Hindawi
Complexity
Volume 2018, Article ID 5809043, 11 pages
https://doi.org/10.1155/2018/5809043

Research Article

Tibetan Weibo User Group Division Based on Semantic Information in the Era of Big Data

Lirong Qiu,[1] **Jia Yu**,[1] **Jie Li**,[1] **and HaoRan Jia**[2]

[1]*School of Information Engineering, Minzu University of China, Beijing, China*
[2]*International School, Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Beijing 100876, China*

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

Received 17 May 2018; Revised 29 June 2018; Accepted 9 July 2018; Published 5 August 2018

Academic Editor: Jaime Lloret

In the era of big data, group division in online social network analysis is a basic task. It can be divided into the group division based on static relationship and the group division based on dynamic relationship. Compared with the static group division, users express their semantic information in all kinds of social network behaviors, and they tend to interact with other users who have the same idea and attitude; this is how different groups are formed. In this paper, aimed at the issue that some Tibetan users use Chinese to publish microblogs on social platforms, a group division method based on semantic information of Tibetan users under the big data environment is proposed. When dividing a large number of Tibetan user groups in a social network, a large amount of semantic information of Tibetan users in the social network is first analyzed. Then, based on the semantic similarity between users, we aggregate the Tibetan users with high similarities into one group, thus achieving the final group division. The experimental results illustrate the effectiveness of the method of analyzing Tibetan user semantic information in the context of big data for group partitioning.

1. Introduction

With the rapid development of the Internet and the arrival of the era of big data, digital social network websites have attracted more and more people to participate in online social networking, and people's communication is no longer blocked by time and space barriers. This has given people an efficient and convenient way to communicate, which has completely changed people's lives. People's behavior on social networks is stored in the website database with the form of data, which provides a great convenience to scholars who are engaged in sociological behavior research. However, the enormous user scale, the large amount of data, and the complex network structure have posed great challenges to the research of social networks. Therefore, under such a big data environment, it is very valuable to study an effective method of group division social network users.

The division of user groups is a prerequisite for online social network analysis. Only when a specific user group is identified can the group be further analyzed. This paper will use big data technology to tap the Tibetan microblogging user community in the Weibo online social network. This work is of great significance. For the government, it can quickly and accurately locate the group via the Weibo platform and understand the group's concerns and behaviors; thus, it reduces the time of offline manual research and document communication between departments, expands the channels and improves the efficiency of communication, and controls real-time feedback of the policy implementation. For the researchers, it can help to collect valid social samples from Weibo and analyze the internal characteristics of Tibetan Weibo users effectively. For enterprises, it has a great potential commercial value, because it can accurately find the Tibetan users' interests and delivery advertisements accordingly.

Some Tibetan users use Tibetan in online social networking platforms, while some Tibetan users use Chinese to express their opinions. For Tibetan users, it can be easily determined as a Tibetan user. In order to effectively dig out Tibetan users from users who use Chinese language, considering that Tibetan users' opinions and interests expressed on social platforms are different from non-Tibetan users, this

paper conducts research from the perspective of semantic information expressed by users. The main contributions of this paper are as follows:

(1) The semantic information expressed by online social network users is analyzed. Since the semantic expressions of the users are all realized through various behaviors in the social network, the topology between users in the online social network is constructed based on the user interaction behaviors that represent the semantic information; the network structure describes the intimacy between users from a semantic level.

(2) When dividing the network group, each of user nodes is regarded as an independent group, and then the groups were aggregated according to their similarity until a large group containing two subgroups with similar internal dimensions is obtained. The two subgroups are the result of group division, and the structure of the subgroups also represents the intimacy degree among users.

2. Related Works

With the development of network communication technology and the arrival of the era of big data, the way of interaction between people is digitized. The research on the social network has also changed from traditional sociological research to data mining research, from social behavior and social relation research to network mathematical statistics and quantitative analysis research [1]. Golbeck and Rothstein [2] used the FOAF (friend of a friend project) model to define the semantic social network. As a result, community discovery research transitions from traditional nonsemantic community discovery to semantic community discovery.

When the semantic community in social networks is divided, the topological structure and semantic information of the network are taken as the research object, and the traditional community discovery theory is used as the basis. If the semantic information expressed by the user is added to the algorithm model, the results of the community division will be more reasonable [3].

