Improvement of LDA Topic Mining Algorithm and Its Application in Short Text

Kai Li, Chunmei Li*
Department of Computer Technology and Application, Qinghai University Xining 810016, China +8618595841812, +8613897207231
*Corresponding Author
likai614020758@126.com, li_chm0422@sina.com

ABSTRACT In order to quickly provide a large number of short text themes, an improved linear discriminant analysis (LDA) topic mining algorithm is proposed in this paper. Firstly, the acquisition method of obtaining the traditional short text themes is first analyzed. Focusing on the shortcomings of the original algorithm, the improvement process of the improved topic mining algorithm based on LDA is introduced. Finally, through the mining of short text themes by college students as an example, the improved LDA algorithm is improved in the accuracy of short text processing.

1. INTRODUCTION
In today’s information age, the development of the Internet can bring a lot of information growth. How to find valuable information in the massive information, organize, manage, and clearly present it to users to help users better use the network information has become an important research topic. In the research of news commentary and similar short-text clustering, some scholars focus on the k-means clustering algorithm’s insignificance of news comment data, select the initial point by constructing the comment similarity matrix and improve the classification method, and use the cosine distance [1, 2]. In addition, some scholars have analyzed the labels of short texts and used the standard LDA algorithm for cluster analysis [3]. Some scholars have used feature vectors to calculate hot topics in a large number of MicroBlog short texts, and used TD-IDF (term frequency–inverse document frequency) to solve word frequency and word vectors [4]. Although extensive research has been done on the topic mining of traditional texts, traditional text mining algorithms cannot well model special short texts [5].

Extracting assessment topics/objects from the results of topic mining is more suitable for news reviews. This is because the topic mining algorithm based on short text clustering is easy to get the comment topic [6]. Some scholars have proposed preprocessing techniques based on the sparseness, multidimensionality and large-scale nature of MicroBlog information, and used LDA for topic mining [7, 8]. Some scholars have found in the news commentary on integrated LDA and clustering that more noise data in the comments has a great influence on the center point of K-means. Outliers can severely distort the distribution of data, and the squared error function can severely degrade the effects, so k-medoids clustering is used [9,10]. Due to the defects of the LDA algorithm itself, the program overflows and the slow speed caused by the high dimension of the short text matrix, and the LDA is more and more prominent in dealing with the segment text problem. Based on this problem, this paper proposes an improved algorithm based on LDA for a series of MicroBlogging topics, which is
combined with the text structure of short text content, and improves the LDA algorithm to find short text keywords of MicroBlog's data.

2. TRADITIONAL LDA THEME MODEL

The LDA topic model is a generation model. Generating a model refers to a given article, determining each word to start with an article, and using a "document-theme" probability distribution to select a topic and selecting a word with a probability distribution based on the topic. Repeat the selection of words with a probability distribution from which the article is generated. Then the probability of the occurrence of each word in the document is calculated as follows:

\[ P = \sum P \cdot P \]

LDA is a three-layer Bayesian probability generation model in which two model parameters "word subject" and "topic document" need to be solved. The probability map of the LDA probability topic model is shown in Figure 1 below.

![Figure 1. LDA probability map model](image)

\( \Phi \) represents the "subject" probability model, which is a Dirichlet distribution with hyperparameters \( \beta \). \( \theta \) represents the "document subject" probability model, which is a Dirichlet distribution with hyperparameters \( a \). \( T \) represents the number of topics. \( D \) represents the number of documents, and \( N \) represents the word length of the document.

