Short Text Topic Discovery Based on BTM Topic Model

Wei-Dong Zhu[1,a] and Wen-Gan Zhou[2,b,*]

[1]Information Center, Beijing Jiaotong University, Beijing, China

[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

[a]wdzhu@bjtu.edu.cn, [b]17120465@bjtu.edu.cn
*Corresponding author

Keywords: BTM, JS distance, Singel-Pass clustering, Short text topic discovery.

Abstract. With the further development of the online social platform, the research techniques of hot topic related to short text data which are represented by Weibo, instant messaging, news commentary and so on, are not extensive enough, and the research efforts are not deep enough either. Moreover, short text data set has many characteristics such as high noise, sparsity, and irregular specification, which makes the performance of traditional topic research techniques insufficient. Therefore, for the data characteristics of short text, this paper uses a short text topic discovery method based on BTM (Bi-term Topic Model) theme model. Firstly, the BTM of the processed short book is modeled to meet the probability distribution of the subject obtained after the data language features of the essay are modeled. Then JS distance is used as the text similarity measure, combined with the improved Single-pass clustering algorithm to find out the hot topic of short text data set. The comparison experiments show that the short text modeling and improved single-pass algorithm use BTM making the clustering efficiency improved, and it can effectively solve the problem of data sparsity in short texts. There has been a remarkable improvement in the quality of the topic discovery.

Introduction

The short text on various social platforms is not only tremendous, but also the growth rate is extremely fierce. Individuals life, work and learning have been fully infiltrated by all kinds of information on the network, affecting the formation of individuals three views. Therefore, timely grasping the topics of concern to netizens will help the public quickly identify and track hot topics in the society, and take appropriate measures to guide them.

The short text is simple and clear, most of which is less than two hundred words, and the data in the short text has a large data sparseness. Taking microblog short text data as an example, the context between texts is not large, and the number of words is small and refined. Commonly used topic models are Vector Space Model (VSM) and LDA (Latent Dirichlet Allocation) topic models, but these models are often applied to traditional long text, without considering the particularity of short text[1,2]. It is not suitable for microblog short text data, thus affecting the quality of microblog short text hotspot discovery.

Combining with the shortcomings of short text modeling, the method of short text topic hotspot discovery based on Bit Topic Model (BTM) is adopted to solve the problem of data sparsity in short text hotspot discovery, so as to improve the quality of topic discovery. The method uses the BTM model to model the obtained short text, the topic discovery stage, the subject-vocabulary distribution and the probability distribution of the topic's distribution vector, and then uses the JS distance as the text similarity measure. Finally, the improved single-pass clustering algorithm for text topic discovery is used.

Related Work

Biterm Topic Model

At The theme model breaks the document-theme layer of the traditional model. By converting the document into a word pair, the word pair refers to two words that are co-occurring after the document

is pre-processed; the word pair of the entire corpus is modeled to overcome the short text. The sparse issue takes into account the semantic relationship between words and better understands short text information than traditional models. The BTM theme model is shown in Figure 1[3,4].

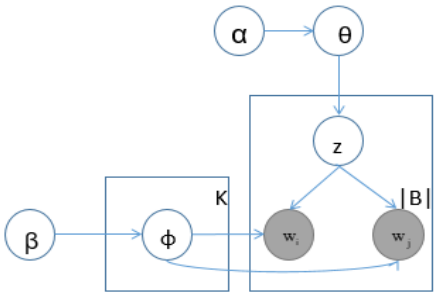

Fig. 1 BTM topic model

In the model, θ stands for the distribution of topics in the entire corpus, φ is the distribution of words under a certain topic, z is the subject of a pair of words, and w_i, w_j stands for two different words that constitute a pair of words. | B | is the number of words contained in the entire corpus, and k is the number of topics in the corpus.

The BTM model is modeled on the basis of the generated word pairs. The effective solution solves the issus of sparseness of short text. For the word pairs in the entire language database, the specific BTM modeling process is described as follows:

For the entire corpus, there is a topic-distribution θ~ Dir(α), and α is a prior parameter.For each topic Z, the word distribution under this topic is ΦZ ~ Dir(β), and β is a prior parameter.

