Improvement of LDA Algorithm Based on Microblog Short Text Hotspot Analysis

Kai Li¹,a, Chunmei Li¹,b,*
¹ Department of Computer Technology and Application, Qinghai University Xining 810016, China
b likai614020758@126.com
 cite:li_chm0422@sina.com
* Corresponding Author

Abstract. In view of the high sparsity of microblog short text, the simple and random text content and the number of typos, and the use of standard LDA algorithm to calculate the result deviation of the keyword of short text in microblog is insufficient. In this paper, an improved LH-LDA model of microblog short text based on LDA algorithm is proposed. The model simultaneously links the heat of short text labels and short text. Through the relationship between the two and the text content, the keywords of a large amount of text are mined. At the same time, the correctness of the model is derived by Gibbs sampling, and the keywords of Weibo content are obtained. This article was adopted Python web crawler microblogging content, comment, forwarding, and praise. The results show that the traditional LDA algorithm is subject to more interference factors and the processing results are not accurate enough. From the results of the improved LH-LDA model processing in this short text, the keywords of this type of short text can be better found.

1. Introduction

With the development of modern society, the network has developed rapidly as a product of the new era, and many network-related industries have emerged. As one of the types, Weibo has developed rapidly. As a contemporary mainstream entertainment social segment, Weibo is popular in the world due to the convenience and speed of Weibo. Weibo users can log in a variety of ways or post their own content to share the content of others. At the same time, the number of Weibo users has reached 400 million, and the number of microblog readings in a single month has reached more than 25 in the field of 10 billion. Due to the huge number of users and the amount of information content, it is more important to mine and analyze the hot topics of microblog short text.

The hot topic of Weibo is that the number of people and attentions participating in a topic in Weibo reaches a certain level in a certain period of time. Therefore, whether the topic is a hot spot is usually determined by the number of participants and the degree of attention. Because the content of Weibo is simple, highly sparse, the text is randomly written, and the sentence structure is not fixed, the information expressed by the text vocabulary is very limited. Therefore, these factors determine that using the standard LDA algorithm to select the topic of the microblog text will cause a large error. At the same time, if the K-Means clustering algorithm is used to calculate the keywords of the short text, the clustering algorithm will produce a large deviation because the similarity between the texts ignores the internal relations of the text.
Combining the characteristics of essay, an LDA algorithm based on label and heat is proposed to improve the standard LDA algorithm. By collecting labels and heat, you can improve the LDA standard and increase the weight value. Secondly, sparse short texts with short texts with heat will make a measure of the importance of each short text. Helps to find keyword headlines for short text and increase accuracy.

2. Overview

In the discovery of the topic of short texts such as Weibo and social media, scholars at home and abroad have done a lot of research. Zhou Fuxing et al. proposed a method for mining hot topics on microblogs using a hybrid approach [1]. Although Weibo is labeled with and without tags, Weibo is a short text with a heat value. The heat of a text can represent the degree of attention of the topic; Huang Chang et al. proposed a BBTM improved model to mine the hot topic of microblogging [2], using the BBTM model to take into account the contextual meaning of short text, and integrated micro The popularity of the blog text, but did not take into account the loss of some hot topics after the word segmentation. Wang Hui used a model based on the POSTTLA model for microblog hot topic discovery [3], and used PSO to optimize the bp neural network to train the prediction model composed of time series. This method only considers the influence of time on the topic heat, and does not consider the attention of Weibo content and the influence of the tag of Weibo content on topic discovery; some scholars improve the LDA algorithm [4, 5], improve the essay The accuracy of the extraction of subject matter in this content. At the same time, some scholars analyze vocabulary [6,7,8,9], and mine microblogs and Twitter hot topics through eigenvalues such as vocabulary heat, label and time.

It is found that Weibo content related to hot topic keywords often has hotspot tags, and the tags are usually created by the official or individual and have a large number of people to discuss. Considering the attention of Weibo content is usually determined by forwarding, comments and praise. This paper proposes a short text topic analysis model based on label and attention. The core content of this model is the potential Dirichlet algorithm. In this paper, the hot topic of microblog short text is analyzed by the optimization of the algorithm.