By calculating the heat of the word, the corresponding topic heat is calculated, and the topic of the short text is sorted on this basis. The calculation of the topic's hotspot is as follows:

\[ H(T) = \sum_{T} T_{w}(w \in T, w \in d) \]

3. IMPROVEMENT OF TOPIC MINING ALGORITHM BASED ON LDA

Since text is treated as an unordered word package model in the semantic analysis processing, all short texts of the same time period can be linked together and treated as text processing blocks. In this way, the text space can be reduced, the features of the original text can be expanded, and the similarity of the text can be improved, thereby solving the sparseness and abnormality of the short text to a certain extent. However, it has a high repetition rate. By simplifying the function, the text with the high coincidence rate is changed to the same text as the general text, which is advantageous for the effect of the LDA processing. The algorithm is improved according to the characteristics of the LDA ignoring the word order in the document and the corpus, as follows:

Step 1: the comment of the same paragraph is turned into a text block with a total of \( m \) text blocks.
Step 2: the first paragraph is represented by all the characteristic words T. Corresponding word frequency is WF: WF (wf1, wf2, … wk). Then express this paragraph with L1 = (log*T1|log*T2|…log*Tk).

Step 3: then repeat step 2 to get L2, … Lm.

Step 4: Use the short text of a total of m lines as input to the LDA's Single-pass incremental clustering algorithm. The improved LDA's Single-pass algorithm usually selects the first document as the center of the first category, and compares the similarity of the two texts with the next document similarity as the initial cluster. The theme center is recalculated each time when a new topic is generated, or when a topic is added to a new document. The cluster center is the average of the document vectors in the cluster. The improved algorithm flow chart is shown in Figure 2.

Figure 2. Improved Single-pass algorithm flow chart

4. MINING QUALITY TOPICS FROM THE CORPUS

4.1 Application of Improved LDA Topic Mining Algorithm in Short Text
The corresponding number of topics will be automatically mined based on the number of clusters pre-entered manually. These topics are automatically obtained based only on the co-occurrence relationship of words in the corpus. In fact, not all topics consisting of related words can represent a real subject. Table 1 below is the related words of the wedding theme in the corpus, and the topics are found according to the co-occurrence relationship. It can be seen that the top seven words with the highest frequency of the theme are "wedding", "plan", "wedding ceremony" and so on. These words are a good representation of the theme of the wedding, and there is no word "Wedding plan" related to the theme, so this theme is a quality topic.

Table 1. High quality theme (wedding theme)

| Key Words     | Wedding | Plan | Wedding Ceremony | Ceremony |
|---------------|---------|------|------------------|----------|
| Probability   | 0.0918  | 0.070| 0.0310           | 0.0202   |
| Key Words     | Event   | Celebrat | flower | … |
| Probability   | 0.0185  | 0.0153| 0.0144           | …       |
4.2 Mining Noise Themes in the Classification Corpus

Entering the number of topics in the algorithm, here set the number of subjects K to 8. The corpus of different topics are tested, and a non-centered topic with a relatively low frequency of corresponding keywords with a minimum degree of relevance is found in the classified corpus, as shown in Table 2 below. It can be concluded that the words “centre” and “service” are the most frequently used words in this topic. However, common words such as "center" and "service" usually do not represent the subject, and the probability distribution is not uniform. The probability of only "center" is 0.3277. Second, in its high-frequency words, there are two different themes, "shopping" and "Internet cafes." Therefore, the topic is a noise theme that cannot be classified as a topic in the text and should be classified as a non-central theme.

| Key words       | Center   | Service | Shopping | Internet cafe |
|-----------------|----------|---------|----------|--------------|
| Probability     | 0.3277   | 0.0503  | 0.0345   | 0.0083       |

4.3 Feature Word Selection

The data source in this test experiment is the 1 million 400,000-dimensional data of Sina Weibo, which will be built in the experiment. We find the probability of the first 200 words under each topic and the probability between 20% and 75%. Can fully represent the corresponding theme. Accordingly, in the screening process of the high quality theme of the LDA model, the first 200 feature words in the ".twords" suffix file output by the LDA model are processed only according to the frequency of each topic.

4.4 Similarity Calculation

The Sim function is defined for independent supervision of each topic. Mainly to calculate the similarity between the two representative words of the subject, that is, the accumulation of the probability product of the co-occurrence words between the two representative words of the subject. In the actual calculation, the first 200 feature words are used to calculate the similarity, and the formula is as follows:

$$
\text{Independence}(i,m) = \sum_{j=1}^{N} \text{Sim}(\text{topic}_i, \text{topic}_j, m)
$$

N is the total number of topics, i.e. the number of artificially predefined LDA clusters; J is the subject of the first document; L is the subject of the second document, and m is the number of texts for the two topics.

