ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp.360-365
doi: 10.18178/wcse.2017.06.062

GBTM: A Short Text Clustering Model Based on Word Pairing

Mengmin Tian [1,2+], Ping Lu [1,2], Jincai Chen[1,2] and Min Wu[1,2]

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, P.R.China

[2] Key Laboratory of Information Storage System (School of Computer Science and Technology, Huazhong University of Science and Technology), Ministry of Education of China

Abstract. With the rapid development of Internet and many kinds of mobile applications, the number of short text has been growing rapidly. Because of the semantic spare problem and the context dependency problem, the traditional semantic mining in social network is inefficiency. At present, semantic mining of short text mainly considers the word correlation, without considering the correlation of word pairs. In order to more in-depth mine the semantic of short text, the GBTM model for short text clustering based on word pairing is proposed, firstly the text - topic probability distribution is obtained by mining the word pairs' correlation, on the basis of this, the topic correlation between the text is calculated using K-means clustering algorithm combined with the JS distance. The experimental results show that, the proposed GBTM model has a certain improvements in the clustering effect Purity (accuracy) and F-measure (precision and recall rate ratio) compared with LDA model and BTM model, Therefore, the mining of the word pairs' correlation can help to improve the efficiency of short text topic clustering.

Keywords: topic modeling, GBTM model, Gibbs sampling, cluster description, word pairs' correlation

1. Introduction

The traditional text clustering model is based on the Bag of Words model [1-2], which is a natural language processing procedure based on a simple hypothesis: the word or phrase of a text is not context-sensitive, their order does not affect the semantics. The model can also cause a lot of data sparse problem [3-4], so short text clustering has more challenger than long text clustering. In order to solve the problem of short text data sparseness, most research methods for short text clustering are based on the improvements of feature extraction, There are two key ideas of abstracting frequent word sets [5] and extracting repeat substrings [6]. "Frequent word set" refers to some of the higher frequency of words or phrases. Using the "frequent word set" some meaningless vocabulary , which is little help for text semantics can be deleted, and it can play an important role in reducing dimension of word vector [7-9], but also a good description of clustering results, however on the quality of clustering can't be significantly improved.

Yan Xiaohui et al proposed the BTM (Biterm Topic Model for Short Texts) Model at International Conference on World Wide Web In May 2013 [10]. The model introduces the similarity of word parsing, overcomes the semantic sparse problem of short texts. But in the interpretation of short texts, not only to consider the meaning itself, but also its background knowledge which is related to it. In 2016, LI et al have proposed a short text clustering model with auxiliary word embedding at the International ACM SIGIR Conference ACM [11]. The model improves the clustering effect by learning the relevant background of the external file, but it also extends the clustering time. In this paper, considering the word pairs' correlation, based on BTM model considered with CRP [12] (Chinese Restaurant Process) introduces the correlation of word pairs, and then comes up with a model named GBTM (Gravity Biterm Topic Model). In the process of

+ Corresponding author. Tel.: + 152-7181-8348; fax: +027-87792284.
E-mail address: luping06@hust.edu.cn.

Gibbs sampling with in the GBTM model, the similarity matrix of word pairs in a whole set of word pairs based on CRP is calculated. The introduction of the word pairs' correlation is not only able to quickly determine the topic of vocabulary, deep mining the semantic of text content, but also can improve the accuracy of topic clustering. Experiments were carried out on the data set of Sougo. The experiment results showed that the GBTM model has a higher clustering effect than the BTM model.

2. Word Parsing Similarity of BTM Topic Model

Most of the existing methods based on the long text modelling need to use the expansion of the external corpus, but the BTM model has no external resources for the short text topic learning [13]. For short text, BTM model is based on word co-occurrence patterns; after the text preprocessing (Chinese word segmentation and remove the low frequency and stop words) [14-15], any two words in the text form a word pair regardless of the order. Finally, we can get all the word pairs set for the entire text data set. The essence of BTM topic model is to learn the theme based on the whole word pairs set [16], and the whole corpus set is considered as the mixed distribution of the topic, instead of considering the individual short texts. From the previous text - topic matrix to the present corpus - topic matrix, the number of rows of the matrix is directly changed from the original N (the number of the texts in the corpus) to one, which can solve the semantic spare problem mentioned above.