3. Microblog short text topic analysis model based on tag and heat

3.1. Traditional LDA algorithm model

First, the LDA algorithm model is based on the PLSA algorithm model. The PLSA algorithm model is a word bag method. The process of generating a document using the PLSA algorithm model is to select the topic of the article first, and then determine the topic to generate the word. The difference from the LDA algorithm model is that the samples of the PLSA algorithm model appear randomly, and the parameters are unknown but fixed. The LDA algorithm sample is fixed and the parameters are unknown, but the parameters are not fixed and are a random variable. The steps to generate a document for the PLSA model are given below:

- The first step: select a random document \(d_j\) according to a certain probability \(P(d_j)\).
- The second step: Select the subject \(z_i\) of the document according to the probability \(P(z_i|d_j)\) in the document \(d_j\).
- The third step: selecting a word \(W_k\) of the subject according to the word frequency \(P(W_k|z_i)\) in the already selected topic \(z_i\).

Therefore, in the use of the PLSA algorithm model to determine the topic of the article is the selected word, backwards to derive the topic to which the word belongs, and finally determine the topic of the article. The probability plot of the PLSA algorithm model is shown in Figure 1.
Figure 1. PLSA algorithm model probability map

Where N is the number of words under the topic distribution and M is the number of articles. LDA is a topic model, which is a Bayesianization of the PLSA algorithm model. It adds a priori function when calculating the topic distribution and word distribution. The a priori function a priori distribution of the subject distribution a, and a word distribution of the a priori function β. It can give the subject words of the document in probabilistic form, so that by analyzing some documents to extract their themes (distribution), topic clustering or text categorization can be performed according to the theme (distribution). Then the probability formula for each word appearing in the article is as shown in Equation 1.

\[ P(a_i, w_i) = P(w_i | a_i) \times P(a_i) \] (1)

Therefore, the LDA model judges the subject of this article by counting the probability of each word appearing. Therefore, the LDA algorithm model probability map can be obtained, as shown in Figure 2.

Figure 2. Probability plot of the LDA algorithm model

Where φ represents the probability of the subject word of the article, θ represents the probability of the topic of the article, z is the topic distribution, N is the number of words under the topic distribution, and M is the number of articles.

3.2. Label hotspot model based on LDA algorithm

Since the content of Weibo is different from the normal document, the short text content of Weibo has its own characteristic values, such as the microblog tag "#", forwarding, comment, praise, and so on. Therefore, a topic analysis algorithm LH-LDA (label hotspot LDA algorithm) for microblog short text is proposed. The algorithm is a comprehensive algorithm model based on LDA algorithm, which is suitable for the mining of microblog topic keywords. The Bayesian probability distribution graph is shown in Figure 3. The micro-blog hotspot tag and attention degree are used to add additional feature values to the short text content of each microblog, and the theme distribution of the content is calculated according to the microblog content and its attached feature values. The Weibo content containing the Weibo hotspot label is classified into one category, and the Weibo content that does not contain the Weibo hotspot label is classified into one category. Then calculate the theme distribution...
using different calculation methods for different categories. The following is a definition of Weibo with Weibo hotspot tags and no hotspot tags:

1) Hot topic Weibo: refers to the microblog that contains two “#” hot topic tags in the Weibo content. This kind of Weibo can respond well to the subject matter of Weibo.

2) General topic Weibo: refers to the microblog that does not contain hot topic tags in the Weibo content. The Weibo may be a microblog that is original or forwarded by the user. This kind of Weibo content is often diverse and needs to be converted into a word vector to analyze its subject word distribution.

The process of the algorithm is introduced by the Bayesian probability graph of the calculation model.