4.5 Choice of High Quality Theme

Computation of similarity although clustering operations can be performed, the word frequency distribution of high quality topical representation words are usually unbalanced. A few core words appear more frequently. However, the representation word s of a noise subject are usually composed of some random words, and the correspondence between these random words and high frequency words is small. To calculate the coverage of feature words, that is, the sum of the calculated probabilities, the choice of high quality topics can be achieved. Therefore, it is necessary to set a certain threshold, and some low-probability eigenvalues are deleted by filtering the threshold. If the threshold is already sorted on this basis, the hotspot formula for the topic is as follows:

$$
\text{Coverage}(i,m) = \sum_{j=1}^{n} wp_{i,j}
$$
WPi,j is the probability of the jth character from the topic.

Secondly, considering the probability distribution of feature words, if the volatility is too small, that is, the probability distribution of all feature words is too balanced, most words are composed of random words, lacking the subject core words, such topics can not be used as high quality. On the other hand, if the fluctuation is too large, that is, the probability distribution between the feature words is not balanced, this theme is mainly composed of a few general words as high-frequency words, and there is no difference between the topics. For example, non-central themes of high-frequency words, such as "services" and "centers", cannot be considered high-quality topics. So the variance is used to measure the deviation between the random variable and its mathematical expectation, which is used to measure the volatility of a set of data. Therefore, the variance of a high quality theme should fall within a range of fluctuations. The calculation formula is expressed by Equation:

\[
\frac{x(1,m)}{m} = \sum_{j=1}^{m} w_{p_{i,j}}
\]

Entropy was originally a concept in thermodynamics, a physical quantity used to express the degree of disorder of molecular states, and then introduced into information theory to express the order of the system. The lower the information entropy, the less information is contained. The information entropy has an accurate formula, and the corresponding information entropy of the probability sequence is calculated as Equation:

\[
H(P_1, P_2, ..., P_n) = -\sum_{i=1}^{n} P_i \log(P_i)
\]

When choosing each high-quality theme, from an information theory point of view, if all the words appearing are considered to be representation words, they are often stable to the whole system, and the smaller the information entropy of the word sequence is better. Similarly, if only a small number of core feature words are selected to calculate the information entropy, since the probability of the core word is high and the probability distribution between the core words is not balanced, the higher the information entropy of the sequence, the better.

5. ALGORITHM IMPLEMENTATION

5.1 Experiment Setup
In this experiment, the word segmentation tool is used for microblog text, complete word deletion and perform part-of-speech annotation work. Sort the text according to part-of-speech annotation, calculate the heat of microblog, and make preliminary selection according to heat. A Single-Pass and improved LDA algorithm is used for the user to select the subject number K value, the similarity threshold, and the number of microblogs per batch to facilitate topic merging in text clustering.

5.2 Database Design
According to the analysis in the previous chapters, five database tables are designed in this experiment: the microblog data table is used to store the text content and heat value of Weibo, the heat value is composed of forwarding, comment, and praise; the text direction scale table is used for storage. Processing the vector value composed of the text content; the theme table is used to store the keyword used to process the data using the improved algorithm; the user table is used to store the user name, region and gender obtained when the text content is obtained; the log table is used to store the processed data generated Logging.

5.3 Data Sources
The test data source of this experiment data is Sina Weibo. From January 1st, 2018 to November 1st, 2018, all Weibo on the topic of "college students" and no central theme.
5.4 Experimental Results

Use Python's re matching characters to clean out the English letters, numbers, punctuation, and special symbols in the Weibo content. Then the data of the cleaning is performed using the THLAC library of Tsinghua University [11] and the stop word list of Harbin Institute of Technology for word segmentation. Use the IDF method to vectorize the data after the word segmentation. The data after vectorization of the words is calculated separately, one part is the data with the subject, and the other part is the data without the subject. The results of the final processing of the data with the subject are shown in Table 3 below.