Steyvers et al. [4] proposed the AT model. This model firstly introduces the LDA model in the field of social network analysis. Through the topic modeling of the topic distribution of user nodes in social networks, the topics at the user node level are extracted. Zhu et al. used the nonsemantic community discovery form of NMF [5, 6] to divide the semantic community and put forward the CCLC (combining content and link for classification) algorithm [7]. Yang et al. proposed the PCL (popularity-based conditional link) model [8], which divides semantic community modeling into semantic modeling and topological relationship modeling [9]. Rios and Munoz and Mattingly et al. proposed the SLTA (speaker-listener topic propagation algorithm) [10, 11]. Wang and Fang used the SLTA to establish a user activity network through semantic data [12]. Kianian et al. [13] proposed a semantic community discovery algorithm based on a label propagation algorithm. Based on the higher degree of intimacy between user nodes in social networks, the topic distribution is more similar [14, 15]. Hu et al. proposed the FT (feature topic) model for user semantic information analysis and the close relationship between users. Analysis of the ST (social topic) model, in which the FT and ST models are independent of LDA models [16]. Natarajan et al. [17] used link community as the starting point to establish a link-content model that uses link-content as the semantic analysis object [18, 19].

The above algorithms give the realization methods of semantic community division in complex networks. However, it ignores the use of the probability topic model to divide the network without considering the overall network topology; it cannot be reasonably explained by the corresponding real groups after getting the network division structure.

Therefore, this paper differs from other semantic partitioning algorithms in that we fully consider the various behaviors of users in online social networks who serve to express their views, attitudes, and emotions; this paper proposes a Tibetan microblog user group division algorithm based on user semantic information in the age of big data.

3. Semantic Analysis

The semantic information of users in online social networks is often expressed through the text content published by users. Semantic analysis takes the text content published by users as the research object and analyzes individual users or texts by mining the semantic information expressed in texts. The semantic similarity (i.e., the similarity of topics) will ultimately connect the users through the calculation of semantic similarity to form a user network structure, achieve group division, and make the division result more rational.

The latent Dirichlet allocation (LDA) topic model is a probabilistic model for textual data modeling that enables the modeling of subject information of textual data. And the LDA topic model can effectively realize the reduced dimension representation of text in the semantic space, and it models the text with the probability of vocabulary, which can alleviate the problems brought by data sparsity to some extent. This article uses the LDA topic model to model the content of the Weibo text.

3.1. LDA Topic Model. The LDA topic model is a hierarchical Bayesian model that first assumes that the words appearing in the text are independent and irrelevant and considers that each document is composed of a number of implicit topics, which are made up of some specific words in the text.

LDA is a typical probability model, determined by the parameter (α, β), which indicates the relative strength between hidden topics in the text set, and β indicates the probability distribution of the hidden topic itself. The generation process of the LDA model is shown in Figure 1, where θ_m represents the subject probability distribution of the text (i.e., to say, there are m kinds of "document-subject" distributions and each document has an θ_m distribution), ϕ_K represents the feature word probability distribution of each topic (i.e., to say, there are K kinds of "subject-word"

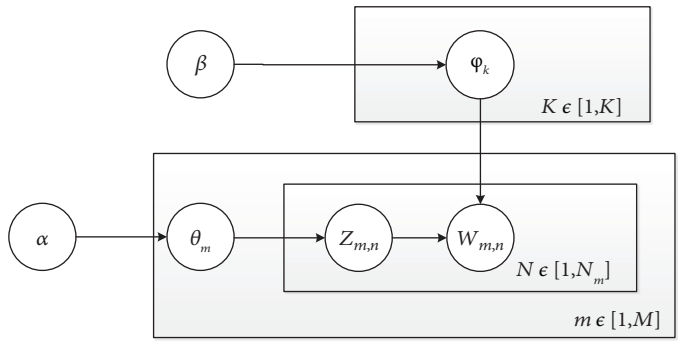

Figure 1: LDA generation process.

distributions), M is the number of documents, K is the number of subjects in all documents, and N is the number of feature words in each document.

The LDA document generation process is shown in Algorithm 1.

The LDA model performs the following process for each document:

(1) Take a sample from the Dirichlet distribution α to generate the topic distribution θ_m of the document.

(2) Take a sample from the polynomial distribution θ_m of the topic to generate the topic $Z_{m,n}$ of the nth word in document m.

(3) Take a sample from the Dirichlet distribution β to generate the feature word distribution ϕ_K of the topic $Z_{m,n}$.

(4) Take a sample from the polynomial distribution ϕ_K of total words to generate the final word $W_{m,n}$.