It is found that LDA (Latent Dirichlet Allocation) topic model [17-20] is starting from a text D, each text here is regarded as a vector of words, and each word in the text is assigned a theme through a certain probability, and then each text generates a topic distribution. Finally, a probability distribution matrix, which takes the texts as the columns and the theme topics as the rows, is generated. The dimension is D×N dimension, D is text number, N is vocabulary number. Since each text is composed of a large number of words, the result of Gibbs iteration sampling should depend on the last data. Therefore, the distribution of other words in the text may affect the distribution of the current vocabulary in this iteration. When the text is very short, the impact will be highlighted, thereby affecting the entire topic model of learning. However, the BTM theme model starts from the overall situation and regards the entire text datasets as a text, so that it has rich semantic information and generates a 1×K dimension of the topic probability matrix without losing the word pairs' correlation, K is the number of topic. At the same time, because the different words in each text are independent, it can infer a text corresponding to different theme of the probability, and then you can deduce the document-topic matrix distribution. BTM model graphics as shown in Fig. 1:

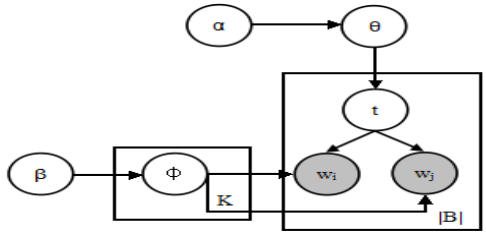


Fig. 1: Description of BTM model diagram

Table 1: Some notations and the description

Notations	Description
θ	The topic probability distribution of BTM model corpus
Ø	The topic-biterm distribution (biterm)
t	The topic of the biterm selected from θ
w_i, w_j	Two words in the biterm
|B|	The number of all biterms in the corpus

The steps of generating all biterms of BTM corpus are as follows, we first define some notations in table I.

(1)For each topic t, generate a word distribution $\theta \sim Dir(\beta)$;

(2)Generate a topic distribution $\theta \sim Dir(\alpha)$ of the entire document data set;

(3)Each biterm operation of the biterm collections B is as follows: hypothesis biterm=(w_i, w_j),extract a topic K from the topic distribution θ of the entire text data set of the short text, that is $K \sim Mulit(\theta)$, then extract two words from this topic K extracted w_i, w_j, w_i, w_j $\sim Mulit(Ø_K)$.

To repeat the above process, all the words in the short text corpus are generated. However, the BTM model is directly mining the word correlation in the biterms after the composition of the biterm, there is no more deeply mining of the correlation between the words, Therefore, considering the correlation between biterms, the GBTM model is proposed, which can deeply mining the semantic of short text.

3. Correlation of the world pairs in GBTM model

3.1. The combination of the word pairs' correlation and CRP

The idea of CRP is to simulate the process of dining in a restaurant (Infinite table) of M customers, it's a Dirichlet process. The customer's distribution process is shown in the formula 1, Customer J on the table J, The probability of customer m choosing table Z_m=j is related to the distribution of customers:

$$p(Z_{m=j} \mid Z_{-m}) = \begin{cases} \frac{C_j}{m-1+\gamma}, \text{Table j has } C_j \text{ customers} \\ \frac{\gamma}{m-1+\gamma}, \text{Table j has no customer} \end{cases} \quad (1)$$

where γ is a priori parameter of Dirichlet.

It is obvious that the CPR process is flawed, and the customer choose the table only depending on the number of other customers on the table, which means the larger number is more attractive to the customer. But there is still a gap with reality, such as if the customer noticed the guests are very noisy on the table, he would like to choose another table. Or if his friends were on another table even with a small number of people, he would choose to sit with his friend to eat. Obviously CRP is not considering the correlation between customers. While the GBTM model is to consider the effect of the primary customer on the following customer, which is influenced by the customer's own properties, and not just by the number of customers on the table. The basic idea of GBTM is adding biterm similarity matrix in the Gibbs sampling process to more in-depth mining the relevance of biterm. For the GBIM model, the dining table is topic and the customer is the biterm.

