Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence

Minglai Shao
School of Computer, Electronics and Information
Guangxi University
Nanning, China
E-mail: shml114@sina.com

Liangxi Qin
School of Computer, Electronics and Information
Guangxi University
Nanning, China
E-mail: qin_lx@126.com

***Abstract*—LDA (Latent Dirichlet Allocation) topic model has been widely applied to text clustering owing to its efficient dimension reduction. The prevalent method is to model text set through LDA topic model, to make inference by Gibbs sampling, and to calculate text similarity with JS (Jensen-Shannon) distance. However, JS distance cannot distinguish semantic associations among text topics. For this defect, a new text similarity computing algorithm based on hidden topics model and word co-occurrence analysis is introduced. Tests are carried out to verify the clustering effect of this improved computing algorithm. Results show that this method can effectively improve text similarity computing result and text clustering accuracy.**

Keywords-topic model; LDA (Latent Dirichlet Allocation); JS (Jensen-Shannon) distance; word co-occurrence; similarity

I. INTRODUCTION

With the rapid development of internet, the amount of information on the internet increases exponentially. How to discover useful information efficiently from the magnanimous text data (one of the main carrier of information) becomes a crying need.

Vector space model (VSM), a classic mode in the text mining area, represents the documents as space vector and computes the similarity among the vectors to measure the similarity among the documents. Hereinto, the TF-IDF (Term frequency-inverse document frequency) is the most widely applied similarity measure method. By this method, word-weighting is expressed by the frequency of a particular word in a particular document and by the inverse frequency of this word in the document set. However, this method ignores the semantic associations among words, leaving it difficult to process the semantic factors. For example, there are no common words between "Steve Jobs left us." and "will the price of Apple products drop?", yet there is certain correlation between them. For another example, when the word appears in two articles describing respectively a fruit and a cell phone brand, the two "apple" are considered as correlated. What's more, this method also bears problems regarding the high-dimensional sparse of data space.

In solving these problems, modeling the text set through LDA topic model and computing the similarity of the text with the JS (Jensen-Shannon) distance have made preferable clustering results. However, JS distance cannot distinguish semantic association among text topics. This may leads to wrong clustering of texts that have similar topic probability yet different topics. For this defect, we introduce the idea of word co-occurrence to analyze the semantic correlation of text themes, since co-occurrence words embody the text topic better. It provides an improved text similarity measure method based on hidden topic model and word co-occurrence analysis.

II. RELATED WORKS

A. Text hidden topic model

Text topics mining have received wide attention and have been extensively applied to text clustering in recent years since topic model can reduce dimensions efficiently and is interpretable.

Currently available hidden topic model includes LSA (Latent Semantic Analysis)[1], PLSA (Probabilistic Latent Semantic Analysis)[2] and LDA, etc.

LSA applies SVD (Singular Value Decomposition) and other mathematical method to discover hidden semantic structures of documents. Its limitation lies in its disability to distinguish polysemy in the documents.

PLSA is a probabilistic model presented by Hofmann on the foundation of LSA. Basing its work on the production model and maximum likelihood estimation method, this model gets results by EM (Expectation Maximization) algorithm. Thus PLSA is prior to LSA in dealing with large-scale date sets.

LDA introduces Dirichlet prior parameters to word layer and hidden topic layer in modeling, which is a ground-breaking extension of PLSA. It solves the problem of overfitting generated by concomitant linear increase of topic parameters at the increase of training documents in PLSI model and LSI model, making it more suitable for large-scale corpus processing.

Shi Jian-hong[3] et al. applied LDA topic model to Chinese micro blog topic and carried out effective micro blog topic discoveries. Li Wen-bo[4] et al. raised a labeled LDA topic model by adding text class information to the LDA topic model, which calculated the distribution of hidden topics in each class and raised the classification ability of the traditional LDA model. Phan[5] et al. adopted hidden topic in text character extension based on the external corpus. Shi Jing[6] et al. achieved a preferable extraction effect by using Shannon information to extract key words in LDA probability distribution. Quan[7] et al. used topic as dependency of third-party words and further mined text similarity. Zhang Zhi-fei[8] et al. raised a text

classifying method based on LDA topic model and an overall consideration of context.

