Hindawi
Complexity
Volume 2018, Article ID 2503816, 10 pages
https://doi.org/10.1155/2018/2503816

Research Article

CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background

Lirong Qiu and Jia Yu

School of Information Engineering, Minzu University of China, Beijing, China

Correspondence should be addressed to Lirong Qiu; qiu_lirong@126.com

Received 5 March 2018; Accepted 11 April 2018; Published 16 May 2018

Academic Editor: Zhihan Lv

In the present big data background, how to effectively excavate useful information is the problem that big data is facing now. The purpose of this study is to construct a more effective method of mining interest preferences of users in a particular field in the context of today's big data. We mainly use a large number of user text data from microblog to study. LDA is an effective method of text mining, but it will not play a very good role in applying LDA directly to a large number of short texts in microblog. In today's more effective topic modeling project, short texts need to be aggregated into long texts to avoid data sparsity. However, aggregated short texts are mixed with a lot of noise, reducing the accuracy of mining the user's interest preferences. In this paper, we propose Combining Latent Dirichlet Allocation (CLDA), a new topic model that can learn the potential topics of microblog short texts and long texts simultaneously. The data sparsity of short texts is avoided by aggregating long texts to assist in learning short texts. Short text filtering long text is reused to improve mining accuracy, making long texts and short texts effectively combined. Experimental results in a real microblog data set show that CLDA outperforms many advanced models in mining user interest, and we also confirm that CLDA also has good performance in recommending systems.

1. Introduction

In the background of today's big data, it is an important part of the enterprise activity planning to accurately excavate the interest preference of the user-specific fields from the large data. Nowadays, the emergence of a social network represented by microblog makes a large number of users more willing to use it to share their interest in various fields. Microblog platform will provide a large number of user emotion data, which can be used to mine the user's interest preferences in specific areas. Therefore, a lot of user data on the microblog platform can effectively mine the user's interest and bring huge commercial value.

At present, most of the research on microblog is based on the analysis of the relationship between users and the community [1], and few studies are on microblog content. Traditional text mining algorithms are mainly used in traditional corpus and do not take into account the special structural information contained in the microblog data texts, so we cannot model the microblog data text very well. Topic models are an effective method of text mining, but traditional topic models such as pLSA [2] and LDA [3] all learn potential topics in the corpus by developing words from the document. As a result, topic models often suffer from severe data sparseness problems when applied to microblog short text. A popular and effective strategy is to overcome this bottleneck by aggregating short texts into long texts based on user information, title categories, and so on [4, 5]. However, these methods are heuristic and highly dependent on the data. In addition, such aggregated long text content is excessively redundant, reducing the accuracy of mining user interest preferences. Therefore, this article is based on these previous studies. In terms of text processing, we combine the characteristics of microblog with introducing time dynamics for each user's short text, short text extensions using user-generated short texts, and information retrieval tools. On the short text and long text problems facing, we propose Combining Latent Dirichlet Allocation (CLDA), a new topic model that can learn the potential topics of microblog short texts and long texts simultaneously. The data sparsity of short texts is avoided by aggregating long texts to assist in learning short texts. Short text filtering long text is reused to improve

mining accuracy, making long texts and short texts effectively combined. We borrow the Gibbs sampling method to derive our model. The experimental results show that this model is superior to many other advanced models in mining interest preference in specific fields in a large number of microblog user data.

The main contributions of this paper are as follows:

(1) A long text acquisition scheme is used to aggregate microblog short text in a user unit and expand microblog short text by using the information retrieval tool.

(2) The dynamic time attribute and Ebbinghaus forgetting curve are integrated into microblog short text, which makes it more reasonable to mining users' interest preferences.

(3) A new topic model, CLDA, is proposed to learn the potential topics of microblog short text and long text at the same time. By using long text to assist the learning task of short text, the short text data is avoided, and the short text is used to filter long text to improve the mining accuracy.

(4) The experimental results demonstrate the superiority of the proposed method and can be extended to recommendations in various fields.

The outline of this paper is as follows: The relevant work is briefly reviewed in Section 2; Section 3 briefly introduces the LDA topic model; then the method we proposed is introduced in detail in Section 4; the experimental results of the actual Sina Microblog data set are given in Section 5; and finally in Section 6 we draw conclusions and introduce the next work.

2. Related Work

The predecessor of probability theme model can be traced back to LSA (Latent Semantic Analysis) [6]. LSA is based on spatial dictionaries, and implicit semantic documents are implemented in low-dimensional representation of space, but it cannot solve the problem of polysemy. Hofmann proposes a PLSA (Probabilistic Latent Semantic Analysis) [7, 8] for the defect of LSA, mainly using the probability distribution corresponding to one dictionary in each dimension. However, PLSA does not provide a probabilistic model at the document level, which leads to overfitting problems easily due to the linear increase in the number of parameters to be estimated in the model with the size of the corpus. LDA (Latent Dirichlet Allocation) [3] is a generation model that uses the Dirichlet a priori distribution of topics to overcome the shortcomings of PLSA. The model can find the semantic structure of the text set, mining the theme of the text.

Xiong et al. [9, 10] used the existing user interest modeling method LSARS to represent the user's interest and fused the LDA topic model with the geospatial attributes to overcome the data sparsity with user interest mining. Specifically, they first divided geospatial attributes into subregions where one's personal interests can be inferred from a set of topics. With regional and thematic interdependence, LSARS combines geopolitical clustering and LDA thematic modeling into a single process. In order to further reduce the data sparsity of user behavior, LSARS further integrates the crowd's preference [11]. However, these approaches that incorporate geolocation attributes are only applicable to mining local user interest preferences, and the effects applied to other areas are less than ideal. Our approach is to incorporate time dynamics, independent of the geographic location of the space.

