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, 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 polysemic 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 groundbreaking extension of PLSA. It solves the problem of overfitting generated by concomitant linear increase of topic parameters at the increase of training documents in PLSA 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 \cite{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 \cite{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\cite{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. \( \alpha \) : \( \beta \) are the document layer parameters of LDA, \( \alpha \) denotes the relative strength of latent hidden topics in the document set and \( \beta \) denotes the probability distribution of all hidden topics. \( \theta \) denotes the topic probability distribution for certain document. \( \phi \) 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):

\[
p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{z=1}^{K} p(z_i | \theta) p(w_i | z_i, \beta)
\]  

The generative process of LDA topic model is as follows:

1) For hidden topic \( i \), calculate \( \phi \) 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 \( \theta \) 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 \( \theta \);
   b) Select a feature word randomly from the polynomial distribution of topic \( z \).
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):

\[
D_{KL}(p,q) = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i}
\]  

(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):

\[
D_{s}(p,q) = \lambda D_{KL}(p,\lambda p + (1-\lambda)q) + (1-\lambda)D_{KL}(q,\lambda p + (1-\lambda)q)
\]  

(6)

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

\[
D_{s}(p,q) = \frac{1}{2}[D_{KL}(p,\frac{p+q}{2}) + D_{KL}(q,\frac{p+q}{2})]
\]  

(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 \( W = \{w_1, w_2, \ldots, w_N\} \) is the feature word of topic \( T_i \). According to co-occurrence formula (8) (as follows), the co-occurrence probability of feature word is \( p_{1i}, p_{12}, \ldots, p_{1N} \).

\[
p_{imw}(w_m, w_n) = p(w_m|T_i)p(w_n|T_i)
\]  

(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 \( w_m \) in topic \( T_i \) is \( p_{im} \), the co-occurrence probability of feature word of text \( w_m \) and \( w_n \) in topic \( T_j \) is \( p_{mn} \) ( \( p_{mn} \) can be obtained from formula (8), then the similarity computing formula of \( w_m \) and \( w_n \) is as shown in formula (9):

\[
correlation(w_m, w_n) = \frac{p_{mn}}{p_{im} + p_{jn} - p_{mn}}
\]  

(9)

According to formula (9), when the value of \( p_{mn} \) is 0, \( correlation(w_m, w_n) = 0 \) , which means feature word \( w_m \) and \( w_n \) is uncorrelated. When \( correlation(w_m, w_n) \neq 0 \), feature word \( w_m \) and \( w_n \) 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):

\[
Similarity(d_i, d_j) = \lambda D_{s}(d_i, d_j) + (1-\lambda) \left[ \sum_{w \in W} (1-correlation(w_{im}, w_{jn})) / |W(V-1)| \right]
\]  

(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 \( \varphi \) and \( \theta \);

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 $\varphi$ and $\theta$;
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):

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

(11)

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

(12)

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

(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: model the document by LDA topic model and do model solve and efficiency analysis by Gibbs sampling algorithm. In the experiment, assign value to $\alpha$ and $\beta$ according to Document [12]. Let $\alpha$ be 50/K, $\beta$ be 0.01, T be 100, which generates the best effect. Do the iteration for 1000 times to get the probability distribution matrix $\theta$ of document-topic and probability distribution matrix $\varphi$ of topic-word.

3) Document similarity computing: measure document similarity by the similarity computing method mentioned in 4.2. Let $\lambda$ be 0.1、0.2……0.9. Repeated comparative testing and analyzing show that the result is the best when the value of $\lambda$ equals 0.7. Thus $\lambda$ 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>
<thead>
<tr>
<th>Category</th>
<th>LDA+JS</th>
<th>LDA+JS+ Word co-occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision Ratio</td>
<td>Recall Ratio</td>
</tr>
<tr>
<td>Art</td>
<td>0.7613</td>
<td>0.7975</td>
</tr>
<tr>
<td>Computer</td>
<td>0.7298</td>
<td>0.7225</td>
</tr>
<tr>
<td>History</td>
<td>0.7818</td>
<td>0.7525</td>
</tr>
</tbody>
</table>
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**