The joint distribution of variables in the LDA model is shown in

$$P(w_m, Z_m, \theta_m, \phi_K|\alpha, \beta) = \prod_{n=1}^{N} p(\theta_m|\alpha) \times p(Z_{m,n}|\theta_m) \times p(\phi_K|\beta) \times p\left(W_{m,n}\middle|\theta_{Z_{m,n}}\right). \quad (1)$$

Finally, the maximum likelihood estimation of the feature word distribution in each document can be obtained by integrating θ_m and ϕ_K from (1) and summing $Z_{m,n}$, which is shown in

$$P(w_m|\alpha, \beta)\int_{\theta_m}\int_{\phi_K}\sum_{Z_m} p(w_m, Z_m, \theta_m, \phi_K|\alpha, \beta), \quad (2)$$

3.2. The Improved LDA Model Based on the Weibo Text. Most of the content of Weibo is short text, which has a serious problem of data sparseness; its text content is generally colloquial, with unstandardized grammar and a large number of cyber languages, symbols, and buzzwords. This makes Weibo text data have large noise, and by combining fast update speed and large text data size of Weibo, the topic modeling of the Weibo text becomes complicated and difficult.

```
1: for all topics k ∈ [1, K] do
2:    sample mixture components φ_k~Dir(β)
3: end for
4: for all documents m ∈ [1, M] do
5:    sample mixture proportion θ_m~Dir(α)
6:    sample document length N_m~Poiss(ε)
7:    for all words n ∈ [1, N_m] do
8:          sample topic index z_m,n~Mult(θ_m)
9:          sample item for word w_m,n~Mult(φ_Zm,n)
10:   end for
11: end for
```

Algorithm 1: LDA document generation process.

The original LDA topic model is an unsupervised model, and the result of directly using the LDA to model the Weibo text will be affected by the Weibo text length. Since the Weibo content is short text with relatively lower occurrences of a single word, it is hard to judge if the two words are related or not. Therefore, modeling the Weibo text directly with the LDA model cannot achieve a satisfying result.

After analyzing the document generation process of the Weibo text topic model, the main reason that it is difficult to obtain the ideal effect directly with the topic model modeling is that the topic distribution ϕ is not reasonable enough. To solve this problem, Hong and Davison [20] trained the LDA model by presorting the Weibo text and then synthesizing the growth text. Based on this, the paper first trains the LDA model through standard news texts and then uses the microblog to optimize the topic-word distribution in the LDA model. The LDA generation process is shown in Algorithm 2.

3.3. Improved LDA Topic Model Based on Weibo Users. When group discovery is based on the semantic information between users in an online social network, it is necessary to know the semantic similarity between users; that is to say, the topic distribution needs to be obtained at the user level, which means the "user-topic" distribution. When topic modeling is carried out on the user level, we, respectively,

Input: *News text data: Document1, Weibo text data: Document2*
Output: *Topic distribution of Weibo short texts: Document_Topics_Matrix*
training LDA Model by Document1
for all *d, d* ∈ *Document*2 **do**
  combine short texts of microblogs belonging to the same topic to long texts Document3
end for
for all *d, d* ∈ *Document*3 **do**
  predict topics of d by LDA Model
  update $w_{m,n} \sim Mult(\varphi_{Zm,n})$ *of LDA_Model*
end for
for all *d, d* ∈ *Document*2 **do**
  predict topics of d by LDA_Model
  get Document_Topics_Matrix of d
end for
return *Document_Topics_Matrix*

Algorithm 2: Improved LDA topic model based on the Weibo text.

Input: *News user text data: User_Document1*
Output: *Weibo users' topic distribution: User_Topics_Matrix*
get LDA_Model by Algorithm 2
for all *d, d* ∈ *User_Document*_1 **do**
  predict topics of d by LDA_Model
  combine short texts of microblogs belonging to the same topic to long texts Document2.
end for
for all *d, d* ∈ *Document*2 **do**
  predict topics of d by LDA_Model
  update $w_{m,n} \sim Mult(\varphi_{Zm,n})$ *of LDA_Model*
end for
for all *user, user* ∈ *User_Document*1 **do**
  predict topics of d by LDA_Model
  get Document_Topics_Matrix of d
end for
return *User_Topics_Matrix*

Algorithm 3: Improved LDA topic model based on Weibo users.

model each Weibo text of every single user to obtain its topic distribution, which is later aggregated to obtain the user-level topic distribution. The experimental results show that the topic distribution obtained by this method is not ideal enough, the analysis found that each user has a large number of Weibo text contents, and most of the microblog texts are independent of each other in the expression of subject information. Therefore, the context of the long text aggregated does not have semantic relevance. In order to solve this problem, this article uses the following method for user-level topic modeling. The LDA generation process is shown in Algorithm 3.