In short text processing, many strategies have been widely used in data mining tasks, especially query extensions with relevant feedback [12, 13], semantic correlation analysis [14, 15], short text classification [16, 17], and interest extraction [18, 19]. However, short texts often have large data sparsity and often do not work well when decimated. Tang et al. [20] proposed an end-to-end solution to the short text sparseness and automatically learned how to extend the short text to optimize a given learning task. A novel deep memory network was proposed to automatically extract relevant information from a long list of documents and to redesign short texts through gated mechanisms to avoid short text sparsity. The expanded text usually takes the form of an interpolation between the original short text and the retrieved document before it is used for other tasks. These methods are only intended to solve the sparseness of short text data and ignore the retrieved documents that contain noise, and the interpolation weights are heuristically set, so these errors may accumulate in the task and compromise the accuracy of the final task result. The difference compared to our work is that we use long texts to help short texts perform learning tasks and thus overcome the data sparsity of short texts. Short texts, on the other hand, can filter extended data sets to reduce noise interference and greatly improve the accuracy of the final result.

Since the LDA topic model was proposed, it has received extensive attention from researchers. Many scholars have continuously improved the LDA model to achieve the desired topic mining effect. RosenZvi et al. [21] proposed the Author-Topic Model (ATM) to aggregate user tags into a large document and analyze and model the user's interests. Ramage et al. [22, 23] proposed that Labeled-LDA modeled the topic of microblog texts, used tags to associate topics with tags, and implemented supervised learning text topics. Weng et al. [4] proposed User-LDA to merge all the text content published by each user on Twitter into one large long text and then use the standard LDA model to extract the user interest on the long text. Zhao et al. [24] believe that microblog is relatively short and proposes a Twitter-LDA that only learns 1 topic for all words in the microblog. However, on a highly dynamic social platform such as microblog, new topic appears constantly and user interest preferences are constantly changing. The topic mining of aggregating long texts does not well represent users' dynamic interest preferences. Each short text of the user also contains user interest information. Therefore, we propose CLDA, a new topic model that can learn the potential topics of microblog short texts, and long texts simultaneously which avoids the data sparsity of short texts by aggregating long texts to assist in learning short texts. Short text filtering long text is reused

(i) For each topic $z \in \{1, \ldots, K\}$,
Choose a multinomial distribution over terms, $\varphi_z \sim \mathrm{Dir}(\beta)$.
(ii) For each document $d \in \{1, \ldots, M\}$
(a) Choose a multinomial distribution over topics, $\theta_d \sim \mathrm{Dir}(\alpha)$
(b) For each of the N words w_n
Choose a topic $z_n \sim p(z \mid \theta)$;
Choose a word $w_n \sim p(w \mid z)$;

Algorithm 1: The generation process of LDA.

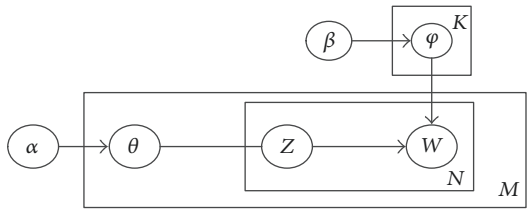

Figure 1: Graphical representation of LDA.

to improve mining accuracy, making long texts and short texts effectively combined.

3. Latent Dirichlet Allocation (LDA)

In the field of topic discovery, public opinion analysis, text categorization, and so on, LDA topic model has become a common way to catch the distribution similarity between words (vocabulary, semantics, or even syntax). The basic idea is to give each document set in the form of probability distribution and to extract the theme distribution through analyzing some documents. After the theme clustering, theme distribution or text classification can be carried out. At the same time, it is a typical bag-of-words model; that is, a document consists of a set of words, and there is no order relationship between each word. In addition, a document can contain multiple topics, each of which is generated by one of the topics. The document generation process is described in Algorithm 1, which corresponds to the graph model shown in Figure 1, where the variables φ and θ and z (assigning word labels to topics) are three sets of potential vectors to infer, with each column of the vectors indicating the probability of each topic occurring in the document, which is a nonnegative normalized vector. As mentioned earlier, hyperparameters α and β are constants in the model and need to be set manually. w_n represents the nth word generated w; z_n represents the selected theme, $p(z \mid \theta)$ represents the probability distribution of the theme z given θ; $p(w \mid z)$ ia similar. The arrows in the figure indicate the conditional dependencies between the variables, while the boxes in the figure refer to the repeat sampling step, with N and M indicating the number of samples. A box around φ means that the word distribution is repeated for each topic z until k topics have been generated.

From the generation of the LDA topic model, we can see that LDA is an unsupervised machine learning technology that can be used to identify the hidden topic information in large-scale document collections or corpus. It is an effective method for mining large data texts. However, the microblog information is characterized by shortness, few representative words, and so on. The application of LDA directly to the short text such as microblog will not play a good role. Since the LDA topic model is a bag-of-words model and does not consider the order between words, it simplifies the complexity of the problem and provides opportunities for improvement of the model.

4. Proposed Method

4.1. The Reason for the Proposed Method. Microblog's large amount of user data information is characterized by shortness and less representative words, so it is much more difficult to learn the topic model directly from the traditional long text in the short text of microblog. For this reason, many scholars have proposed to train topic models on a syndicated long text in the same field and then infer the essay to help short text learning tasks [25, 26]. However, on the highly dynamic social platforms such as Weibo, new topics are constantly appearing and user preferences change constantly. Therefore, it is particularly important to better grasp the preferences of users. In this section, we describe a way to better mine the user's interest preferences by designing a new topic model (called Combining Latent Dirichlet Allocation (CLDA)). When learning topics from short texts, you can use long text as auxiliary features to solve the data sparseness problem of short texts. When learning topics from long texts, you can use short text to filter long texts to improve the accuracy of mining user interest preferences. The model can well combine the advantages of short text and long text and can optimally choose hyperparameters α and β. We will handle all the long text and short text in the corpus and output of a T-dimensional themed vector. Our long text aggregation strategy is to aggregate each user's short text content into each user's long text. Faced with the characteristics of the existing sparse short text theme modeling, the strategy we adopted is a combination of information retrieval technology and Wikipedia data. The main approach is to build a search engine based on Wikipedia data, cluster the short text into a keyword query result, and return a short feature extension from the query. Thus, the original short text set and the auxiliary long text set are constructed. Our inspiration for CLDA comes from the fact that long text indexes are more modeled than short texts [27].

