A text retrieval algorithm based on the hybrid LDA and Word2Vec model

Xue Mu
Liaoning Jianzhu Vocational College；Liaoyang 111000,China
NaLu132456@163.com

Abstract—**Text retrieval is a fundamental problem in information retrieval, and it is greatly influenced by the text feature representation. In this paper, we propose a novel text feature representation model, which integrates the LDA and Word2Vec model, and then utilize this model to handle the text retrieval problem. Latent Dirichlet Allocation represents a generative probabilistic model of a corpus, and texts are organized as random mixtures of latent topics. In particular, the proposed algorithm computes the distance between document and topics, and then each document is represented as a feature vector, in which each dimension denotes the distance between this document and a specific topic. To test the effectiveness of the proposed algorithm, several related methods are made performance comparison, and experimental results demonstrate that the proposed solution performs better than other methods, and it can achieve high accuracy for text retrieval.**

Keywords- Text retrieval, Latent Dirichlet Allocation, Word2Vec, Feature space

I. INTRODUCTION

In recent years, most information on Internet is represented as texts. The rapid growth of text information provides a great challenge to information retrieval, and then let it increasingly difficult to seek useful information on Internet with high accuracy [1][2]. For famous information retrieval algorithms, the keywords matching are utilized, that is to say, keywords matching denote the explicit representation. However, it is difficult to find what we need in text database, due to the uncertainty in natural languages, such as synonym and polysemy [3]. Moreover, it is also difficult for us to represent what they really want to retrieve just with keywords [4].

Furthermore, with the rapid development of search engine, people are able to get the information they require via the Internet at anytime and anywhere. But, web data are growing fast and we have higher requirements for information retrieval technology [5][6]. It is of great importance to make the search engine more intelligent. If we illustrate the relationship between query terms and documents, existing information retrieval systems only consider the word literally matching without using the correlation information of word semantic expression. Hence, there is a gap between search results and user requires [7][8].

Based on the above analysis, information retrieval is designed to recognize and acquire information from the set of information, and it can play a key role in information technology [9]. Furthermore, information retrieval has been an efficient approach for us to develop and exploit all sorts of information resources effectively [10].

II. OVERVIEW OF THE LDA TOPIC MODEL

Latent Dirichlet Allocation (denoted as LDA) means a generative probabilistic model of a corpus [11]. In the LDA model, texts are organized as random mixtures of latent topics, and each topic refers to a word distribution [12][13]. The initial structure of the LDA model is listed in Fig. 1.

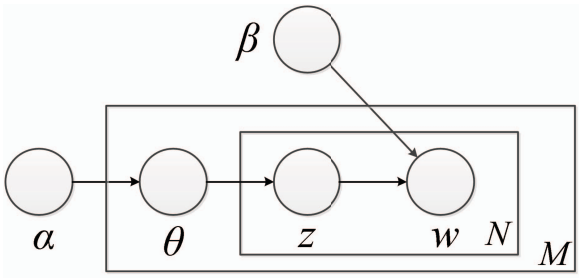

Figure 1. Initial structure of the LDA model.

From Fig. 1, we can see that LDA is represented as a probabilistic graphical model with three levels, among which the outer plate represents documents, and the inner plate denotes the repeated choice of topics and words in a text [14][15]. Texts are associated with multiple topics in LDA model, and the generative process for each document w in a corpus D is defined as follows.

1) Select $N \sim Poisson(\xi)$
2) Select $\theta \sim Dir(\alpha)$
3) For each word w_n, $n \in [1, N]$
   i) Select a topic $z_n \sim Mul(\theta)$
   ii) Select a word w_n from $p(w_n | z_n, \beta)$, which means a multinomial probability conditioned on the topic z_n.

Particularly, we provide several assumptions for the LDA model, that is, a) the dimensionality k of the Dirichlet

978-1-7281-1307-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICITBS.2019.00098

distribution is fixed, and b) the word probabilities are parameterized by a $k \times V$ matrix β, where $\beta_{ij} = p\left(w^j = 1 \middle| z^j = 1\right)$. Furthermore, N is not dependent of all the other data generating variables, such as θ and z.

For parameters α and β, the joint distribution of topics z, word w, topic mixture θ is defined in Eq. 1 [16].

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (1)$$

Based on the above definition, the probability of a corpus is obtained utilizing the product of the marginal probabilities of single document [17], which is shown in Eq.2.

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn}, \beta)\right) d\theta_d \quad (2)$$

III. THE PROPOSED ALGORITHM

In this section, we will demonstrate the hybrid LDA and Word2Vec model, which is able to compute the text similarity.

Word2vec denotes a set of related models which are exploited to generate word embeddings. These models are shallow, 2-layer neural networks which are trained to build linguistic contexts of words. In particular, a large corpus of texts is input to the Word2vec model, and then constructs a vector space with several hundred dimensions. In addition, each unique word in the corpus is allocated to a corresponding vector in the space. Word vectors are located in the vector space, hence, words that share common contexts in the corpus are positioned in close proximity to one another in the space

We utilize the continuous bag of words (denoted as CBOW) to illustrate how the Word2Vec model works, and CBOW is used to forecast words by exploiting the contexts of surrounding texts.