First, the first step is to determine whether the short text of the microblog has a hotspot label, and if the microblog has a hotspot label, that is, a hot topic microblog. The relationship $\theta_i$ between the hot topic $H_i$ and the different topics is extracted from the distribution of the attention parameter $\eta_i$, and $\theta_i$ is assigned to $\theta_j$. If the Weibo does not have a hot topic tag, that is, the general topic Weibo, the relationship $\theta_i$ between the microblog and different topics is extracted directly from the subject distribution of the alpha parameter. In the entire algorithm model, the probability distribution formula of $\theta$ is shown in Equation 2 below.

$$P(\theta|\eta,\alpha,H) = \frac{P(\theta,\eta,\alpha,H)}{\sum_{T=1}^{T} P(\theta,\eta,\alpha,H)}$$

The formula for calculating the degree of interest $\eta$ [10], as shown in Equation 3:

$$\eta = \log_{10}(a) \times 0.1283 + \sqrt{c} \times 0.4082 + F \times 0.6827$$

Where $a$ represents the number of microblogs; $c$ represents the number of Weibo comments; and $F$ represents the number of microblog forwardings. Multiplying the value at the end of the expression is the result of a survey of 10,000 Weibo users. The heat value of Weibo can be approximated by the attention parameter $\eta$. At the same time, for the Weibo content, we use the word bag model without considering the relationship between the upper and lower semantics of the vocabulary content, so when the microblog text has $m$ words You can use Equation 4 to calculate the heat value of each word:

$$\eta(W) = \frac{\eta(W)}{m}$$

In the second step, the probability distribution formula of $\theta$ is solved. For the hot topic microblog, in the different topics included in the microblog, the theme $Z$ to which $\theta$ belongs and the word $W$ to which the theme $Z$ belongs are determined by the Dirichlet distribution of the parameter $\lambda$. For the general topic microblog, the generated word $W$ is extracted by the word polynomial distribution parameter $\varphi$. Finally, the joint distribution of vocabulary and subject results in Equation 5:

$$P(Z,W|\theta,\lambda,\beta) = P(\lambda)P(Z|\theta)P(W|Z,\theta)$$
The process of generating microblogs by LH-LDA algorithm model is used to inversely derive the distribution of microblog topic keywords. Therefore, the microblog hot topic analysis process diagram based on the LH-LDA algorithm model is shown in Figure 4:

![Figure 4. Microblogging hot topic analysis process diagram](image)

3.3. Application of LH-LDA algorithm in short text subject mining

The traditional LDA algorithm model solves the word distribution and can be based on the distribution $Z$ from the subject and the priori Dirichlet parameter $\beta$. Find the word distribution $\psi$ Formula 6:

$$P(w|z,\psi) = \prod_{i=1}^{W} P(w_i|z_i)$$

However, the proposed LH-LDA algorithm model estimates the unknown parameters differently from the traditional LDA variable EM algorithm to estimate the unknown parameters. The LH-LDA algorithm model derives unknown parameters using Gibbs Sampling, which is Gibbs sampling. Gibbs sampling [11] is a method for observing unknown parameters in a series of specified multidimensional probability distributions (joint probability distributions of multiple random variables) in Markov chain Monte Carlo theory (MCMC). For a document $D_i$ in the LH-LDA algorithm model, the attention parameter $\eta$, the Dirichlet distribution $\alpha$, the Dirichlet distribution $\beta$ and the hotspot label parameter $H$ are empirically obtained a priori parameters, while other variables The parameter word topic distribution parameter $Z$, the topic distribution parameter $\theta$, and the word distribution parameter $\phi$ are unknown variable parameters. The relational expression 7 of the word topic distribution parameter $Z$ under the vocabulary $W$ is derived by the word topic joint distribution formula, where $M$ is the number of words of the text:

$$P(Z_i = T | Z_{-i}, W_i) \propto P(Z_i = T, W_i = M | Z_{-i}, W_{-i})$$

Finally, the formula 8 of the distribution parameter $\theta$ of the microblog in the subject relationship and the formula 9 of the distribution parameter $\phi$ of the subject in the vocabulary are obtained by stepwise iterative sampling:

$$\phi = \frac{\sum_{\theta} \sum_{\psi} M_{\psi}^{(i)} + \beta + H_{T} + \eta_{e}}{\sum_{\varphi=1}^{V} \sum_{\psi} M_{\psi}^{(i)} + V \beta + H_{T} + \eta_{e}}$$

$$\theta = \frac{\sum_{\theta} \sum_{\psi} M_{\psi}^{(i)} + \varphi + H_{T} + \eta_{e}}{\sum_{\varphi=1}^{V} \sum_{\psi} M_{\psi}^{(i)} + V \beta + H_{T} + \eta_{e}}$$
The improved LDA algorithm can better calculate the microblog with hotspot tags and heat values, and can find the keyword that best represents the content of the text in the short text of Weibo.