4.2. Long Text Processing Method. Microblog information is short, representative words with fewer features. LDA modeling using short text has serious data sparsity problems. Therefore, many scholars use the original short text q as a query to search a large set of potentially relevant long documents Cq from an external large set C. This file will be used by the model as a "raw" for text extensions. The goal of this step is to get the relevant documents and a high recall rate. Existing technologies, such as reverse indexes used in information retrieval, sensitive areas of high-dimensional data points, and APIs directly from existing search engines can be utilized to implement the process [26]. If you want to make sure the recall rate is high, you have to set the number of returned documents to be quite large, for example, tens or hundreds of documents, but the resulting long text has a significant noise disturbance. We extract a long document L_i randomly from the long document Cq returned by the query and aggregate all the original short texts of the user into a long text Ls to form a group of long document vectors $\overrightarrow{L} = \{L_1, L_2, \ldots, L_d, L_S\}$ to assist short texts.

4.3. Short Text Processing Method. A user's interest preferences in a specific field may be generated by the user's historical behavior record and the keyword interest distribution of microblog. However, all the microblog posts published by the user are sent at different time periods, so the temporal attributes of the short text can well reflect the degree of the user's preference of interests in a particular field at a certain time point. It can be expressed by a set of weight vectors as shown in

$$P_S = \sum_{i=1}^{d} y_i * D_i * C = \{\mathsf{S}_{T1}, \mathsf{S}_{T2}, \ldots, \mathsf{S}_{Tm}\}, \quad (1)$$

where d is the total number of short texts published by the user; D_i is the subject keyword distribution of the user microblog; S_{Ti} denotes the user's preference degree of interest to the ith topic under the specific field; m is the subject number, and C is a constant; y_i is the weight of interest preferences over time in a particular area of the user. y_i follows the Ebbinghaus forgetting curve [28]. Ebbinghaus forgetting curve is a curve used to describe the change of human's memory over time. Previously, some scholars have used it to model the preference model, which shows that the user's preference changing process follows the same rule as the curve process [29, 30]. With the increase of time, the user's preference of interest declines sharply at first and decays to a certain extent and then shows a steady decline. y_i is shown by the formula below:

$$y_i = \lambda + (100 - \lambda)\, e^{-|t-t_1|} \quad (0 \leqslant \lambda \leqslant 99), \quad (2)$$

where t is the current time and t_1 is the time at which the user posted the document. Due to the different speed of change in the user's interest preferences, we have added a dynamic active parameter λ to establish a different Ebbinghaus forgetting curve for each user. Each user activity sliding window mechanism is as follows:

(1) Set the initial value $\lambda = 7$; that is, the minimum observation period of the active sliding window is 7 days.

(2) Calculate the total number of microblogs originally created and forwarded by the user within $[t - \lambda, t]$, denoted as N.

(3) If the value of N is less than 30 in the time sliding window, the time window is expanded, $\lambda = \lambda + 2$.

(4) Repeat steps (2) and (3) until $N >= 30$ or $\lambda = 99$.

4.4. Combining Latent Dirichlet Allocation (CLDA). In the CLDA model, we mainly use two key ideas on how to combine short and long texts.

(1) We can create two different approaches for short texts and long texts (as shown in Sections 4.2 and 4.3) and establish a new thematic model that can be used for auxiliary long text data and target short text data. This approach captures the main topics in each of the two data sets separately, derives the interest preferences in each short text of one user and the interest preferences in the auxiliary long text, and filters out irrelevant or inconsistent topic interest preferences in the auxiliary data.

(2) We can also use different build procedures for auxiliary long text and target short text, respectively, so that the model facilitates more accurate mining of user interest preferences in specific fields and the use of topic generated documents belonging to their field.

In order to better combine the long text with the short text, we use CUI to express the preference of a user. It includes the relationship between the user's preference of all short text and long text. CUI is defined as

$$\mathrm{CUI} = W_1 \sum_{i=1}^{m} S_{Ti} + W_2 \sigma \left(W_1 + W_2 = 1\right), \quad (3)$$

where S_{Ti} is the weight vector of each short text of a user; σ is the value of mining interest of aggregated long texts, and when the aggregated long texts have no preference of users in a particular field, the value is 0; on the contrary, the value will be a fixed value C_L; W_1 and W_2 represent the weight of short text and long text, respectively.

We propose a new model for auxiliary long text data and target short text data, called Combining Latent Dirichlet Allocation (CLDA), which considers the relationship between short text and long text of microblog users based on LDA.

