents and their probability distribution, feature words and their probability distribution.
- Step 4: Training word vector with word2vec. Taking the preprocessed document set as input, train using word2vec to obtain the output words vectors. Using generated word vectors, calculate the similarity between words by cosine similarity, and change the training results, which expresses the semantic information of word items in document, to vector information and save them.
- Step 5: Calculating topic relevancy. For each term of document set, use word2vec to calculate the cosine similarity between it and all feature words under the i-th topic generated by LDA modeling. Through (5), calculate the probability weighted sum of each feature word, then calculate the probability weighted sum of all topics using (6). Finally, accumulate the topic similarities got from each word item using (7), and filtrate off-topic composition according to predetermined threshold.

Above algorithm combined the advantage of LDA and word2vec, expressed semantic relationships between words in a document more accurately after word2vec training, and after LDA modeling, it can be more effectively judged whether the topics of document itself keep to the point. We can get the topic similarity of document in a lower dimensional semantic space, and off-topic document can be effectively detected with the similarity.

V. Experimental Results and Analysis

We collected 1230 college English compositions of 6 different themes, including 205 articles for each theme. Each composition has been artificially scored, 15 points full, and under each topic there existed a certain number of off-topic compositions. If the artificial score is less than 5 points, we regard it as off-topic composition. Comparison was made between the off-topic document detected in experiment and the off-topic document identified by artificial scoring result, and comprehensive evaluation and analysis was made on base of accuracy, recall, and F-measure, to verify the effectiveness and practicability of the proposed algorithm.

In experiment, Gibbs sampling was adopted for LDA model. During document topic modeling, we assumed that the number of topics K is 2; hyper parameter α took the empirical value 50/K [11], which varies with the quantity of the topic; hyper parameter β takes fixed empirical value 0.01[12]. In order to make sure the accuracy, the sampling iteration time was set to 1000 times.

When word2vec trains document set, on base of different parameters and their meanings [13], the result of parameter settings were as shown in "Table I".

TABLE I. Word2vec Parameter Settings

Hyper parameter	Parameter specification	Value
size	the dimension of the word vector	50
window	The size of context window	5
min-count	the minimum threshold of a word	1
CBOW	Whether use CBOW model (0 for using)	1

The accuracy, recall, and F-measure of off-topic detecting results for 6 themes are as shown in "Table II".

TABLE II. Test Results when the Number of Topic is 2

	Theme1	Theme2	Theme3	Theme4	Theme5	Theme6	Average value
Accuracy	94.74%	93.33%	93.75%	86.67%	61.54%	75%	84.17%
Recall	94.74%	100%	100%	86.67%	80%	75%	89.40%
F-measure	94.74%	96.55%	96.77%	86.67%	69.57%	75%	86.55%

In order to optimize the effect of the off-topic detecting, we obtained the variation trend of F-measure according to topic number and through changing the quantity of document topic, determined the best quantity of topic K in the LDA model, and obtained the final experiment results according to determined optimal K. The result of mean F-measure with different quantity of topic is as shown in "Fig. 2".

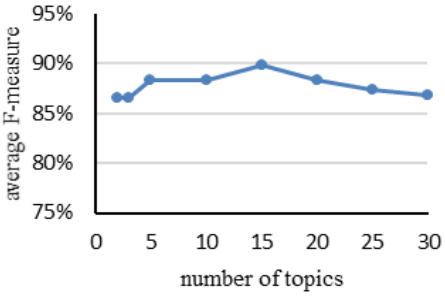

Fig. 2. Average F-measure for different topics.

It can be found in "Fig. 2" that the average F-measure is reached the highest value for the quantity of topic K=15, which is determined optimal topic quantity. Meanwhile it was found that, as the quantity of topic increases, the iteration time of the experiment will also increase. When the K value is changed, the value of hyper parameter α changes with a law of inverse proportion, i.e. the larger the value of K is, the smaller the value of α, which indicates each composition contains more topics. As regards feature words under each topic, [14] has verified that the optimal effect will be obtained when 5 feature words are selected. Therefore, we selected 5 feature words under each topic of each document in the experiment.