In addition, the LH-LDA algorithm can be applied not only to the extraction of keyword texts in short texts, but also to the filtering of spam messages and emails. Through the analysis of SMS and email, find the keywords that best represent SMS and email, and filter SMS and email according to the keyword.

4. Experiment and result analysis

4.1. Short text dataset and data preprocessing

The short text dataset collected in this article is a microblog text content obtained by simulating web browsing using python's selenium library and using Xpath. The microblog short text data set collects 229,900 microblog contents from January 1, 2018 to November 1, 2018. The key words of the data set are mined using the LH-LDA algorithm.

The data preprocessing is first to clean the data collected by the microblog short text. Data cleaning is the deletion of characters and text that are not related to the research content. The data collected in this article is the text collected by the crawler, which may contain URLs, emoticons, numbers, English, Japanese, Korean, etc., so the regular expression is used to clean the collected microblog text. Keep only the text content. At the same time, in the process of cleaning the data, the microblog text with characters less than 5 characters is deleted, because the text is too short to display the theme content well.

Then, the cleaned data is classified. Contains the characters "#....#", that is, hot topic labels, which are classified into one category, as shown in Table 1; the Weibo content that does not contain the characters is divided into one category, as shown in Table 2. Then, the text contained in the middle of the character "#....#" is taken as a hot topic by a specific character extraction, as shown in Table 3.

Table 1. Weibo with hot topic tags

| Hot Topic       | Content Description                                                                 |
|-----------------|--------------------------------------------------------------------------------------|
| #Wang Junkai#   | 18 years of desire for the college entrance examination smoothly and like people to test an ideal university together, and he is not far from the two or three friends to see a relatively stable life, see Wang Junkai |
| #Pure art#      | Wimbledon's painting alumni @YUCHU's rich little expressions, very playful oh#London Art University UAL# |
| # Colorful Garden# | On the celebration party, Miss Chu Yuxi’s vibrating video was praised 80W a day, and videos from universities such as Tsinghua University and the People’s Public Security University were rated as excellent campus videos. |
| #new skill get# | How to pass the College English Level 4 and College English Level 6 exams in one month? Need to take the test of the students of the 4th and 6th grade |
| # Renmin University of China # | # Rookie return plan # A campus environmental awareness campaign advocating recycling and reuse of waste packagings was launched at Renmin University of China in Beijing recently. |
Table 2. Weibo without hot topic tags

| The New Year’s Day is more than a year. I have to go to the University for the New Year’s Day. I have to review the exams. |
| Today, I am very happy at noon. I have a good time at the grandmother’s house. I have a very good grapefruit super sweet afternoon. I’m playing a mobile game with my brother. You’re still like a junior high school. |
| Hard work, small fresh meat arrived early in the morning to pick up the university friend Xixi in the light valley small gathering |
| Have your friends recently indulged in a game called jumping? Let’s let the woods that have passed through the battle to tell everyone how to get the high scores. |

Table 3. Hot topics extracted

| Tsinghua University |
| Campus things |
| Warm video |
| I am a scientist |
| a moment of ease |

Secondly, the Python library of Tsinghua University's word segmentation tool THULAC [12] is used to perform word segmentation on the preprocessed data set. Then use Harbin Institute of Technology's stop word list to remove the stop words from the data set after the word segmentation. The appearance of stop words has a greater impact on the short text of Weibo, because the number of words in Weibo content does not exceed 140 words, and the number of stop words occupies a certain amount, they do not help the topic analysis.