CLDA generation process is shown in Algorithm 2; Bayesian network diagram is shown in Figure 2, where t represents the temporal effect of short text, and its value is determined by the P_S; R stands for the relationship between long text and short text, obeying the binomial distribution with parameter x, which a priori obeys the Dirichlet distribution $p(x) \sim \mathrm{Dirichlet}(N)$. The main role lies in the use of long text to help short text learning tasks and the value of the CUI decision; N_S, M_S, K_S indicate the number of samples in the short text. N_L, M_L, K_L indicate the number of samples in the long text. The left side of Figure 2 is the topic generation process for short texts. The right side is the generation process

(1) For each topic $z \in \{1, \ldots, K\}$,
  Choose long text multinomial distribution over terms, $\varphi \sim \mathrm{Dir}(\beta_L)$.
  Choose short text multinomial distribution over terms, $\varphi \sim \mathrm{Dir}(\beta_S)$.
(2) For each document $d \in \{1, \ldots, M\}$
  (a) If there is a long text of a microblog text associated with a interest preference under a specific area of the current user, the subject of the multi-distribution of long text and short text is selected together, $\theta \sim \mathrm{Dir}(\alpha_L)$, $\theta \sim \mathrm{Dir}(\alpha_s)$. Conversely, only the topic of the multinomial distribution of the short text relational documents $\theta_1, \theta_2, \ldots, \theta_n$, $\theta \sim \mathrm{Dir}(\alpha_s)$
(3) For each of the N words w_n
  Choose a topic $z_n \sim p(z \mid \theta)$;
  Choose a word $w_n \sim p(w \mid z, \beta_L, \beta_S)$;

Algorithm 2: The generation process of CLDA.

for long texts, and the middle is combining long texts and short texts. First, the short text and the long text in the left and right sides of Figure 2, respectively, select a topic word distribution (φ) for each topic from the hyperparameters β_s, β_L of the Dirichlet distribution. This process corresponds to step (1) of Algorithm 2. Second, when generating documents, the model selects the topic distribution θ from only the Dirichlet distribution hyperparameters α_s of the short text, if the long text does not have the microblog text associated with the interest preference in the current user-specific domain. If there is a long text of a microblog text associated with a hobby preference under a specific area of the current user, the topic distribution θ is selected from the Dirichlet distribution hyperparameters α_s, α_L of each short text and long text. This process corresponds to step (2) of Algorithm 2. Finally, according to the probability distribution of topic θ, select the topic for each document, and then select one word from the topic word distribution. Repeat until the long and short texts have their own documents and place them in R. Calculate the joint probability of the final long text and short text through CUI. This process corresponds to step (3) of Algorithm 2.

In the CLDA model, the topic distribution of microblog texts is shown in

$$\begin{aligned} &p(\theta \mid \alpha_s, \alpha_L, e, x) \\ &\quad = \{(1-e) + e(1-R)\, p(R \mid x)\}\, p(\theta \mid \alpha_L)\, e \times R \\ &\qquad \times p(R \mid x)\, p(\theta \mid \alpha_s), \end{aligned} \tag{4}$$

where α_s follows the following formula:

$$\alpha_s = \frac{1}{k}\left[\frac{\sum_{i=0}^{n} R_i(\theta_i)}{\sum_{i=0}^{n} R_i}\right], \tag{5}$$

where $\theta_1, \theta_2, \ldots, \theta_n$ are the topic distributions of the respective microblog short texts of the related user documents. Therefore, the joint probability distributions of the N topics z and N words w for given parameters α_S, α_L, β_L, β_s, e, and x are as shown in

$$\begin{aligned} &P(w, z \mid \alpha_S, \alpha_L, \beta_s, \beta_L, e, x) = p(w \mid z, \beta_s, \beta_L)\, p(z \mid \theta) \\ &\quad \cdot p(\theta \mid \alpha_S, \alpha_L, e, x) = p(w \mid z, \beta_s, \beta_L)\, p(z \mid \theta) \\ &\quad \cdot \{\{(1-e) + e \times p(R \mid x)(1-R)\}\, P(\theta \mid \alpha_L) + e \\ &\quad \times p(R \mid x) \times R \times P(\theta \mid \alpha_s)\}. \end{aligned} \tag{6}$$

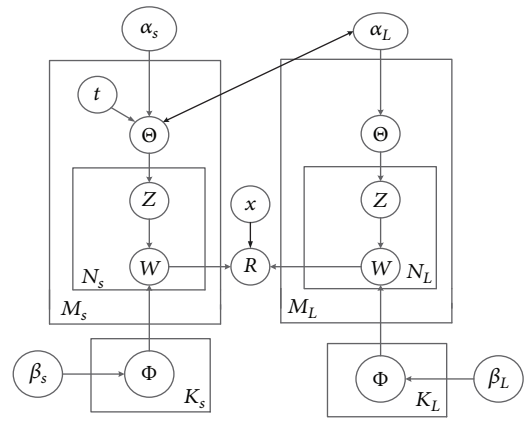

Figure 2: CLDA Bayesian network diagram.

4.5. Model Inference. We used the Gibbs sampling method to derive the CLDA model. The Gibbs sampling method, one of the most widely used methods of the Markov Chain Monte Carlo (MCMC) method, is used to obtain a series of joint probability distributions approximately equal to a given multidimensional probability distribution (such as 2 or more random variables) observing the algorithm of the sample.

In the conditional distribution, a word is randomly sampled as a new topic distribution. Based on the distribution of all potential variables, we can infer a potential distribution of topics as shown in

$$\begin{aligned} &P(z_j = k \mid z_{\neg j, w}) \\ &\quad = \frac{n^x_{k,\neg j} + \beta x}{\sum_{x=1}^{v} n^x_{k,\neg j} + \beta x} \frac{n^k_{j,\neg j} + \alpha_k^{1-f} \cdot \alpha_k^f}{\sum_{k=1}^{K}\left(n^k_{j,\neg j} + \alpha_k^{1-f} \cdot \alpha_k^f\right)}, \end{aligned} \tag{7}$$

where z_j indicates that the jth word in the document is assigned to the topic k; $z_{\neg j}$ represents all distribution subject

assignments except the jth word; v is the total number of words in the dictionary; $n^x_{k,\neg j}$ is the number of times that items other than the jth word are assigned to the topic k and the dictionary V; $n^k_{j,\neg j}$ represents the number of occurrences in topic k in document d except for the jth term topic; f indicates whether this Weibo text is empty.