(1) Use Algorithm 2 to model the topic of each user's microblog, and mine the topic distribution.

(2) The Weibo texts with similar theme distribution are aggregated to obtain a plurality of long Weibo texts with a single theme.

(3) Implement topic modeling to the aggregated long Weibo text, and use the generated topic distributions as the new topic distribution in the model.

(4) Aggregate all the Weibo texts of this user to obtain a long text, use the new topic model to carry out the topic modeling on the long text to obtain the topic distribution, and obtain the final topic distribution of this user.

3.4. Semantic Similarity Calculation. In order to construct the social network structure according to a user's semantic information, that is to say, to construct a topological structure which is capable of representing the strength of the intimacy between the users, we first need to calculate the similarity of the semantic information expressed by the users. This paper achieves by calculating the similarity of topic distribution of users and their Weibo texts.

Input: *Topic(user) distribution matrix of the text: Matrix1, Matrix2*
Output: *Semantic similarity: Semantic_Similarity*
1: **for all** p_i, q_j, p_i = *Matrix1*, q_j = *Matrix2*, **do**
2: *get* $D_{KL}(p, q)$ *by formula* (3)
3: **end for**
4: *get* $D(p, q)$ *from* $D_{KL}(p, q)$ *by formula* (5)
5: *Semantic_Similarity = 1- D(p, q)*
6: **return** *Semantic_Similarity*

Algorithm 4: Calculation of the semantic similarity.

Probabilistic topic distribution obtained through topic modeling is a mapping in the vector space, and their similarity can be obtained by calculating the relative entropy between the topic distributions; that is, to calculate the Kullback-Leibler divergence, the KL distance measures the difference between the two probability distributions in the same time space, as shown in

$$D_{\mathrm{KL}}(p, q) = \sum_{j=1} p_j \ln \frac{p_j}{q_j}. \tag{3}$$

For any j, when $p_j = q_j$, there is always $D_{\mathrm{KL}}(p, q) = 0$. However, because of its asymmetry, which means $D_{\mathrm{KL}}(p, q) \neq D_{\mathrm{KL}}(q, p)$, the correction method is generally adopted in actual use, as shown in

$$\begin{aligned} D_\lambda(p, q) = {} & \lambda D_{\mathrm{KL}}(p, \lambda q + (1-\lambda)q) \\ & + (1-\lambda) D_{\mathrm{KL}}(q, \lambda p + (1-\lambda)q). \end{aligned} \tag{4}$$

When $\lambda = 1/2$, (4) becomes JS distance, and the range of JS distance is $[0, 1]$, as shown in

$$D(p, q) = \frac{1}{2}\left[D_{\mathrm{KL}}\left(p, \frac{q+q}{2}\right) + D_{\mathrm{KL}}\left(q, \frac{q+q}{2}\right)\right]. \tag{5}$$

When the semantic similarity between users is higher, the relative entropy of the probability topic which represents the semantic information in the vector space is smaller; that is to say, the value of JS distance is smaller.

Since the calculated $D(p, q)$ and the semantic similarity are negatively correlated, it is adjusted and modified in order to simplify the subsequent calculation, and the final semantic similarity calculation formula is shown in

$$S(p, q) = [1 - D(p, q)] \times 100\%, \tag{6}$$

where $S(p, q)$ represents the similarity between probability topic distribution p and q.

The calculation of the semantic similarity is shown in Algorithm 4.

4. Network Construction Based on User Semantics

This paper assumes that the semantic information between users in online social networks is generated along with different behaviors of users. That is to say, semantic similarity exists only when there is an interaction between users.

4.1. Semantic Link Analysis Based on Following Behavior. Take the users as the nodes in the semantic network, and the semantic similarity between users is the weight of the edge between users. When the influence of different weights from different users' nodes in the network is not considered, take the user as the node and the following connection between users as the edge to construct a directed graph and then calculate the similarity between the users according to the topic distribution and the semantic similarity calculation method obtained from the Section 2 topic modeling on the user level. The similarity is the weight of the link between user nodes, and the specific calculation method is shown in

$$w - B^{(1)}(i, j) = S(i, j), \tag{7}$$

where i and j represent user nodes and $S(i, j)$ represents the semantic similarity at the user level between i and j.