B. *Word co-occurrence*

Word co-occurrence analysis is a successful use of natural language processing in information retrieval. Its core concept is that the co-occurrence rate of words can to some extent reflect the semantic correlation of them.

Word co-occurrence analysis is being increasingly applied to text analysis. Geng Huan-tong[9] et al. raised the topic word extraction algorithm based on word co-occurrence, expanding extraction scope of the original topic word by mining the co-occurrence word of candidate words. Chang Peng [10] did a deep analysis of the inner link between text topic representation and word co-occurrence and designed a new method of co-occurrence word extraction. He also raised a new document representation model on this basis. Yuan Li-chi[11] proposed to measure word similarity based on the Mutual information. This method effectively eliminated word indeterminacy.

III. APPLICATION OF LDA TOPIC MODEL IN TEXT REPRESENTATION

A. *LDA topic model*

LDA (Latent Dirichlet Allocation) topic model is a three-layer Bayesian probability model composed of word, topic and text. Its basic idea is that every document is a mixture of several hidden topics and each hidden topic is a mixture of several words. The relation between document and topic follows Dirichlet prior distribution and the relation between topic and word follows polynomial distribution. The generative process of LDA is as shown in Figure (1):

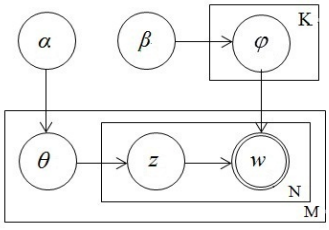

Figure 1. LDA generation probability diagram

Among the variables, M denotes the number of documents, K denotes the number of hidden topic, N denotes the number of words in a document. α、β are the document layer parameters of LDA, α denotes the relative strength of latent hidden topics in the document set and β denotes the probability distribution of all hidden topics. θ denotes the topic probability distribution for certain document. φ denotes the word distribution for certain hidden topic. Rectangle denotes repeated sampling process, unilateral circle denotes hidden variables. Bicircle denotes observable variables. The computing formula of probability model is as shown in formula (1):

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

The generative process of LDA topic model is as follows:

1) For hidden topic i, calculate φ polynomial distribution of feature word of its hidden topic according to Dirichlet distribution;

2) Obtain the number of words N in the document according to Poisson distribution;

3) Calculate the topic probability distribution θ for each text;

4) For each feature word of each document of each document set:

a) Select a hidden topic z randomly from the topic probability distribution θ;

b) Select a feature word randomly from the polynomial distribution of topic z.

B. *Gibbs sampling*

Parameter estimation is needed in LDA modeling. Here Gibbs sampling is used. It is easy to understand, easy to realize and can effectively select topics from large-scale documents. The main idea of computing is that, for a certain feature word w_i , use Gibbs sampling to extract the approximation of the posterior distribution $p(z_i = j \mid z_{-i}, w_i)$ of word from a hidden topic z_i . The computing formula is as shown in formula (2):

$$p(z_i = j \mid z_{-i}, w_i) \propto \frac{n_{-i,j}^{(w)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \bullet \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\bullet}^{(d_i)} + T\alpha} \quad (2)$$

Among the variables, $n_{-i,j}^{(w)}$ denotes the number of word tokens of feature word w_i assigned to the hidden topic j. $n_{-i,j}^{(\bullet)}$ denotes the number of word tokens assigned to the hidden topic j. $n_{-i,j}^{(d_i)}$ denotes the scale of feature words in document d_i that are assigned to hidden topic j. $n_{-i,\bullet}^{(d_i)}$ denotes the number of feature words in document d_i that are assigned to hidden topic j. T and W denote nonnegative weighting. In iteration extracting process, parameter θ and φ is estimated separately according to formula (3) and formula (4).

$$\hat{\theta}^{(d)} = \frac{n_j^{d_i} + \alpha}{n_\bullet^{d_i} + T\alpha} \quad (3)$$

$$\hat{\varphi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\bullet)} + W\alpha} \quad (4)$$