4.2. Vectorization of words

The tool for the word vector used in this article is Word2vec. Word2vec is a toolkit for Google to acquire word vectors for open source in 2013. Due to the high sparsity of Weibo content, and the LH-LDA algorithm model is a word bag model, the word vector used to construct Weibo texts with Word2vec is more accurate. Using Word2vec to perform word frequency statistics on the preprocessed data set, the word frequency is formed into a word frequency matrix, and the vectorization operation of the data set is completed. The result is shown in Table 4, which is the word frequency vector of each word in all documents.

Table 4. Data set conversion to word frequency vector results

| (230125, 54857) |
| (230125, 42018) |
| (230125, 143516) |
| (230125, 45527) |
| (230125, 4599) |
| (230125, 70529) |
| (230125, 151658) |
4.3. Microblogging hot topic analysis experiment of LH-LDA algorithm model

The experimental environment used in this article is the Intel Core i7-8700@3.20Ghz CPU, 32GB of RAM PC. The experimental tool used is python 3.7.2.

First, the a priori Dirichlet parameters $\alpha$ and $\beta$ are determined according to the LH-LDA algorithm model derived above. For the traditional LDA algorithm model, set $\alpha = 1 / K$, $\beta = 1 / K$. For the LH-LDA algorithm model of this paper, $\alpha=50/K$ and $\beta=0.01$ are set according to the empirical value, where $K$ represents the number of topics. Here we have a default number of topics of 10. The traditional LDA algorithm and the LH-LDA algorithm model have a default iteration of 1000 times. Table 5 lists the Top 10 topics after the LDA algorithm and the LH-LDA algorithm process the data set.

| Algorithm model | Top10 topic |
|-----------------|-------------|
| LDA             | 1. Education  
2. Graduation  
3. Play  
4. Weibo  
5. Games  
6. University  
7. Student  
8. Campus  
9. World Cup  
10. Teacher |
| LH-LDA          | 1. Game  
2. Work  
3. High school  
4. Youth  
5. Admissions  
6. Education  
7. Graduation  
8. Campus  
9. Dormitory  
10. World Cup |

Through the experimental results, it can be found that the LDA and LH-LDA algorithms are based on the theme of “university students” in general theme types, which proves that the analysis of microblog topic topics using LDA algorithm is correct. The calculation formula 10 of uncertainty is introduced below, where $u$ represents an uncertain subject vocabulary, and topic represents the total number of vocabulary words obtained:

$\text{Uncertain}(\text{topic}) = \frac{\sum_{n=1}^{U} u}{\sum_{n=1}^{\text{topic}}}$  \hspace{1cm} (10)

From the results of the traditional LDA algorithm model calculation data set, the words "Weibo", "Student" and "Teacher" show the "university" theme, but its uncertainty is relatively large, so it is not Can more accurately derive the keywords in the dataset. The Top10 keywords obtained by the LH-LDA algorithm are subject words related to college students, and the results calculated using the LH-LDA algorithm are less uncertain, and the obtained keywords are more accurate. Table 6 shows the uncertainty of the LDA and LH-LDA algorithms.

| Algorithm model | uncertainty |
|-----------------|-------------|
| LDA             | 0.4         |
| LH-LDA          | 0.2         |
4.4. Analysis of results for qualified topic content

This paper analyzes the accuracy of different algorithms from a given limited theme “Xi'an Smoking”. Collected 8,000 Weibo data through python crawlers. Firstly, the microblog content is manually labeled, and then the data is processed by the LDA and LH-LDA algorithms respectively. The final result is shown in Table 7.