Comparing between off-topic document detected by proposed algorithm and off-topic document determined by artificial scoring, it was finally obtained that the average accuracy of off-topic detection under 6 different themes is 91.86%, the average recall is 88.78%, and average F-measure is 89.81%.

Comparative experiment has been made between the results in the paper and TF-IDF algorithm based on VSM using the same corpus. Cosine similarities between the compositions to be tested and 5 model compositions were calculated separately. Then we carried out mean processing for the similarity results, and screened out off-topic document corresponding to the threshold. The experimental result was evaluated with F-measure, and through 6 groups of experiments, the mean F-measure of off-topic detection is reached 77.4%, which is lower than that of proposed method. The off-topic compositions detected by proposed algorithm can be more than 88%, and the accuracy is also comparatively high. Meanwhile, the algorithm is more effective than TF-IDF algorithm in vector space model, and can screen out the off-topic compositions effectively within short time, so that save much time for teachers going over examination papers.

VI. CONCLUSION

In this paper, LDA was adapted to model document, and word2vec was used for training. LDA and word2vec were employed to carry out topic relevancy evaluation for documents. The experimental result indicates that the proposed algorithm can effectively detect the off-topic compositions. The algorithm possesses intelligent auxiliary evaluation ability for English teaching, including the paper inspection of English competition. With computer, the algorithm can help teachers filter off-topic compositions quickly, objectively, fairly and automatically, can reduce the influence of subjective factors in teachers' reading, and thus can improve the efficiency of paper inspection.

The proposed method only uses F-measure as reference when it uses LDA modeling to determine the quantity of topic, while it doesn't research more other calculation method for determining the quantity of topic. The next step of our work is to continue research, and improve the method of document modeling and determining topic quantity.

REFERENCES

[1] P. Deane, On the relation between automated essay scoring and modern views of the writing construct, Assessing Writing. 18.1 (2013) 7-24.

[2] Z. Zhi-fei, M. Duo-qian, G. Can, Short text classification using latent Dirichlet allocation, Journal of Computer Applications. 33.6 (2013) 1587-1590.

[3] W. Zhen-zhen, H. Ming, D. Yong-ping, Text Similarity Computing Based on Topic Model LDA, Computer Science. 40.12 (2013) 229-232.

[4] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, M. Zhu, A Practical Algorithm for Topic Modeling with Provable Guarantees, International Conference on Machine Learning. (2013) 280-288.

[5] K. Farrahi, D. Gaticaperez, Discovering routines from large-scale human locations using probabilistic topic models, Acm Transactions on Intelligent Systems and Technology. 2.1 (2011) 1-27.

[6] W.A. Link, M.J. Eaton, On thinning of chains in MCMC, Methods in Ecology and Evolution. 3.1 (2012) 112-115.

[7] M. Hai-yun, A producing test case technology research based on Gibbs sampling, Automation and Instrumentation. 2 (2011) 11+14.

[8] T. Ming, Z. Lei, Z. Xian-chun, Document Vector Representation Based on Word2Vec, Computer Science. 43.6 (2016) 214-217+269.

[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, Computer Science. (2013).

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, Advances in neural information processing systems. (2013) 3111-3119.

[11] W. Peng, G. Cheng, C. Xiao-mei, Research on LDA Model Based on Text Clustering, Information Science. 33.1 (2015) 63-68.

[12] H. Ji-ming，C. Guo, Mining and Evolution of Content Topics Based on Dynamic LDA, Library and Information Service. 58.2 (2014) 138-142.

[13] Z. Lian, Exploration of the Working Principle and Application of Word2vec, Sci-Tech Information Development & Economy. 25.2 (2015) 145-148.

[14] W. Kai, W. Ying, The Initial Exploration on Microblogger Knowledge Discovery with User Mention Relations, Library and Information. 2 (2015) 123-127.