For each word pair in the pair B, b = (w_i , w_j):

a. Randomly extract a topic Z from the topic distribution θ of the entire corpus, then Z ~ Multi(θ).

b. From the extracted topic Z, randomly extract two different words w_i , w_j , which form the word pair b, and w_i , w_j ~Multi(φz).

JS (Jensen-Shannon) Distance

KL distance [5] measures the difference between two probability distributions, which is defined based on the concept of information entropy, so it is also called relative entropy and cross entropy. Let p(x) and q(x) be two probability density functions on X, and the KL distance between them is defined as:

$$KL(p,q) = \sum_{x \subset X} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{1}$$

The KL distance does not have symmetry, ie $KL(p,q) \neq KL(q,p)$, for this issus. The JS distance is better and the distance is defined on the closed interval of [0, 1]. The formula is as follows:

$$JS(p,q) = \frac{1}{2}\left[KL(p, \frac{p+q}{2}) + KL(q, \frac{p+q}{2}) \right] \tag{2}$$

Single-pass Clustering Algorithm

The algorithm assigns the latest data is input to the most similar class otherwise creates a new class each time according to the pre-entered threshold according to the order of data input. The Single-pass clustering algorithm is simple and efficient, and does not require doing sets the number of clusters, which is suitable for processing dynamically growing data streams. The algorithm steps are as follows: 1) input new text d: 2) calculate the similarity between d and each document in the existing topic classification, obtain the topic with the greatest similarity with d and obtain the similarity value T; 3) provided that T is greater than the valve Value, subsequently document d is sorted into a known topic category, otherwise as a new topic category; 4) the clustering process ends, waiting for new text, and so on.

The Single-pass algorithm usually selects the first short-text as the center of the first category as the initial clustering, which facilitates the comparison of similarities with the documents that come next. So the selection of the center of the cluster determines the consequences of this clustering algorithm, and each time a new topic is generated, the topic center requires doing recalculated each time. In the Single-pass algorithm, when a new short-text is coming, the similarity calculation is not performed with all the short-texts, but just the similarity calculation is performed with the topic center. When the JS distance is greater than the preset threshold, the document is returned to the topic cluster, and when the JS distance is less than the preset threshold, the document is placed in the queue to be processed. Create a new topic cluster until the queue is empty.

Short Text Topic Discovery Model

The short text topic discovery is to discover the implicit relationship between the text content through the theme-model is processed, text clustering and other methods, and find the social hot-topics that the user discusses otherwise pays attention to. The main steps of the topic model used in this paper are as follows: firstly, the acquisition and pre-processing of microblog short text data; subsequently, based on the obtained pre-processed short text data, the BTM topic model is used for short text modeling to obtain the topic probability distribution; on this basis, combined with the Single-pass clustering algorithm suitable for topic discovery; finally, using JS distance as a measure of text similarity, get hot topics.

Short Text Modeling

The experiment combines the characteristics of the microblog short text to model the microblog short text. The modeling process is as follows.

Short Text Data Collection and Preprocessing. The experiment uses web crawlers to automatically crawl relevant microblog short text. The obtained short text is subjected to text preprocessing to convert the text data into structured data. The purpose of short text pre-processing is to reduce short-text data noise and prepare for subsequent BTM topic modeling.

BTM Topic Model Modeling. The BTM model modeling process into a BTM corpus[6], and subsequently through the generated corpus modeling, the microblog short text into a word pair, and thus rise to the level of the entire corpus to describe the topic Z, that can be expressed between words different topics can maintain the relativeness between words. The joint probability of the word pair b = (w_i , w_j) can be expressed by the following formula:

$$P(b) = \sum_z P(z)P(w_i \mid z)P(w_j z) \tag{3}$$

The probability of the entire corpus in the BTM model is represented by the following formula:

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{4}$$

Two important parameters Φ and θ in the BTM topic model are two implicit variables of the BTM topic model.The conditional probability formula for bi-term is as follows:

$$P(\mathrm{z} \mid z_{-b}, B, \alpha, \beta) \propto (\mathrm{n}_z + \alpha) \frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2} \tag{5}$$