Table 7. Qualified topic LDA, LH-LDA algorithm solution results.

| Calculation method | Topic Words                                      |
|--------------------|--------------------------------------------------|
| Manual labeling    | Xi'an, Non-smoking, Comprehensive, Public Places, Health, Smoking, Smoke-free, Control, Management |
| LDA                | Method, Xi'an, Non-smoking, Tourist, Public places, Smoking, caveat, fine, control, country |
| LH-LDA             | Method, Xi'an, Non-smoking, Tourist, Public places, Smoking, caveat, fine, control, country, The strictest |

The results of the processing of different algorithms are compared with the keywords that are manually labeled. It can be found that the traditional LDA algorithm selects a large number of vocabulary words, but does not consider whether the vocabulary itself belongs to the hot tag content, and does not consider the heat value of the vocabulary, so there will be some irrelevant subject vocabulary, such as "The tourists", "the most strict" and other factors affecting the subject matter. LH-LDA processes a large amount of microblog data, and labels and heats the vocabulary of the word segmentation so that each vocabulary has two attributes. Therefore, from the experimental results, the difference between LH-LDA and manual labeling is no more than 30%. Through the above analysis of the microblog texts of the undefined subject and the microblog text analysis of the limited topics, it can be seen that the annotation of the vocabulary by the LH-LDA algorithm to calculate the keywords of a large number of microblog short texts can make up for the lack of context semantics of the microblog content. High sparsity and other characteristics, better mining of microblog short text content.

5. Conclusion

Due to the particularity of microblog short text, that is, the sparsity is high, the text content is simple, and the context semantics are lacking, the traditional standard LDA is not very effective for the discovery of keyword words in microblog short text content. Aiming at this problem, this paper proposes an improvement of LDA algorithm based on microblog hotspot label and heat value. Combining Word2vec algorithm to convert vocabulary into word frequency vector, the microblog hotspot label and heat value are given different labels to microblog vocabulary, effective the keywords of the microblog short text are calculated, and the algorithm can be used for various types of short text topic analysis. Second, the algorithm converts a large amount of high-sparse short text content into low-sparse text content. The LH-LDA algorithm can be applied to a variety of short text heat analysis and keyword analysis, such as SMS, email, Weibo text, video barrage, etc. for subject extraction analysis. Future research content will continue to optimize short text content processing and LH-LDA algorithm optimization. Also consider adding parallelization operations in the data processing process to speed up the short text processing.

References

[1] Zhou Fuxing, Chen Xiuzhen, Ma Jin, Li Shenghong. A new method for mining hot topic topics based on tag semantics[J]. Computer Engineering, 2018: 1-11.
[2] Huang Chang, Guo Wenzhong, Guo Kun. Research on Improved BBTM Model for Hot Spots of Microblogs[J]. Computer Science and Exploration, 2018: 1-14.
[3] Ma Xiaoning, Wang Hui. Topic trend prediction based on BP neural network optimized by PSO[J]. Computer Engineering and Design, 2018, 9: 036

[4] Wu Wankun, Wu Qinglie, Gu Jinjiang. Hot topic discovery of e-commerce microblog based on EM-LDA synthesis model[J]. Data Analysis and Knowledge Discovery, 2016, 31(11): 33-40.

[5] Zhang Chenyi, Sun Jianbiao, Ding Yiqun. Mining of Weibo Theme Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.

[6] Li Hui, Wang Liting. Research on Hot Spots of Microblog Based on Term Heats[J]. Information Science, 2018, 36(4): 45-50.

[7] Li Jing, Yin Jian, Liu Shaopeng, et al. Mining of Weibo Theme Based on Topic Tag[J]. Computer Engineering, 2015, 41(4): 30-35.

[8] Shang Qingxia. Chinese microblog hot event detection and sentiment analysis based on TH-LDA model [D]. Southwest University, 2017.

[9] Ishikawa S, Arakawa Y, Tagashira S, et al. Hot topic detection in local areas using Twitter and Wikipedia[C]//ARCS Workshops (ARCS), 2012. IEEE, 2012: 1-5.

[10] Yan Yueming. Evolution analysis of hot topics on Weibo based on topic model [D]. Xidian University, 2017.

[11] HAN Dong, WANG Chunhua, XIAO Min. Text Classification Method Combining Semi-supervised Learning and LDA Model[J]. Computer Engineering and Design, 2018, 39(10): 3265-3271.

[12] Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, Zhiyuan Liu. THULAC: An Efficient Lexical Analyzer for Chinese. 2016.