4.2. Semantic Link Analysis Based on Comment Behavior. In an online social network, if there is a comment between two users, a directed edge is generated between the commented user and the commenting user based on the semantic information expressed by this commentary behavior. For example, one user publishes a Weibo and another user comments, and both the Weibo content and the comment would generally express the user's semantic information. That is to say, the comment between users produces a semantic link between users. Take the comment behavior between users A and B as an example; when user A posts a Weibo and user B comments on it, the content of the text posted by both A and B includes the user's semantic information.

The construction of semantic links based on comment needs to calculate the semantic similarity; that is to say, the semantic similarity between the commented text content and the commentary content needs to be calculated. It can be obtained by topic distribution and semantic similarity calculation formula from topic modeling on the Weibo text, and the specific calculation method is shown in

$$w - B_n^{(2)}(i, j) = S(i, j)_n, \tag{8}$$

where n represents the nth comments between users.

Because there may be more comments among users, the links between users need to be constructed according to their overall average influence. If the multiple arithmetic means of comment link weight are simply taken, the trend of the relationship between users in the long cycle can be hardly reflected. This article uses the exponential moving average

Input: *Weibo user data set: User_DataSet*
Output: *Network adjacency matrix of user semantics: User_Semantic_Matrix*
1: **for all** *users, user* ∈ *User_DataSet* **do**
2: *get User_Sematic_Similatrity* B^1 *by formula* (7)
3: *get User_Content_Sematic_Similatrity* B^2 *by formula* (9):
4: *get User_Repost_Sematic_Similatrity* B^3 *by formula* (10):
5: *get User_Like_Sematic_Similatrity* B^4 *by formula* (11):
6: **end for**
7: *get node weight set W by* Algorithm 3
8: **for all** $B, w, B \in [B^1, B^2, B^3, B^4], w \in W$ **do**
9: get *User_Semantic_Matrix from S by formula* (12)
10: **end for**
11: **return** *User_Semantic_Matrix*

Algorithm 5: Network construction based on user semantics.

as the link weight generated by the user comments, and the specific calculation method is shown in

$$w - B^{(2)}(i,j) = \mathrm{EMA}\left[\sum w - B_n^{(2)}(i,j)\right]. \tag{9}$$

4.3. Semantic Link Analysis Based on Forwarding and Point-Like Behavior. In the Weibo social network, the forwarding behavior between users is different from that of other social networks. The user can comment while forwarding, and the influence of the comment while forwarding on the semantic links is essentially the same as that of comment only. Therefore, this paper includes the influence on semantic links of comment while forwarding into comment-only behaviors. When considering semantic links based on the forwarding behavior, only analyze the semantic information based on the forwarding times.

If the semantic links from the comment content are not taken into consideration, the semantic links based on forwarding are only related to forwarding times; then, it is not hard to find that this is similar to the "likes" between users.

Since the forwarding and the like behavior may occur many times between users, if the number of forwarding or like behaviors is linearly calculated, the link weights of the entire network will be generated mostly by these two behaviors. Considering that forwarding or like behaviors among users do not happen every time, we treat it as a probability event, and the probabilities are that the forwarding or like behaviors are taken as the weights of the links.

Treat the user's forwarding or like behavior as an independent repeat event, and calculate the total Weibo number N from user A and the number of forwarding times R and likes L from user B to A. The probability of the forwarding and like behavior from users B to A can be obtained, and the specific calculation method is shown in

$$w - B^{(3)}(i,j) = P_R^{ij} = \frac{R^{ij}}{N_j}, \tag{10}$$

$$w - B^{(4)}(i,j) = P_L^{ij} = \frac{L^{ij}}{N_j} \tag{11}$$

where P_R^{ij} and P_L^{ij} represent the probability of forwarding and like behavior from user i to user j, R^{ij} represents the forwarding time from i to j, L^{ij} represents the like time from i to j, N_j is the total Weibo number of user j, and $w - B^{(3)}(i,j)$ and $w - B^{(4)}(i,j)$ mean the semantic link weight based on forwarding and likes.

4.4. Network Construction Based on User Semantics. The previous chapters describe the different semantic links generated by different communication behaviors and give their specific calculation methods. When constructing the semantic network based on the semantic information generated by the user communication, the semantic information expressed by users through different behaviors is essentially the same, which is the opinion or concept that one user wants to express to another. Therefore, this paper assumes that the semantic links generated by different behaviors between users are equivalent. That is to say, the semantic links generated by different user behaviors have the same influence on constructing the overall semantic networks.