The derivation of $p(\phi \mid z, w, \beta_s, \beta_L)$ and $p(\theta \mid z, \alpha_S, \alpha_L)$ is as follows:

$$p(\phi \mid z, w, \beta_s, \beta_L) = \frac{\prod_{x=1}^{v} \phi_{kx}^{n_k^x + \beta_x - 1}}{\int \prod_{x=1}^{v} \phi_{kx}^{n_k^x + \beta_x - 1} d\phi},$$
$$p(\theta \mid z, \alpha_S, \alpha_L) = \frac{\prod_{k=1}^{K} \theta_j^{n_j^k + \alpha_L^{1-f} \cdot \alpha_S^f - 1}}{\int \prod_{k=1}^{K} \theta_j^{n_j^k + \alpha_L^{1-f} \cdot \alpha_S^f - 1} d\theta}. \quad (8)$$

Due to $\int \prod_{x=1}^{v} \phi_{kx}^{n_k^x + \beta_x - 1} d\phi = \Delta(n_k + \beta_L \cdot \beta_S)$ and $\int \prod_{k=1}^{K} \theta_j^{n_j^k + \alpha_L^{1-f} \cdot \alpha_S^f - 1} d\theta = \Delta(n_j + \alpha_L^{1-f} \cdot \alpha_S^f)$ we can deduce two equations:

$$p(\phi \mid z, w, \beta_s, \beta_L) = \mathrm{Dir}(\phi \mid n_k + \beta_L \cdot \beta_S) \quad (9)$$

$$p(\theta \mid z, \alpha_S, \alpha_L) = \mathrm{Dir}\left(\theta \mid n_j + \alpha_L^{1-f} \cdot \alpha_S^f\right), \quad (10)$$

where n^x_k represents the number of times the term v in the dictionary is assigned to the topic k; n^k_j represents the number of occurrences of the jth term in topic k in document d; f indicates whether this Weibo text is empty. βx represents the xth text element of vector β.

5. Experiments

5.1. Data Set. The data set used in this article comes from Sina Microblog. We crawled in six areas and a total of 600 users posted 234,687 microblog texts from January 2017 to November 2017.

As a first step, we filtered Sina Weibo through language codes. We mainly obtained Chinese Weibo. Subsequently, we performed some basic cleanups, such as replacing usernames and clearing labels, URLs, numbers, and common symbols. Finally, we use the Ansi Chinese word segmentation tool for text segmentation, removing all punctuation marks from strings and English markers. We have also formed a stop-word list to eliminate very common and rare words.

We process the completed data in user units into 600 user-long texts. After that, we did the following for the short text:

(i) When the number of short text words is less than 2, the short text is deleted.

(ii) When the number of short text words is between 2 and 50, we use a novel end-to-end extended feature information retrieval technology to lengthen short texts into long texts and retain the original short texts.

(iii) When the number of short text words is greater than 50, keep the original short text without any operation.

Finally, the number of valid raw short texts and effective auxiliary long texts we have obtained and the field selected for them are shown in Table 1.

Table 1: Experimental data distribution.

Field	Number of users	The number of valid short texts	The number of valid long texts
Education field	150	38155	15488
Entertainment field	150	21955	7556
Medical field	150	22356	10655
Traffic field	150	27898	12568
Total	600	110364	47673

5.2. Baseline Methods and Evaluation Criteria. In this paper, the Micro-$F1$ value is used as the evaluation index of the interest preference of each model in mining specific areas. We use CLDA for direct referrals and accuracy values as a measure of recommendation accuracy.

We compare CLDA with the following methods:

LDA: it is the standard LDA method, used to gather short and long texts to learn LDA directly.

LDA-L: it includes aggregate long texts according to the method presented in Section 4.3 and LDA from long texts.

LDA-S: the short text is processed according to the method presented in Section 4.4 and the LDA is learned from the short text.

MB-LDA: it is a generative model based on LDA for theme mining on Weibo [31, 32].

VSM: vector space model is a model that simplifies the processing of textual content to vector operations in vector space, and it expresses semantic similarity using spatial similarity. We use the VSM method to process all the user U_i text data to get the word weight vector $U_i = \{w_{i1}, w_{i2}, \ldots, w_{ij}\}$. Where w_{ij} is the weight of word j in user U_i text data, we use TF-IDF to calculate the weight value. In the recommendation, the method of calculating the similarity between users adopts the conventional angle cosine value of the following formula:

$$\mathrm{Sim}(U_i, U_j) = \frac{U_i \cdot U_j}{\left|U_i^2\right| \left|U_j^2\right|}. \quad (11)$$

The micro-precision rate (Micro-P) is defined as in (12); micro-recall rate (Micro-R) is defined as in (13); micro-F value (Micro-$F1$) is defined as in (14). Among them, TP is the correct classification of the text to the user with a certain number of interest preferences; FN is the number of model errors that classify the text into user interest preferences; FP is the number of incorrectly categorized texts that interest the user's preferences in the model into other user interest preferences. n is the total number of all user interest preferences in a particular area. The higher

the value of Micro-$F1$, the better the classification performance.

$$\text{Micro-}P = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i} \tag{12}$$

$$\text{Micro-}R = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i} \tag{13}$$

$$\text{Micro-}F = \frac{2\,\text{Micro-}P \times \text{Micro-}R}{\text{Micro-}P + \text{Micro-}R}. \tag{14}$$

In the recommendation system, recommendation is often based on the similarity between users. We take t users before extraction as recommended list to the user U_i to form the recommended set $U_t = \{U_1, U_2, \ldots, U_j, \ldots, U_t\}$. For each user U_j in the recommended set, it is determined whether the user U_i is in the same specific area, and if so it is considered correct to recommend U_j to U_i. The recommended accuracy of a single user U_i is shown in formula (15). In certain area A user recommended accuracy is as shown in (16), where N is the total number of users under the domain where user U_i is located.