3.2. Word pairs' correlation and Gibbs sampling

The correlation between the word pairs in GBTM model is obtained through the process of Gibbs sampling. Gibbs sampling is an algorithm used to obtain a series of approximate and equal to the probability distribution of the observed samples (For example, the joint probability distribution of 2 or more random variables) in the Markov Monte Karl theory(MCMC). When using the GBTM model to carry on the topic clustering, the document - topic matrix θ is mainly used, Each row in the matrix represents the topic probability distribution of the document P(z|d). If you want to get the probability distribution of the current sample of the words W_{di} to the topic Z_{di}, the similarity matrix *sim*(*sim*=X^T**X*, X is the word matrix of the whole corpus.)of all words must be calculated first. And the final matrix is the probability distribution of the words to the topic Z_{di}. The formula of probability calculation of B distribution to the j topic in the rest of the topic is shown as follows:(The left probability P is proportional to the right).

$$P(z_{di} = j \mid w_{di} = B, z_{-i}, w_{-i}, sim) \propto \frac{\sum_{B' \setminus B} (C^{WT}_{B'j} \cdot sim_{B'B}) + C^{WT}_{Bj} + \beta}{\sum_{B'} C^{WT}_{B'j} + V\beta} \cdot \frac{C^{DT}_{dj} + \alpha}{\sum_{j'} C^{DT}_{dj'} + T\alpha} \quad (2)$$

The symbolic meaning is shown in table II. The formula (2) $Ø_{Bj}$ is the probability of the word B in the topic J, θ_{dj} is the probability of the text d in the topic J.

Table 2: Symbolic meaning in formula

Symbolic	Description
Z_{di}	a topic number Z is randomly selected for a certain probability; and then select the topic-word dictionary which is numbered as Z; A word i in the d of the document is selected with the probability of 1/K.
W_{di}	the words B is made up of word i and the word of document D.
α	value of the prior distribution parameters of topic distribution, here α takes the experience value 0.5
β	value of the prior distribution parameters of word distribution, here β takes the experience value 0.1.
T	The number of topics, V is the number of words. At this point W is the words B.
$C^{WT}_{B'j}$	The number of word pairs(Apart from the word B) obtained in the topic - word frequency matrix(WT) on the topic J.
C^{WT}_{Bj}	The frequency of the word B on the topic J in WT.
C^{DT}_{dj}	The number of word pairs obtained in the text- topic frequency matrix(DT) on the topic J.
$C^{DT}_{d'j}$	The number of word pairs obtained in the text- topic frequency matrix(DT) on the topic except for J.

B assigned to a new theme j=T. The formula of probability calculation is as follows:

$$P(z_{di} = j \mid w_{di} = B, z_{-i}, w_{-i}, sim) \propto \frac{0 + \beta}{0 + V\beta} \cdot \frac{0 + \alpha}{\sum_{j'} C_{dj'}^{DT} + T\alpha} \tag{3}$$

This will get a text - topic probability distribution matrix and a topic-word pair probability matrix, which directly affects the accuracy of the text - topic probability distribution matrix. Therefore, we can get the text topic probability distribution matrix by mining the correlation between words, and provide more reliable semantic clustering results for the following topic clustering.

4. Topic Clustering of GBTM Model

Because the GBTM model is the topic model, the text representation is presented in the probability distribution of the topic. Therefore, the similarity between two texts needs to be computed by the probability distribution vector of the topic, and that is the text topic probability distribution matrix, which is said in the 3.2 section.

To calculate the distance between two probability distributions of the classical algorithm is KL distance [22], assuming that there are two probability distributions p and q, The probability value of the distribution is p(z | d), p and q can be calculated by the 2.2 section formula (1). Their KL distance formula is as follows:

$$D_{KL}(p \| q) = \sum p_i \log \frac{p_i}{q_i} \tag{4}$$

But the problem of KL distance is about asymmetry, that is, $D_{KL}(p \| q)$ and $D_{KL}(q \| p)$ is not same, therefore there is a distance: JS distance, which is proposed to solve the asymmetry of KL distance, the formula is as follows:

$$D_{JS}(d_i, d_j) = \frac{1}{2} D_{KL}(d_i \| \frac{d_i + d_j}{2}) + \frac{1}{2} D_{KL}(d_j \| \frac{d_i + d_j}{2}) \tag{5}$$

The text clustering of GBTM model is to calculate the distance of two probability distribution using K-means clustering algorithm combined with JS distance. Specific algorithm is as follows:

(1)Select K texts from the all texts as the K centers of the K clusters.

(2)Measuring JS distance of each remaining text to each cluster center, if the JS distance is minimum from the text to the center, then classify the text.

(3)After classifying the all texts, then recalculate the new K centers of the all text.