As a result, without considering user behavior's influence on the user network structure, the semantic network structure between users can be obtained. The calculation method of the adjacency matrix of the user network is shown in

$$w - B(i,j) = \sum_{z=1}^{4} w - B^{(i)}(i,j). \tag{12}$$

The concrete process of network construction based on user semantics is shown in Algorithm 5.

4.5. Group Division Algorithm for Tibetan Weibo Users. This paper claims that users in online social networks can be divided into different groups according to their semantic information, which means their interested topics. Due to the influence of national culture, there is a big difference between the interested topics between Tibetan users and non-Tibetan users. Therefore, the semantic similarity between users can be used to find Tibetan user groups in social networks.

In the initial stage of the algorithm, each user node is regarded as an independent group structure, and then by

Input: *Network adjacency matrix of user semantics: User_Semantic_Matrix*
Output: *The result of group division.: Communities*
1: **for all** *communities, communities* ϵ *User_Sematic_Matrix* **do**
2: *get Community_Similatrity from by User_Semantic_Matrix formula* (13)
3: *Combine the two groups with the largest similarity into one larger group: New_Matrix*
4: *User_Semantic_Matrix = New_Matrix*
5: **end for**
6: *get Communities from User_Semantic_Matrix*
7: return *User_Semantic_Matrix*

Algorithm 6: Division of network groups of semantic construction of Tibetan Weibo users.

calculating the similarity between different groups, the two groups with the highest similarity are selected and merged into a larger group until the entire network is merged into a large group.

The similarity calculation between groups is shown in

$$M = \frac{1}{2m} \sum_{i \in C_1, j \in C_2, i \neq j} \left[B(i,j) - \frac{w_i \times w_j}{2m} \right], \tag{13}$$

where C_1 and C_2 represent two different group structures, $B(i, j)$ denotes the elements in the network adjacency matrix, w_i and w_j are the weights of nodes i and j, and m denotes the number of edges in the network structure. The specific steps are shown in Algorithm 6.

The specific steps of the algorithm are as follows:

(1) Enter the current network structure, take each node in the network as the initial group structure, and use (13) to calculate the similarity between each group structure.

(2) Select the two groups with the highest similarity, merge them into the same group structure, and calculate the similarity between the new group and other groups.

(3) Repeat the second step until the entire network merges into a large group.

Because this algorithm calculates iteratively to obtain a larger group structure with greater semantic similarity, when the algorithm terminates, the whole network will be divided into two different larger subgroups, which are the Tibetan user group and non-Tibetan user group.

5. Experiment and Analysis

5.1. Experiment Data. At present, there is no unified standard for the discovery of Tibetan user groups on online social media, and there is no public evaluation data set. Therefore, the experimental dataset needs to be obtained and marked manually. The experimental dataset of this paper is from Sina Weibo, the user data are crawled via Weibo crawling tools, and the crawling process is as follows:

(1) Randomly select non-Tibetan user and Tibetan Weibo user ID to build the initial seed set.

(2) Select the Weibo of the seed users and crawl the following information: the text content, the comment content, the comment user ID, the forwarding content, the forwarding user ID, the liking user ID, the follower's user ID, and the following user ID; then, determine whether these IDs have been crawled; if not, they will be added to the sequence to be crawled.

(3) For each user ID in the sequence to be crawled, request to obtain the user's Weibo text data, including the microblog text content posted by the user, the comment content of other users on the microblog text, and the corresponding comment user ID, and get the user ID that likes or forwards the microblog; also, add these IDs to the user series to be crawled after deduplication.

(4) The obtained attribute data and microblog data of each user are stored in the MongoDB database as the "key value."

This paper has crawled a total of 5000 Weibo users, as well as the related 600,000 Weibo text data. To verify the validity of this method, we need to do experiments with the manually labelled data. After cleaning the crawled data, a total of 200 Weibo users and 5000 Weibo data are randomly selected and manually labelled. Among them, there are 121 non-Tibetan users and 79 Tibetan users. When labelling the textual data, we classify the main semantic information into its related topic category.

5.2. Experiment Design. Since the group division algorithm proposed in this paper is based on semantic similarity, in order to verify the effectiveness of the proposed method, we first need to verify the validity of the semantic analysis algorithm used in this paper and then verify the validity of the algorithm.

(1) Semantic analysis algorithm validation:

Experiment 1. Use the original LDA model to extract semantic information at the Weibo text level.

Experiment 2. After aggregating the microblogs published by users into long texts, use the LDA model to extract the semantic information.

Experiment 3. Use the improved LDA model to extract semantic information at the Weibo text level.