Suppose that there is a sentence $S = \{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\} \in \mathbb{R}^m$, where w_t refers to the target term. Then, the input layer is defined as follows.

$$c(v(w_t)) = \{v(w_{t-2}), v(w_{t-1}), v(w_{t+1}), v(w_{t+2})\} \in \mathbb{R}^m \quad (3)$$

where $c(v(w_t))$ refers to the context of the term $v(w_t)$. Next, the projecting layer is used to construct a contextual vector $v(w_t)$ as follows.

$$v(w_t) = \sum_{i=t-2}^{t+2} c(v(w_i)) \quad (4)$$

Then, in the output layer, the word in dictionary is regarded as a leaf node in terms of the occurrence in the corpus

For a set of texts $X = \{x_1, x_2, \cdots, x_n\}$, of which the dictionary is made up of N terms $\{w_1, w_2, \cdots, w_n\}$. After training the texts dataset X, LDA model is able to generate topics set $\{t_1, t_2, \cdots, t_T\}$. For the topic t_i, its topic vector $v(t_i)$ is computed as follows.

$$v(t_i) = \sum_{n=1}^{h} \lambda_{in} v(w_{in}) \quad (5)$$

where $\lambda_i = \gamma_i \Big/ \sum_{n=1}^{h} \gamma_n$, and γ_{ij} means the j^{th} word in the topic t_i

Then, the feature vector of each text can be computed as follows.

$$v(x_i) = \frac{1}{k}\sum_{n=1}^{k} v(w_{in}) \quad (6)$$

Afterwards, each text is able to be described as a distance distribution from a text to all different topics in a same space. The distance between a text and a topic is calculated as follows.

$$Dis(v(x_i), v(t_j)) = |v(x_i) - v(t_j)| \quad (7)$$

From the above, we can see that the proposed algorithm can estimated the distance between document and topics, and then each document is able to be described by a vector, in which each dimension refers to the distance between this document and a specific topic.

IV. EXPERIMENT

To test the performance of the proposed algorithm, in this section, we choose the 20Newsgroups dataset to make performance evaluation. 20Newsgroups contains 18846 newsgroup documents collected by Ken Lang, which is organized into 20 various newsgroups. The 20 Newsgroups data set denotes a set of almost 20,000 newsgroup documents, partitioned evenly across 20 various newsgroups. In addition, this dataset is constructed by Dr. Ken Lang, probably for his Newsweeder: Learning to filter newspaper, though he does not explicitly mention this collection. Particularly, the 20 newsgroups collection has

been a popular data set for experiments in text retrieval and text classification. News classes in the 20Newsgroups dataset are shown in Table. 1.

Table. 1 News classes in the 20Newsgroups dataset.

ID	Class name	ID	Class name
1	comp.graphics	11	sci.electronics
2	comp.os.ms-windows.misc	12	sci.med
3	comp.sys.ibm.pc.hardware	13	sci.space
4	comp.sys.mac.hardware	14	misc.forsale
5	comp.windows.x	15	talk.politics.misc
6	rec.autos	16	talk.politics.guns
7	rec.motorcycles	17	talk.politics.mideast
8	rec.sport.baseball	18	talk.religion.misc
9	rec.sport.hockey	19	alt.atheism
10	sci.crypt	20	soc.religion.christian

Firstly, we should test how many topics is the most suitable for the proposed algorithm. Then, we test the F1 score of the proposed algorithm with different number of topics, and the results are shown in Fig. 2.

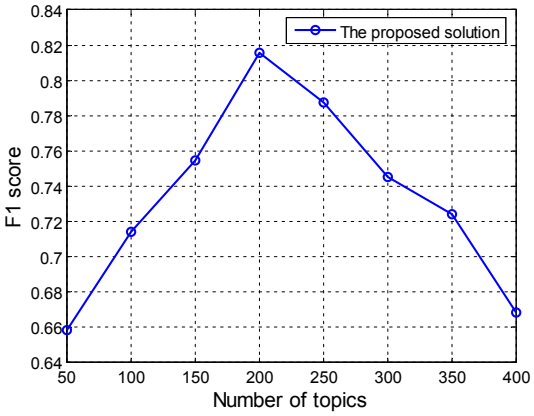

Figure 2. F1 score with the number of topics varying.

From Fig. 2, we can see that for different number of topics, the proposed algorithm achieve its highest value when the number of topics is set to 200. Therefore, in the following part, we set the number of topics to 200.

To demonstrate the effectiveness of the proposed hybrid LDA and Word2Vec model, we select other feature extraction methods: 1) TF-IDF, 2) LDA, and 3) Word2Vec. Firstly, we show the F1 value for each news classes

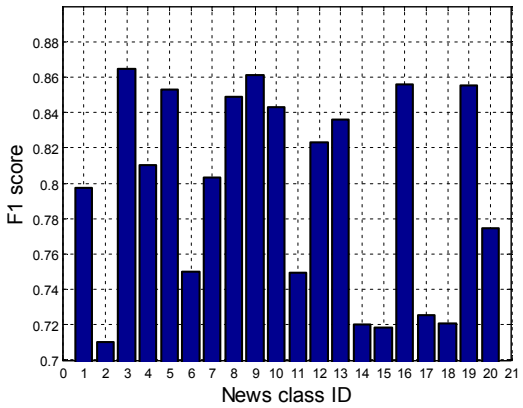

Figure 3. F1 score for each news class

Afterwards, we compare our proposed solution with other types of document features, and experimental results are shown in Table. 2, and the number of topic is set to 200.