$$\text{Accuracy}\left(U_i\right) = \frac{\sum_{j=1}^{t} f\left(U_i, U_j\right)}{k}, \quad t \leqslant N-1, \quad f\left(U_i, U_j\right) = \begin{cases} 1 & U_i, U_j \text{ belong to the same area} \\ 0 & U_i, U_j \text{ do not belong to the same area} \end{cases} \tag{15}$$

$$\text{Accuracy}\left(A\right) = \frac{\sum_{i=1}^{N} \text{Accuracy}\left(U_i\right)}{N}. \tag{16}$$

5.3. Result. Figure 3 shows the classification performance of each topic number K on Sina microblog data set. We set the parameter α to 50/K and the parameter β to 0.01. The abscissa K is a variable, and we adjust the effect of each model by changing the size of K. The ordinate is the Micro-$F1$ rating, which shows the performance of each model in obtaining a user's preference for interest.

As shown in Figure 3, LDA-S can achieve better results when the number of topics K is smaller, and when $K = 4$, the maximum value of Micro-$F1$ reaches 61.8%. As the number of topics K increases, LDA, MR-LDA, LDA-L, and CLDA will reach relatively high values and then decrease gradually. This result shows that when the number of topics is too large or too small, the impact of each model will be affected. The Micro-F values of MB-LDA, LDA, and LDA-L reached the maximum of 73.8%, 66.7%, and 70.8%, respectively, when K reached 8 and 10, respectively. Because CLDA combines the advantages of both LDA-S and LDA-L, CLDA achieves relatively good performance at both small and large K values. When $K = 10$, the Micro-$F1$ value of CLDA reaches a maximum of 76.1%, which is higher than other models.

Figure 4 shows the relationship between the Micro-$F1$ value of the CLDA model and the weight $W1$ of the short text in the CUI when k is set to ten. From Figure 4, W_1 value of 0.5 with the best performance can be seen. It shows that the proportion of long text and short text in CLDA model is the same, occupying the same position.

Since CLDA models can recommend users with similar interests, we use the user recommended accuracy values described in Section 5.2 to measure the quality of the model. The results of the comparison of accuracy values between CLDA and VSM in the education, entertainment, medical, and traffic fields are illustrated in Figures 5–8. The average results of the accuracy values in the various fields of CLDA and VSM are shown in Figure 9.

In Figures 5–9, we show the results of $t = 10$, 20, and 30 users before extraction, respectively. The number of topics in CLDA takes the optimal result of $K = 10$. On the whole, the recommendation effect of CLDA is better than that of VSM. Especially in Figures 5 and 6, the results of recommendation in the field of education and entertainment are outstanding, and the accuracy of 75% and 72.1% is obtained at $t = 10$, respectively.

However, in Figures 5–9, we can see that the best recommendation effect of each model of Sina data set we collected is $t = 10$, and as the value of t increases the recommendation effect begins to decline. CLDA is more affected by the t-value, while VSM have less effect. Therefore, CLDA is more susceptible to t-value when it is recommended. This is where we will improve in the future.

In Figures 7 and 8, we also found deviations from the recommended results in the medical and traffic fields. Analyzing the microblog of users in these fields, we found that this may be due to the relative unpopularity of these fields compared with other fields and the relatively few users to discuss, so that the microblogging published by the users is relatively broad, the content and the topics involved are complicated, and the theme mining the interference is relatively large. How to reduce this kind of data to user's recommendation interference is the focus of work in the future.

6. Conclusion and Future Work

In this paper, aiming at the short text of Weibo data, combined with LDA model, we propose a novel theme model. The model can learn the potential topics of short texts and long texts simultaneously, by aggregating long texts to assist short text learning tasks, to avoid short text data sparsity. Finally, short text filtering long text is used to improve mining accuracy, making the long text and short text have effective joint use. The experimental results show that our model can outperform many advanced models, not only effectively mining the topics of interest to users, but also having the ability to be applied to the recommendation system. In future

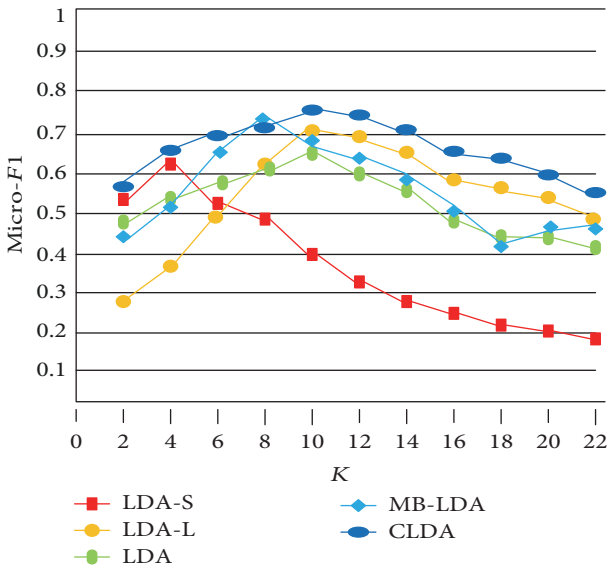

Figure 3: Comparison of Micro-$F1$ for each model.

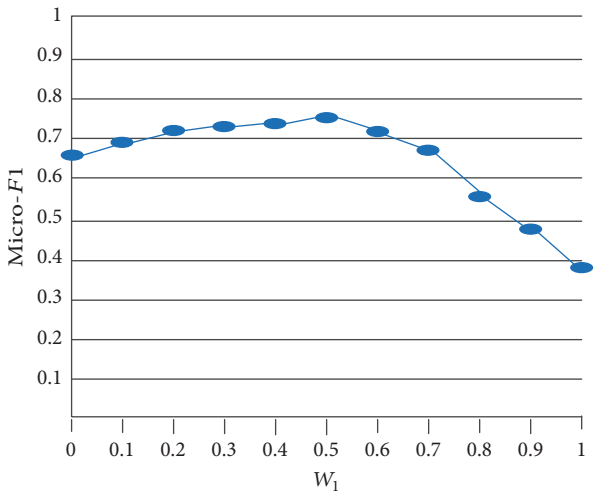

Figure 4: The importance of long text and short text in CLDA.