(4)Repeat steps 2-3 until the new center with the original center is less than or equal to a specified threshold, then the convergence ends.

The K-means clustering algorithm is the first to cluster the documents in the corpus. Then, it uses JS distance to calculate the semantic distance between the texts, to get better clustering results.

5. The Results and Analysis

5.1. Experimental data

The data set is from Sogou laboratory area, the content is all about journalism, which is divided into ten categories, ten categories are as follows: C000007 car; C000008 finance; C000010 IT; C000013 health; C000014 sports; C000016 tourism; C000020 education; C000022 recruitment; C000023 culture; C000024 military. There are 100 texts in each category, after the truncation, and then through to carve words and stop words, ending with the length of each text is not more than 200 characters.

5.2. Evaluation Method

The most common evaluation methods: Purity value; F value.

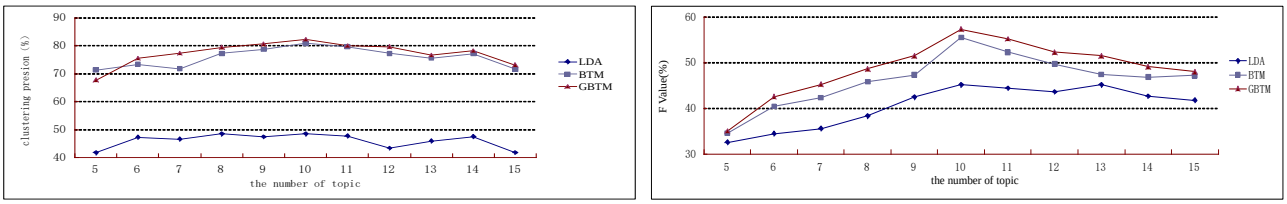
(1)Purity value

Purity method is to calculate the proportion of text number of correct clustering to the text of the whole text data in each cluster, the formula is as follows:

$$purity(\Omega, C) = \frac{1}{N} \sum \max | \omega_k \cap c_j | \tag{6}$$

In the formula (5), $\Omega=\{\omega_1, \omega_2,\ldots,\omega_k\}$ is the entire data sets, ω_k is the k^{th} clusters, $C=\{c_1, c_2,\ldots, c_j\}$is the texts sets which are correctly clustering, c_j is the j^{th} text, N is the number of the texts of the entire text sets. Since we know the data sets are divided into 10 categories in advance .Therefore, it is possible to infer the best number of topics is likely to be 10. So the Purity value of clustering is calculated by using 5, 6, 7, 8, 9, 11, 12, 10, 13, 14, 15 topics respectively. We can observe whether the effect of clustering under different number of topics can be improved after introducing the correlation between words.

Through the analysis of the difference in classes and between classes of clustering results the corresponding topics in the model, as well as comparison of LDA model and BTM model clustering effect, it can be seen that the accuracy of the GBTM model has been improved 3.26% on the average. Purity value of clustering effect is shown in Fig.2 (a):



(a) Purity values (b) F values

Fig. 2: Comparison of Purity values and F values of GBTM, BTM and LDA models

(2)F value

Before introducing the F value, First look at the precision (P) and recall (R). Ideally, the clustering effect is best when both the precision and recall values are maximized, but in fact, precision and recall rates are inversely proportional to the general conditions[23], that is, the higher the accuracy, the lower the recall rate, and vice versa. Therefore, there is need to adjust precision and recall rate according to the demands, then F value appears at this time, the value of α could be changed, the formula is as follows:

$$F_\alpha = \frac{(\alpha^2+1)P\cdot R}{\alpha^2(P+R)}(0<\alpha\le 1) \tag{7}$$

Compared the F value of each topic model, it shows that the F value is the biggest when using GBTM model. Compared with BTM model, the F value has been proposed 5.6%. Therefore, the relative effect of clustering have a certain promotion, F value of each model is shown in Fig. 2(b):

(3)comparative analysis

From Fig. 2 we can see that the effect of GBTM on K=10 (the number of theme) is the best. The effect of BTM in K=10 is also optimal. The effect of BTM is better than LDA, and the effect of GBTM is better than that of BTM. For GBTM, when the number of topics is too large, the clustering effect will be a slight decrease. Originally, each short text containing fewer words, the number of words formed is also less, so when the theme is too large, it is possible to cause the words-topic probability matrix to be split， which not only will affect the distribution of the topic probability after modeling, but also will affect the expression of the document, and then the clustering effect will be affected.