The conditional probability distribution of each word pair is $P(z \mid z_{-b}, B, \alpha, \beta)$, where z_{-b} represents the distribution of the topics of all pairs of words except the word pair b in the entire corpus, and B represents the set of word pairs in the corpus, that is, contains all pairs of words, α , β is the first Test parameters, $\mathrm{n}_{w|z}$ represents the number of times the word pair is assigned to the topic Z; , $\mathrm{n}_{w_i|z}$, $\mathrm{n}_{w_j|z}$ represents the word i, j assigned to the topic Z; M represents different words in the corpus

generated by the entire model Number of;The probability formula for the distribution of topics in the corpus obtained after the entire modeling is as follows:

$$\theta_z = \frac{n_z + \alpha}{|B| + K\alpha} \tag{6}$$

The probability formula for the topic-word distribution in the corpus is as follows:

$$\varphi_{w|z} = \frac{n_z + \alpha}{\sum_w n_{w|z} + M\beta} \tag{7}$$

The document-topic distribution $P(z|d)$ is calculated from the biterm-topic distribution $P(z|b)$ and the document-biterm distribution $P(b|d)$. The formula is as follows:

$$P(z|d) = \sum_b P(z|b)P(b|d) \tag{8}$$

The prior parameters of this experiment and the empirical values are set to 50/K and 0.01 respectively, respectively, where K is the number of topics set in advance. After the above steps, the word vector space of the short text is mapped into the topic vector space of the short text, and the theme-modeling of the short text is fulfilled .

Determination of the Optimal Number of Topics K

When training the BTM topic model, it is necessary to determine the optimal number of topics K. The value of this parameter directly affects the quality of the consequence. In order to strengthen the accuracy of the results and strengthen the performance of the algorithm, the parameters are optimized using the confusion in this experiment.

Confusion is generally used to measure the probability of a probability distribution or a probabilistic model to predict a sample. A low-probability probability distribution model can better predict a sample and is an important indicator for measuring the generalization ability of a model. The formulas for calculating the degree of confusion are as follows:

$$perplexity(D) = \exp\left\{-\frac{\sum \ln P(b)}{\sum_{m=1}^{M} N_m}\right\} \tag{9}$$

N_m represents the set of words in the mth document, and P(b) represents the probability that the BTM model produces the word pair w, where P(b) is calculated as follows:

$$P(b) = \sum_z p(z) \times p(w_i|z) \times p(w_j|z) \tag{10}$$

In this experiment, according to the number of topics that may appear in the microblog short text, set K=10~100 as the limit, and then compare the degree of model confusion under the number of topics. Figure 2 shows the trend of confusion of the BTM theme model at different K values.

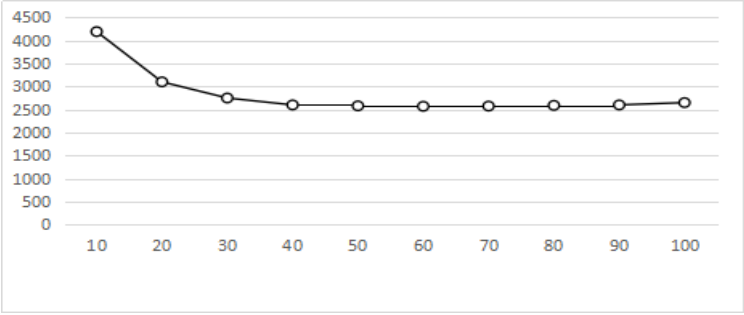

Fig. 2 Confusion under different themes

According to the above figure, as the number of topics increases, the confusion becomes lower and lower, and then the confusion is lowest when the K value is around 70, and afterward the confusion increases with the number of topics. Therefore, K = 70 was taken in this experiment.

Experiment and Result Analysis

The short text used in the experiment was collected from Sina Weibo, and the validity and correctness of the BTM theme model combined with the improved Single-pass algorithm were demonstrated from the experimental results. In the first set of experiments, the combination of the VSM model & classic single-pass algorithm, the combination of the LDA model & classic single-pass algorithm, and the combination of the BTM model & the classic single-pass algorithm were compared. In the second set of experiments, the BTM model & classic single-pass algorithm was compared with the BTM model & improved single-pass algorithm to demonstrate the effectiveness of the experiment.