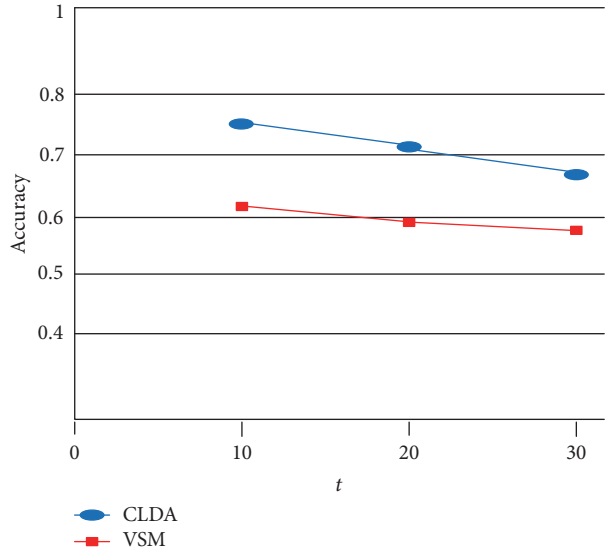

Figure 5: Recommended accuracy in education field.

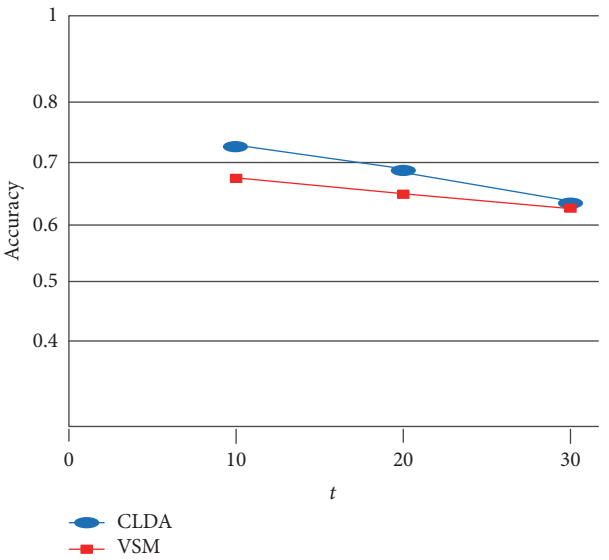

Figure 6: Recommended accuracy in entertainment field.

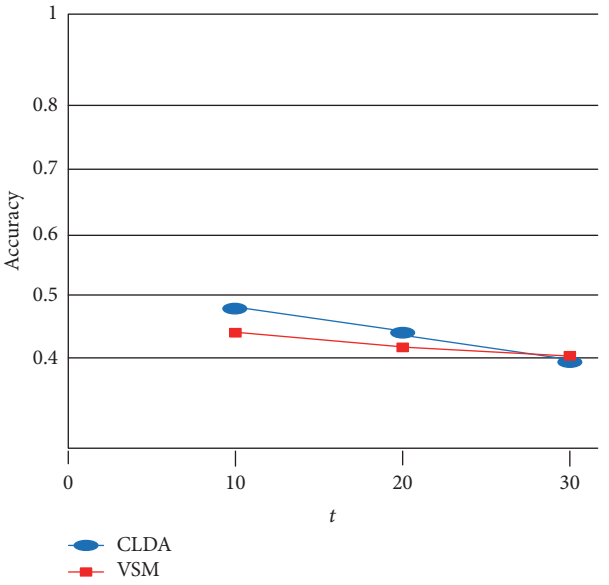

Figure 7: Recommended accuracy in medical field.

research work, we will continue to optimize the effectiveness and efficiency of the CLDA model and reduce the interference of the nonmeaningful Weibo on the topic mining so as to adapt to various fields. Try combining more social network features and real-time microblogging data processing.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. Also all the mentioned funding did not lead to any conflicts of interest.

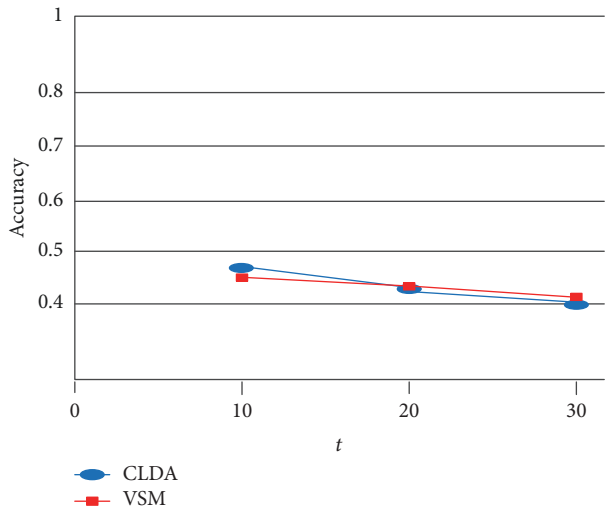

Figure 8: Recommended accuracy in traffic field.