Experiment 4. Use the improved LDA model to extract semantic information at the Weibo user level.

(2) Group division algorithm validation:

Experiment 5. Construct the user network structure only based on the "following" relationship between Weibo users, and then use the modularity maximization algorithm to divide the group.

Experiment 6. Build the network structure based on Weibo user semantic information, and use the modularity maximization algorithm to divide the group.

Experiment 7. Build the network structure based on Weibo user semantic information, and use the proposed group division algorithm based on similarity degree to divide the group.

5.3. Evaluation Index. The data used in this experiment are all labeled data, so the accuracy, recall rate, and $F1$ values are used as the experimental evaluation indexes. The accuracy refers to the ratio of the user number with the correct group division results in the entire experimental user sample. The recall rate refers to the ratio of the user number that has the correct group division result in the same type of users. The F value is the harmonic average of the accuracy and the recall rate. The specific formulas are as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\sum T_i}{N}, \\ \text{Recall} &= \frac{\sum T_i}{\sum (T_i + F_i)}, \\ F1 &= \frac{2 \times \text{Accuracy} \times \text{Recall}}{\text{Accuracy} + \text{Recall}}. \end{aligned} \tag{14}$$

Among them, T_i refers to the number of the correct divisions, F_i refers to the number of wrong divisions, N is the total number of samples in the experiment.

5.4. Experimental Results and Analysis. Figures 2–5 are comparison graphs of the result of semantic analysis experiments from Experiment 1 to Experiment 4.

The experimental comparison results in Figure 2 show that when using the original LDA topic model to extract text semantic information, the user-published microblogging aggregated as long text can obtain more effective results than extracting semantic information directly on the microblogging text level.

The results of Figure 3 demonstrate that by using manually selected long text training the LDA to obtain the topic distribution and updating it with the aggregated labelled

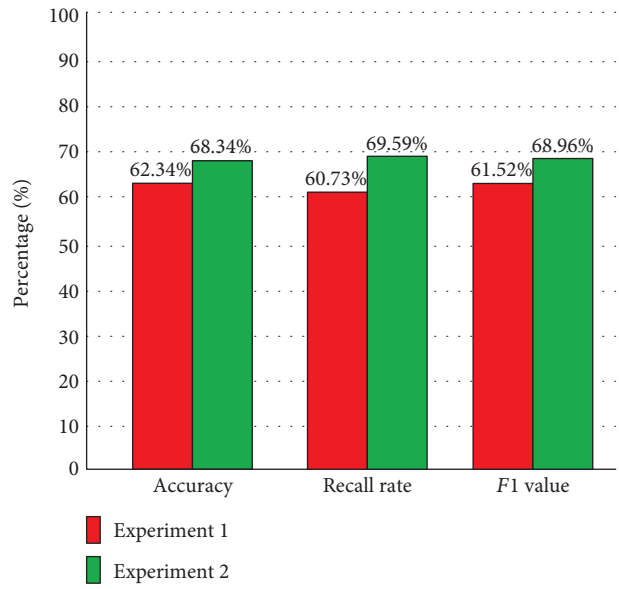

Figure 2: Comparison of Experiment 1 and Experiment 2.

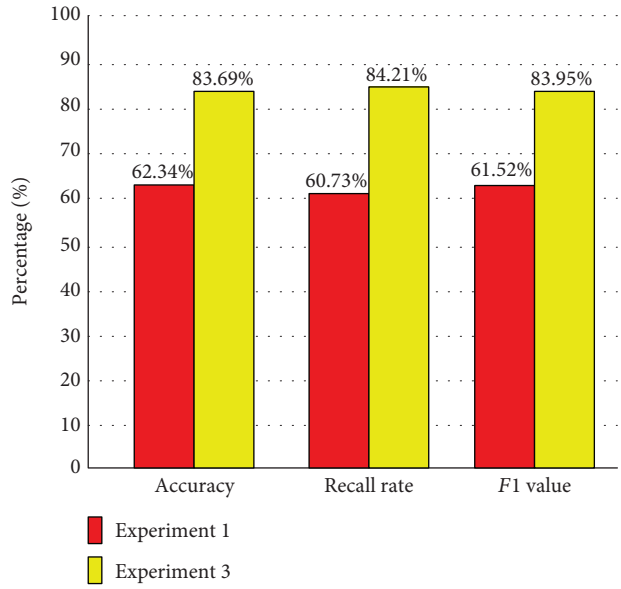

Figure 3: Comparison of Experiment 1 and Experiment 3.