Table 3. represents the center theme (college students)

| Subject Words | College Student | Game | Work | High School |
|---------------|----------------|------|------|-------------|
| Probability   | 0.5718         | 0.2102 | 0.1452 | 0.0311 |

| Subject Words | Youth | Graduation | World Cup |
|---------------|-------|------------|-----------|
| Probability   | 0.0227 | 0.0100 | 0.0090 |

Table 3 shows that the two most frequent keywords are the vocabulary of the topic specified in the collection of data, "college student" and "game" Therefore, in terms of the improved algorithm, the accuracy of keyword extraction is higher for short text content with themes. Then, the collected MicroBlog data without the center theme is processed. Can get the following table 4:

Table 4 represents a no-center theme

| Subject Words | Victory | Trade | Asian Games | China |
|---------------|---------|-------|-------------|-------|
| Probability   | 0.4302  | 0.3281 | 0.1101      | 0.0514 |

| Subject Words | USA | Commerce | ZTE |
|---------------|-----|----------|-----|
| Probability   | 0.0349 | 0.0677 | 0.0221 |

Observing Table 4, we can see that the topic-free corpus is about the topic of "Asian Games wins" and "China-US trade" "ZTE". By manually searching for 10 months of topics, these three types of topics are indeed in 10 Among the most discussed topics since the month, the improved LDA algorithm improves accuracy in the processing of short text. Reduce the impact of irrelevant vocabulary

6. CONCLUSION

An improved LDA topic mining algorithm is proposed in this paper. It can improve the shortcomings of the original LDA algorithm applied in the short text data. Aiming at the special text structure of microblog short text, an improved LDA algorithm is proposed to effectively discover the subject words of short text topics. In theory, it is first discovered that the improved LDA subject evolution algorithm can be applied to a series of microblog heat analysis. Secondly, it is found that unlike the evolutionary features of low-dimensional dense vectors, for high-dimensional sparse short text vectors, it needs to be converted into a file structure close to the microblog text. This provides an improved direction for short text LDA topic mining: simplifying a large amount of short text data into general low-dimensional text. In practice, the topic extraction model can be applied to short text heat analysis, such as microblogging, post bar, chat history, and video barrage. Improved algorithms facilitate topic mining due to high precision, which helps managers and decision makers make use of critical information to make decisions.

REFERENCES

[1] Zhao D, He J, Liu J. An improved LDA algorithm for text classification. International Conference on Information Science, Electronics and Electrical Engineering. IEEE, 2014, pp. 217-221.
[2] Zhang C, Yang M. An Improved Collaborative Filtering Algorithm Based on Bhattacharyya Coefficient and LDA Topic Model. International CCF Conference on Artificial Intelligence. Springer, Singapore, 2018, pp. 222-232.

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

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

[5] Wang D, Wang S. Improved 2DLDA Algorithm and Its Application in Face Recognition. IEEE, International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, 2015, pp. 707-713.

[6] Qiuqiu L I, Yang H, Feng J, et al. Face recognition algorithm based on 2DPCA+2DLDA and improved LPP. Computer Engineering & Applications, 2015, 90(11), pp. 777-81.

[7] Ma S, Jiang Z, Zhang T. The improved multi-scale Retinex algorithm and its application in face recognition. Control and Decision Conference. IEEE, 2015, pp. 5785-5788.

[8] Hu D, Chen L. A High Efficient Recommendation Algorithm Based on LDA. Human Centered Computing. Springer International Publishing, 2016, pp. 668-675.

[9] Fralenko V P. Localization of text fragments on mixed background: short scientific review. Programmnye Sistemy Teoriya I Prilozheniya, 2014, pp. 33 – 45.

[10] Lee J Y, Dernoncourt F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. 2016, pp. 515-520.

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