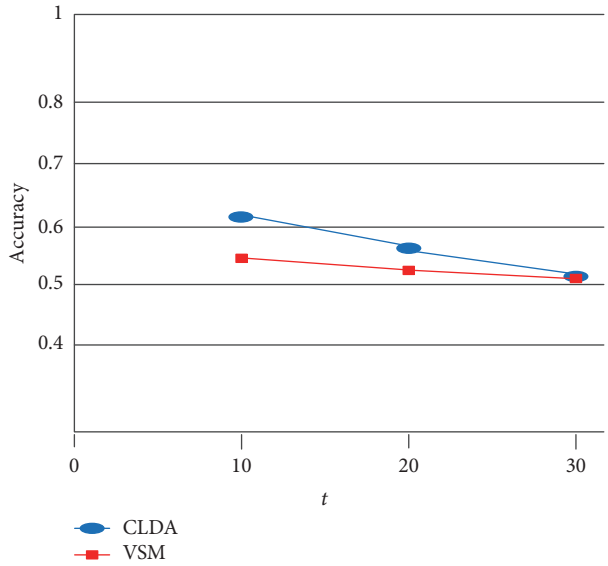

Figure 9: Recommended accuracy averages in various fields.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 61672553) and the Project of Humanities and Social Sciences of the Ministry of Education in China (Project no. 16YJCZH076). This work is supported by the Open Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education (Grant no. 2018KF01).

References

[1] W. Ding and Zhaoyun, "Mining user interest in microblogs with a user-topic model," *China Communications*, vol. 11, no. 8, pp. 131–144, 2014.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pp. 50–57, Berkeley, Calif, USA, 1999.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] J. Weng, E. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential twitterers," in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pp. 261–270, New York, NY, USA, February 2010.

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," *Workshop on Social Media Analytics (SIGKDD)*, pp. 80–88, 2010.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the Association for Information Science and Technology*, vol. 41, no. 6, pp. 391–407, 1990.

[7] S. Deerwester, "Indexing by latent semantic analysis," *Journal of the Association for Information Science & Technology*, vol. 41, no. 6, pp. 391–407, 2010.

[8] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[9] H. Xiong, "A location-sentiment-aware recommender system for both home-town and out-of-town users," in *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1143, 2017.

[10] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, "GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014*, pp. 831–840, USA, August 2014.

[11] L.-C. Chen, "An effective LDA-based time topic model to improve blog search performance," *Information Processing & Management*, vol. 53, no. 6, pp. 1299–1319, 2017.

[12] M. Efron, P. Organisciak, and K. Fenlon, "Improving retrieval of short texts through document expansion," in *Proceedings of the 35th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012*, pp. 911–920, usa, August 2012.

[13] Y. He, Y. Li, J. Lei, and C. H. C. Leung, "A framework of query expansion for image retrieval based on knowledge base and concept similarity," *Neurocomputing*, vol. 204, pp. 26–32, 2016.

[14] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in *Proceedings of the International Conference on World Wide Web (WWW '06)*, pp. 377–386, Edinburgh, Scotland, 2006.

[15] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, pp. 806–814, 2016.

[16] X. Hu, N. Sun, C. Zhang, and T. S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proceedings of the Acm Conference on Information & Knowledge Management*, pp. 919–928, 2009.

[17] S. Amir, A. Tanasescu, and D. A. Zighed, "Sentence similarity based on semantic kernels for intelligent text retrieval," *Journal of Intelligent Information Systems*, pp. 1–15, 2016.

[18] N. Schlaefer, J. Chu-Carroll, E. Nyberg, J. Fan, W. Zadrozny, and D. Ferrucci, "Statistical source expansion for question answering," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM'11*, pp. 345–354, gbr, October 2011.

[19] J. Dalton, L. Dietz, and J. Allan, "Entity query feature expansion using knowledge base links," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014*, pp. 365–374, aus, July 2014.

[20] J. Tang, Y. Wang, K. Zheng, and Q. Mei, "End-to-end learning for short text expansion," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017*, pp. 1105–1113, can, August 2017.

[21] M. RosenZvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04)*, pp. 487–494, 2012.

[22] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume Association for Computational Linguistics*, pp. 248–256, 2009.

[23] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM '10)*, pp. 130–137, Washington, Dc, Wash, USA, 2010.

[24] W. X. Zhao, J. Jiang, J. Weng et al., "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*, pp. 338–349, Springer-Verlag, Heidelberg, Germany, 2011.

[25] X. H. Phan, C. T. Nguyen, D. T. Le, L. M. Nguyen, S. Horiguchi, and Q. T. Ha, "A hidden topic-based framework toward building applications with short web documents," *Transactions on Knowledge Data Engineering*, vol. 23, no. 7, pp. 961–976, 2011.

[26] T. L. Luong, Q. T. Truong, H. T. Dang, and X. H. Phan, "Domain identification for intention posts on online social media," *Symposium on Information and Communication Technology*, pp. 52–57, 2016.

[27] X. Wu, L. Fang, P. Wang, and N. Yu, "Performance of using LDA for Chinese news text classification," *Electrical and Computer Engineering*, vol. 2015, pp. 1260–1264, 2015.

[28] K. K. Kopiske, N. Bruno, C. Hesse, T. Schenk, and V. H. Franz, "The functional subdivision of the visual brain: Is there a real illusion effect on action? A multi-lab replication study," *Cortex*, vol. 79, pp. 130–152, 2016.

[29] W. Meng, L. Lin, W. Jing, P. Yu, J. Liu, and X. Fei, *Improving Short Text Classification Using Public Search Engines. Uncertainty in Knowledge Modelling and Decision Making*, Springer, Heidelberg, Germany, 2013.

[30] H. Palangi, L. Deng, Y. Shen et al., "Deep Sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 4, pp. 694–707, 2016.

[31] C. Zhang, J. Sun, and Y. Ding, "Topic mining for microblog based on MB-LDA model," *Computer Research and Development*, vol. 48, no. 10, pp. 1795–1802, 2011.

[32] Z. Wang, Y. H. Shao, L. Bai, C. N. Li, L. M. Liu, and N. Y. Deng, "Mb-lda: a novel multiple between-class linear discriminant analysis," *Information Sciences*, vol. 369, pp. 199–220, 2016.