Table. 2 Performance comparison for different methods.

Method	F1 score
TF-IDF	0.652
LDA	0.729
Word2Vec	0.803
The proposed solution	0.814

Integrating all the experimental results above, we can see that the proposed solution performs better than other methods, and it can achieve high performance in the task of text retrieval.

V. CONCLUSION

This paper proposes a new text feature representation model, which combines the LDA and Word2Vec model together, and then uses this model to solve the text retrieval problem. In particular, the hybrid LDA/Word2Vec model to calculate the distance between document and topics, and then each document is represented as a feature vector. In addition, each dimension represents the distance between this document and a specific topic. In the end, experimental results demonstrate the effectiveness of the proposed algorithm.

REFERENCE

[1] S. Amir, A. Tanasescu and D. A. Zighed, Sentence similarity based on semantic kernels for intelligent text retrieval,

Journal of Intelligent Information Systems, 2017, 48(3): 675-689

[2] D. Datta, S. K. Singh and C. R. Chowdary, Bridging the gap: effect of text query reformulation in multimodal retrieval, Multimedia Tools and Applications, 2017, 76(21): 22871-22888

[3] K. Ghosh, A. Chakraborty, S. K. Parui and P. Majumder, Improving Information Retrieval Performance on OCRed Text in the Absence of Clean Text Ground Truth, Information Processing & Management, 2016, 52(5): 873-884

[4] S. Hakak, A. Kamsin, P. Shivakumara and M. Y. I. Idris, PARTITION-BASED PATTERN MATCHING APPROACH FOR EFFICIENT RETRIEVAL OF ARABIC TEXT, Malaysian Journal of Computer Science, 2018, 31(3): 200-209

[5] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal and A. Valencia, Information Retrieval and Text Mining Technologies for Chemistry, Chemical Reviews, 2017, 117(12): 7673-7761

[6] M. T. Lechuga, J. M. Ortega-Tudela and C. J. Gomez-Ariza, Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts, Learning and Instruction, 2015, 40: 61-68

[7] K. Nguyen and M. A. McDaniel, The JOIs of Text Comprehension: Supplementing Retrieval Practice to Enhance Inference Performance, Journal of Experimental Psychology-Applied, 2016, 22(1): 59-71

[8] R. S. Renu and G. Mocko, Computing similarity of text-based assembly processes for knowledge retrieval and reuse, Journal of Manufacturing Systems, 2016, 39: 101-110

[9] P. P. Roy, A. K. Bhunia and U. Pal, Date-field retrieval in scene image and video frames using text enhancement and shape coding, Neurocomputing, 2018, 274: 37-49

[10] W. Song, B. Wang, Q. Wang, Z. Y. Peng, W. J. Lou and Y. H. Cui, A privacy-preserved full-text retrieval algorithm over encrypted data for cloud storage applications, Journal of Parallel and Distributed Computing, 2017, 99: 14-27

[11] D. Backenroth, Z. H. He, K. Kiryluk, V. Boeva, L. Pethukova, E. Khurana, A. Christiano, J. D. Buxbaum and I. Ionita-Laza, FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications, American Journal of Human Genetics, 2018, 102(5): 920-942

[12] C. Chen, A. Zare, H. N. Trinh, G. O. Omotara, J. T. Cobb and T. A. Lagaunne, Partial Membership Latent Dirichlet Allocation for Soft Image Segmentation, Ieee Transactions on Image Processing, 2017, 26(12): 5590-5602

[13] S. Lee, S. Kim, S. Lee, J. Choi, H. Yoon, D. Lee and J. R. Lee, LARGen: Automatic Signature Generation for Malwares Using Latent Dirichlet Allocation, Ieee Transactions on Dependable and Secure Computing, 2018, 15(5): 771-783

[14] X. X. Li, Z. Y. Ma, P. Peng, X. W. Guo, F. Y. Huang, X. J. Wang and J. Guo, Supervised latent Dirichlet allocation with a mixture of sparse softmax, Neurocomputing, 2018, 312: 324-335

[15] Z. J. Li, H. J. Zhang, S. Z. Wang, F. R. Huang, Z. P. Li and J. S. Zhou, Exploit latent Dirichlet allocation for collaborative filtering, Frontiers of Computer Science, 2018, 12(3): 571-581

[16] S. Momtazi, Unsupervised Latent Dirichlet Allocation for supervised question classification, Information Processing & Management, 2018, 54(3): 380-393

[17] Y. K. Tang, X. L. Mao and H. Y. Huang, Labeled Phrase Latent Dirichlet Allocation and its online learning algorithm, Data Mining and Knowledge Discovery, 2018, 32(4): 885-912