6. Conclusion

In this paper, we propose a GBTM model based on the correlation between word pairs. In the process of modeling the Gibbs sampling, the word pairs set is used to calculate the similarity matrix of the word pairs, the model not only takes into account the correlation between words, but also takes into account the semantic relationship between word pairs, so more semantic content of the text will be dug up. But the GBTM model influences the probability distribution of text-topic through probability distribution of topic-words. Based on Text-Topics probability distribution, using K-means clustering algorithm to calculate the semantic distance (topic correlation) between the texts with JS distance of probabilistic graphical models, and then cluster the results. Experiments were carried out on the data set of the Sogou. After analysis of the results, the feasibility of the GBTM model in the short text data set was confirmed, and the cluster results had improved than BTM model. The research can be widely used in the discovery of hot topic of micro blog, the acquisition and question recommendation of the different users' interest in the interactive question answering system and the generation of user specific content so on.

7. References

[1] Y. X. Wang, H. Guo, C. Q. He. Bag of Spatial Visual Words Model for Scene Classification[J] . *Computer Science*, 2011, 38(8):265-268.

[2] F. Y. Cao, W. T. Niu. A Block Data Clustering Algorithm Based on the Bag of Word Model. *Journal of Shanxi University (Natural Science Edition)*, 2016, 39(2):216-222.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. *In Proceedings of 16th World Wide Web Conference (WWW16)*,2007,pp. 757-766.

[4] B. K. Wang, Y. F. Huang, X. Li. Short text classification based on strong feature thesaurus. *Journal of Zhe-jiang University Science C*, 2012, 13(9): 649-659.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research, 2003*, 13(3): 993-1022.

[6] J. X. Hu, Y. Liu. Short-text Clustering Based on Repeats. *The Eighth National Conference on Computational Linguistics*, Nanjing, 2005,pp. 27-29.

[7] J. Turian, L. Ratinov, Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, 2010,pp. 384-394.

[8] T. Mikolov, W. T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013,pp.746-751.

[9] K. Zhou, Y. Liu, J. K. Song, et al. Deep Self-taught Hashing for Image Retrieval. *In 23th ACM International Conference on Multimedia*, Brisbane, Australia, 2015,pp. 1215-1218.

[10] X. H. Yan, J. F. Guo, Y. Y. LAN, et al. A Biterm Topic Model for Short Texts. *International Conference on World Wide Web*, 2013,pp. 1445-1456.

[11] C. L. Li. Topic Modeling for Short Texts with Auxiliary Word Embedding. *The International ACM SIGIR Conference ACM*, 2016,pp. 165-174.

[12] X. P .Zhang, X. Z. Zhou. A Topic Model Based on CRP and Word Similarity. *Pattern Recognition And Artificial Intelligence* , 2010, 23(1): 72-76.

[13] J. L. Yang. Research on the short text clustering based on external knowledge. Tianjin, Nan kai University, 2010.

[14] http://www.ictclas.org. ICTCLAS.

[15] F. Li. Research on several key technologies of text mining[M]. Beijing University of Chemical Technology, 2010.

[16] Y. Zhang. Short text similarity computation based on BTM theme model feature extension[M]. Anhui University, 2013.

[17] X. H. Phan. learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections.*WWW2008/Refereed Track: Data Mining - Learning*. 2009, 12(3): 23-34.

[18] S. Q. Zhao, T. Liu, S. Li. A Topical Document Clustering Method .*Journal of Chinese Information Processinc*, 2007,pp. 57-62.

[19] H. Xie, H. Jiang. Improved LDA model for microblog topic mining Journal of East China Normal University (*Natural Science*) 2013, 6(6): 93-101.

[20] P. J. Zhang, L. Song. Overview on Topic Modeling Method of Microblogs Text Based on LDA. *Library and Information Service*, 2012, 56(24): 120-126.

[21] P. P. Liang, Y. M. Chai, L. M. Wang. Relational Text Classification Algorithm Based on iTopic Model. *Computer Engineering*, 2011, 37(21): 124-125.

[22] C. Zheng, H. Li. Text Clustering of K-means Based on LDA. *Computer and Modernization*, 2013, 8(8) :78-80.

[23] Z. T. Zhou. Quality Evaluation of Text Clustering Results and Investigation on Text Representation. BeiJing, CAS Institute of computing, 2005.