IV. IMPROVED TEXT SIMILARITY COMPUTING

A. JS distance

Since the topic distribution of a text is a simple mapping of text space, similarity of two texts can be measured by computing the topic distribution. KL distance is the measurement of "difference" between two probabilities. Some people have used KL distance as the criterion of similarity computing. Let p(x) and q(x) be two probability density functions, the KL distance between this two can be defined as shown in formula (5):

$$D_{KL}(p,q)=\sum_{i=1}^{T} p_i \ln \frac{p_i}{q_i} \tag{5}$$

However, $D_{KL}(p,q)\neq D_{KL}(q,p)$ means the KL distance is asymmetrical. So here its symmetrical version is used as shown in formula (6):

$$\begin{aligned} D_{\lambda}(p,q)=&\lambda D_{KL}(p,\lambda p+(1-\lambda)q)\\ &+(1-\lambda)D_{KL}(q,\lambda p+(1-\lambda)q) \end{aligned} \tag{6}$$

When λ=1/2, the above formula turns into JS distance. Assigning the value [0, 1] to it, the results is as shown in formula (7):

$$D_{js}(p,q)=\frac{1}{2}[D_{KL}(p,\frac{p+q}{2})+D_{KL}(q,\frac{p+q}{2})] \tag{7}$$

B. Improved text similarity computing based on word co-occurrence

JS Distance can't distinguish the semantic relation between topics when it is used to carry out similarity computing. For this defect, an improved similarity computing method is proposed, which analyzes the semantic correlation between topics from a word co-occurrence angle and adds a semantic correlation computing of topic feature words to the original JS measuring method. Details are as follows:

Assume T_i is the topic of text D_i, word set $W=\{w_{i1},w_{i2},\cdots\cdots w_{iN}\}$ is the feature word of topic T_i. According to co-occurrence formula (8) (as follows), the co-occurrence probability of feature word is $p_{11},p_{12},p_{13}\cdots p_{NN}$.

$$p(w_{im},w_{in})=p(w_{im}\mid T_i)p(w_{in}\mid T_i) \tag{8}$$

After the computing of co-occurrence probability of topic feature word, here follows the discussion of the semantic correlation between feature words from topic T_i and topic T_j. If the probability of feature word w_{im} in topic T_i is p_{im}, the co-occurrence probability of feature word w_{im} and w_{jn} in topic T_i is p_{mn} (p_{mn} can be obtained from formula (8), then the similarity computing formula of w_{im} and w_{jn} is as shown in formula (9) :

$$correlation(w_{im},w_{jn})=\frac{p_{mn}}{p_{im}+p_{jn}-p_{mn}} \tag{9}$$

According to formula (9), when the value of p_{mn} is 0, $correlation(w_{im},w_{jn})=0$, which means feature word w_{im} and w_{jn} is uncorrelated. When $correlation(w_{im},w_{jn})\neq 0$, feature word w_{im} and w_{jn} is correlated.

When taking a comprehensive consideration from the angle of probability distribution of hidden topic and from the angle of feature word co-occurrence of hidden topic, it is known that when the similarity degree of hidden topic probability distribution is high and the topic feature word is correlated, the text similarity degree is the highest and these documents should be placed in one category. When the similarity degree of hidden topic probability distribution is low and the topic feature word is not correlated, the text similarity degree is the lowest and these documents should not be placed in one category. When the similarity degree of hidden topic probability distribution is high and the correlation degree of topic feature word is low, similarity between texts should be reduced. When the similarity degree of hidden topic probability distribution is low and the correlation degree of topic feature word is high, similarity between texts should be enhanced.

To sum up, a new text similarity computing method is proposed, which is shown in formula (10):

$$\begin{aligned} Similarity(d_i,d_j)=&\lambda D_{js}(d_i,d_j)\\ &+(1-\lambda)\left[\sum_{m,n=1}^{V}(1-correlation(w_{im},w_{jn}))\Big/(V(V-1))\right] \end{aligned} \tag{10}$$

Among the variables, d_i and d_j denote arbitrary texts from the document set, w_{im} and w_{jn} denotes the feature word of d_i and d_j separately, V denotes the number of feature word of this selected document. $\lambda\in[0,1]$ denotes a correlation coefficient assigned to this document. The smaller the value of $Similarity(d_i,d_j)$ is, the more similar the two texts d_i and d_j are.