Experimental Evaluation Index

In the field of natural language processing [7], the accuracy of the (Precision), recall (Recall) [8], F-value (F-Measure) is generally used to evaluate the text clustering results, the larger the F value, the clustering the better the effect, so this paper uses these three indicators to evaluate the performance of the clustering algorithm. Used to evaluate the effect of subject detection, this paper uses the three indicators of False Alarm (FA), Missing Rate (Miss) and Normalized Detection Cost in the TDT evaluation standard. The smaller the normalized detection cost is, the overall performance of the experiment is better

The accuracy rate P represents the proportion of the number of documents belonging to a topic correctly retrieved by the algorithm to the total number of documents identified (including the number of documents identifying the error). The recall rate R indicates that the number of documents of a topic that is correctly retrieved accounts for the proportion of the original number of documents in the topic in the actual total number of documents in the corpus. Since the accuracy rate and the recall rate are often mutually influential, the F value is generally used to comprehensively measure the accuracy rate and the recall rate.The formula for the f value is:

$$\mathrm{F}=\frac{2\times P\times R}{P+R} \tag{11}$$

The false positive rate indicates that the number of documents that are not related to the topic is detected as a percentage of the actual number of documents in the corpus that are not related to the topic. The false negative rate indicates that the number of unreported documents related to the topic accounts for the proportion of the actual number of documents in the corpus related to the topic. The index of the comprehensive performance of the normalized detection cost evaluation topic detection algorithm is calculated by comprehensively calculating the false alarm rate and the false negative rate. The formula is:

$$\left(C_{Det}\right)_{\mathrm{Norm}}=\frac{C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{Fa}\cdot P_{Fa}\cdot P_{\neg target}}{\min(C_{Miss}\cdot P_{target},C_{Fa}\cdot P_{\neg target})} \tag{12}$$

Experimental Data Description

This experiment uses crawler technology crawl micro-blog content, and the test data include 10 topics and 10,000 microblogs to verify the experiment. Microblog short text data has an army of redundant information, some useless information and special data format, and removes useless information through specific regular expressions, such as @user, hyperlink, etc., using jieba tools to perform word segmentation processing, according to integration optimization the stop vocabulary removes the stop word.

Analysis of Results

Firstly, the results of the first set of comparative experiments are analyzed to verify the impact of the BTM topic model on the topic discovery compared to the traditional VSM model and the LDA topic model. Since this group of comparative experiments mainly verified the influence of the text representation model on the topic discovery effect, the classic Single-Pass clustering algorithm was adopted in all three experiments, so we only used the false negative rate, false positive rate and topical normalization detection. The value is used to evaluate the results. In order to analyze the results more intuitively, a histogram of the normalized value of a set of experiments on 10 topics was drawn, as shown in Figure 3. The histograms are VSM model, LDA model, and BTM model, respectively.

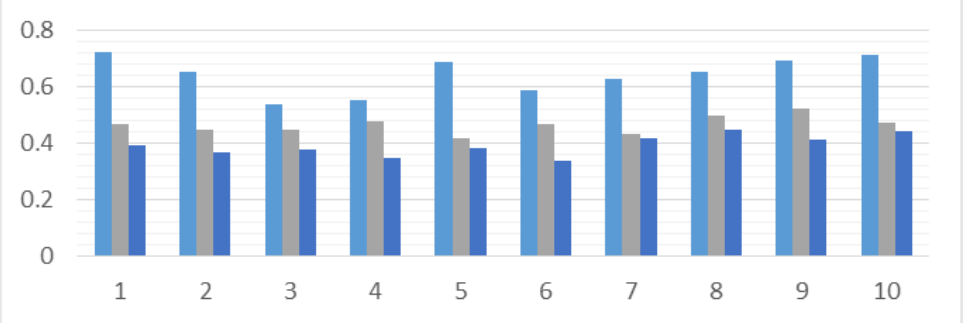

Fig. 3 Experimental iterative process

From the above experimental comparison results, the BTM theme model is better than the traditional VSM model and the LDA model. Average Cost of Detection per topic is reduced by 5.01%, verifies the superiority of BTM model.

In the comparison experiment of clustering algorithm, in the case of the same text representation model, the classic Single-Pass clustering algorithm and the improved Single-Pass clustering algorithm is implemented respectively. Next we will use the accuracy, recall-rate and F value. The results of the clustering were evaluated, and the experimental results were compared as shown in Table 1 below.