Detailed steps of this improved computing method are as follows:

Computing method: Improved text similarity computing method.

Input: arbitrary text d_i and d_j, probability distribution φ and θ;

Output: similarity between d_i and d_j: $Similarity(d_i,d_j)$

Step 1: extract the first N letters of highest document probability distribution as the feature word of text d_i and d_j,

based on the distribution of word in probability distribution φ and θ;

Step 2: extract feature word based on formula (8) and Step 1, calculate the co-occurrence probability of text feature word;

Step 3: calculate the correlation between arbitrary feature words based on formula (9);

Step 4: calculate similarity between d_i and d_j, which is $Similarity(d_i, d_j)$, based on formula (7) and results from Step 3.

V. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

A. Evaluation criterion

This paper measures text similarity and clustering effect with a clustering analysis of text, adopting F Metric, Precision Ratio and Recall ratio. F Metric is a balance index for information retrieval combining Precision Ratio and Recall ratio.

Precision Ratio $P(i, j)$, Recall ratio $R(i, j)$ and F Metric $F(i, j)$ are defined respectively in formula (11), (12) and (13):

$$P(i, j) = \frac{N_{ij}}{N_i} \tag{11}$$

$$R(i, j) = \frac{N_{ij}}{N_j} \tag{12}$$

$$F(i, j) = \frac{2P(i, j) * R(i, j)}{P(i, j) + R(i, j)} \tag{13}$$

Among the variables, N_{ij} denotes the number of text from category i in cluster j. N_i denotes the number of text from category i. N_j denotes the number of text from cluster j.

B. Corpus choice

This method is tested in the Chinese Corpus of Fudan University. In the experiment, three subsets were extracted and were named as C3-Art、C7-History、C19-Computer. From each subset, 400 pieces of text were extracted, with a total number of 1200.

C. Experimental procedure and main parameters selection

1) Preprocessing of document: mainly includes word segmentation and the elimination of Stop words, etc. Word segmentation is carried out with the ICTCLAS system developed by Institute of Computing Technology in the Chinese Academy of Sciences.

2) Document modeling: Document modeling: model the document by LDA topic model and do model solve and efficiency analysis by Gibbs sampling algorithm. In the experiment, assign value to α and β according to Document [12]. Let α be 50/K, β be 0.01, T be 100, which generates the best effect. Do the iteration for 1000 times to get the probability distribution matrix θ of document-topic and probability distribution matrix φ of topic-word.

3) Document similarity computing: measure document similarity by the similarity computing method mentioned in 4.2. Let λ be 0.1、0.2…… 0.9. Repeated comparative testing and analyzing show that the result is the best when the value of λ equals 0.7. Thus λ is assigned to the value of 0.7.

4) Document clustering: carry out text clustering through hierarchical clustering algorithm and analyze the clustering result to evaluate the degree of accuracy of the computing..

D. Analyses of experimental results

This paper does a comparative analysis of the original "LDA+JS" computing method and "LDA+ JS+ Word co-occurrence" computing method proposed in this paper. Experimental results are as shown in TABLE I, Figure 2, Figure 3 and Figure 4. Results prove that the Accuracy rate and Recall ratio of "LDA+ JS+ Word co-occurrence" computing method proposed in this paper is higher.

This result owes to the fact that analyzing co-occurrence word as a whole can better represent the text topic. At the basis of adopting JS Distance in the measuring of text similarity, topic correlation analysis based on word co-occurrence is added, thus effectively solving problems concerning polysemy, synonym and context dependency, better representing text similarity and effectively reducing mis-clustering of texts that have similar topic probability yet different topics. Test results prove that the similarity computing method proposed in this paper is feasible.