Table1 Feature inclusion part in the first iteration

Topic	Classic Single-Pass			Improve Single-pass		
Number	P	R	F	P	R	F
1	0.8741	0.7102	0.7837	0.9000	0.8133	0.8544
2	0.8534	0.7279	0.7857	0.8739	0.7647	0.8157
3	0.7627	0.6923	0.7258	0.8475	0.7692	0.8065
4	0.7115	0.7255	0.7184	0.78	0.7609	0.814
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.7778	0.6885	0.7304	0.8426	0.7459	0.7913
Average	0.7762	0.7057	0.7383	0.8344	0.7925	0.8117

According to the above table, the improved Single-Pass clustering algorithm not only greatly improved the accuracy, but also greatly improved the recall rate, and its average accuracy increased from 77.62% to 83.44%. It rose by 5.82%, and the average recall rate increased from 70.57% to 79.25%, an increase of 8.68%. The average F value increased by 7.33%. Therefore, the improved Single-Pass clustering algorithm is feasible for topic discovery research.

BTM modeling was performed on 10,000 microblog short text datasets crawled in this experiment. After text clustering, five typical topics were extracted from ten topics, as shown in Table 2 below.

Table 2 Feature inclusion part in the first iteration

Topic	Top8 Keywords
1	Li Yong, host, bid farewell, Havin, CCTV, US, anticancerr
2	E-sports, China, S8, Games, Finals,Winning, Korea, IG
3	Jin Yong ,martial arts, HK, passing away ,willingness ,rivers and lakes, martial arts,
4	Chongqing, bus, accident, driver, falling river, body, driving, passenger
5	Tang Yan, Luo Jin, Hu Ge, wedding, groomsman, bridesmaid, marriage, groom

According to the table above, "Li Yong's death event" has received extensive attention. "S8 Finals, Chinese Team IG Wins Event" are closely followed, and the time for the topic microblogging short text is also consistent. The hot words extracted in the method can accurately and comprehensively describe the meanings expressed by related topics. The vocabulary excavated by the model is the topic-related hot words in the microblog short text datasets.

In summary, it can be seen that after using the BTM theme model, the cost of normalized detection of the topic is reduced. On the basis of the improved Single-Pass clustering algorithm, the effectiveness of the algorithm is further improved. It proves the feasibility of Ideas for improvement proposed in the topic representation and topic clustering process.

Conclusion

It is prove by experiments that the BTM topic model can overcome the high sparsity issue generated by the LDA model when processing short text, and does not require doing the external patching data source for the semantic completion of short text. Experiments using the JS distance function to calculate the text similarity are better than using traditional cosine similarity clustering. And through the combination of the improved single-pass algorithm and the BTM theme model, the quality of the clustering results is improved comprehensively. However, based on experimental data, when the number of short text increases, the time ratio of modeling and clustering will increase, and further improvement is needed.

References

[1] GUO Qing-lin,LI Yan-mei,TANG Qi.Study on text similarity calculation based on VSM[J].Application Research of Computers,2008(11):3256-3258.

[2] Latent Dirichlet allocation. Blei D M,Ng A Y,Jordan M I. Journal of Machine Learning Research . 2003.

[3] Tang Qiulian. Short text clustering based on BTM [D]. Anhui University, 2014.

[4] Zhou Xiaotang. Research on fast Gibbs sampling topic inference algorithm for topic model [D]. Jilin University, 2018.

[5] Unsupervised Learning by Probabilistic Latent Semantic Analysis[J] . Thomas Hofmann. Machine Learning . 2001 (1).

[6] A Biterm topic model for short texts. Yan X,Guo J,Lan Y,et al. Proceedings of the 22nd international conference on World WideWeb . 2013.

[7] XU Ge,WANG Hou-feng.The Development of Topic Models in Natural Language Processing[J].Chinese Journal of Computers,2011,34(08):1423-1436.

[8] Zhu Yuqi,L ü Linyuan.A Review of Recommended System Evaluation Indicators[J].Journal of University of Electronic Science and Technology of China,2012,41(02):163-175.