TABLE I. EXPERIMENTAL RESULT

Category	LDA+JS			LDA+JS+ Word co-occurrence		
	Precision Ratio	***Recall Ratio***	***F Metric***	***Precision Ratio***	***Recall Ratio***	***F Metric***
Art	0.7613	0.7975	0.7789	0.7835	0.8050	0.7941
Computer	0.7298	0.7225	0.7252	0.7500	0.7537	0.7353
History	0.7818	0.7525	0.7669	0.7928	0.7750	0.7734

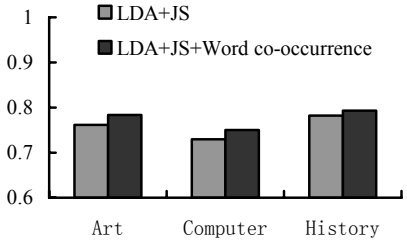

Figure 2. Precision Ratio

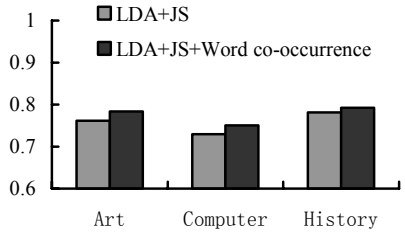

Figure 3. Recall Ratio

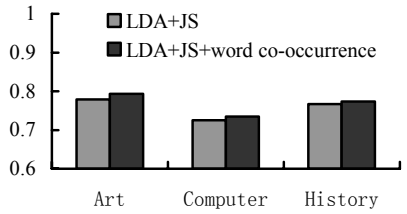

Figure 4. F Metric

VI. CONCLUSION

In this paper, we present a research into text similarity computing from the two angles of text hidden topic probability distribution differences and of semantic correlation of text feature words. Modeling documents by LDA hidden topic model greatly reduces text dimension and improves the computing efficiency. Analyzing semantic correlation of text feature word from a word co-occurrence angle based on LDA model enhances the use of text topic information and effectively improves text clustering result.

Since LDA topic model is highly expandable, follow-up works will be centered on new text modeling method and text similarity computing method. Ideas may include modeling text by replacing the single word in LDA model with co-occurrence word combination. This topic-based processing method has much significance for Data Mining and other disciplines.

REFERENCES

[1] Deerwester S, Dumais S, Landauer T. Indexing by latent semantic analysis [J]. Journal of the American Society of Information Science, 1990,41(6):391-407

[2] Hofmann T. Probabilistic latent semantic indexing[C] // Proc of the 22nd Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1999:50-57

[3] Shi Jian-hong, Chen Xing-shu, Wang Wen-xian. Discovering topic from microblog based on hidden topics analysis [J]. Application Research of Computers. 2014, 31(3):700-704

[4] Li Wen-bo, Sun Le, Zhang Da-kun. Text classification based on Labled-LDA model [J]. Chinese Journal of Computers, 2008, 31(4):620-627

[5] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large－scale data collections[C] In: Proceedings of the 17th International Conference on World Wide Web (WWW08). NewYork: ACM, 2008: 91-100

[6] Shi Jing, Li Wan-long. Topic words extraction method based on LDA model [J]. Computer Engineering, 2010,19(36): 81-83

[7] Quan X J, Liu G, Lu Z. Short text similarity based on probabilistic topics [J]. Knowledge Information System, 2010, 25(3):473-491

[8] Zhang Zhi-fei, Miao Duo-qian, Gao Can. Short text classification using latent Dirichlet allocation [J]. Journal of Computer Applications, 2013,33(6):1587-1590

[9] Geng Huan-tong, Cai Qing-Sheng, Yu Kun, Zhao Peng. A kind of automatic text key phrase extraction method based on word co-occurrence[J]. Journal of Nanjing University (Natural Sciences), 2006 ,42(2): 156-162

[10] Chang Peng. Research on terms co-occurrence based models and algorithms for Text Mining [D]. Tianjin: Tientsin University, 2009: 30-37

[11] Yuan Li-chi. A word clustering method based on mutual information [J]. Systems Engineering, 2008,26(5): 120-122

[12] Huang Bo. Research on microblog topic detection based on VSM model and LDA model[D]. Chengdu: Southwest Jiaotong University, 2012:36-40