Weibo texts to get the ideal topic distribution, the Weibo text content can be reasonably explained to a large extent.

From the experimental results shown in Figures 3 and 4, we can see that the improved LDA model presented in this paper has a significantly higher improvement than the original LDA model when it comes to topic extraction. This result validates that the improved LDA model presented in this paper has data sparsity, which compromises the effectiveness of topic extraction from Weibo data.

The results of Figure 5 show that it is effective to express the interested topics at the user level by aggregating the Weibo text multiple times and using it to train the LDA model to obtain a better distribution of topics.

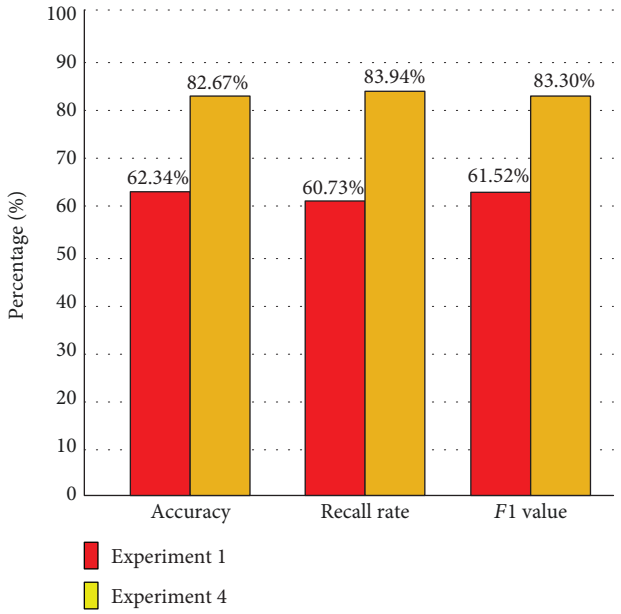

Figure 4: Comparison of Experiment 1 and Experiment 4.

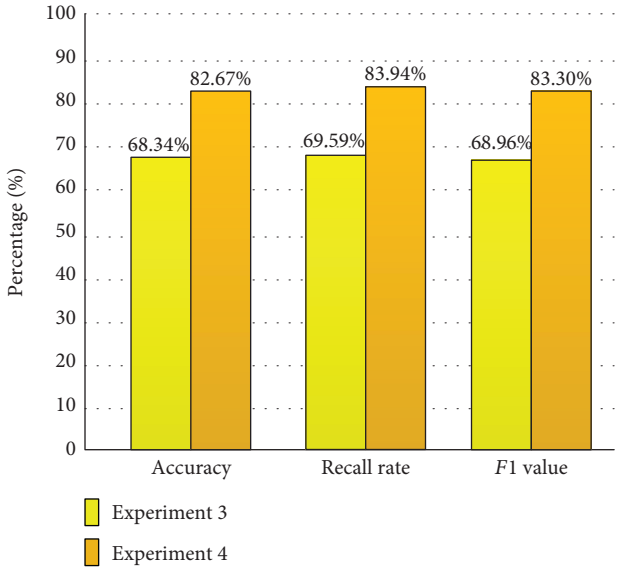

Figure 5: Comparison of Experiment 3 and Experiment 4.

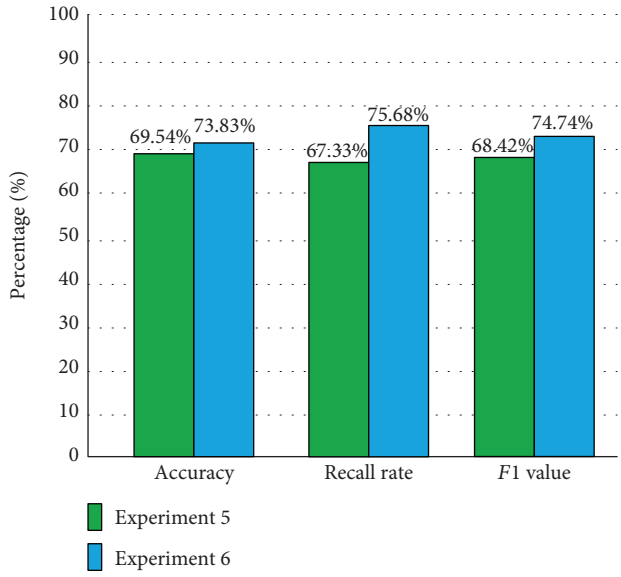

Figure 6: Comparison of Experiment 5 and Experiment 6.

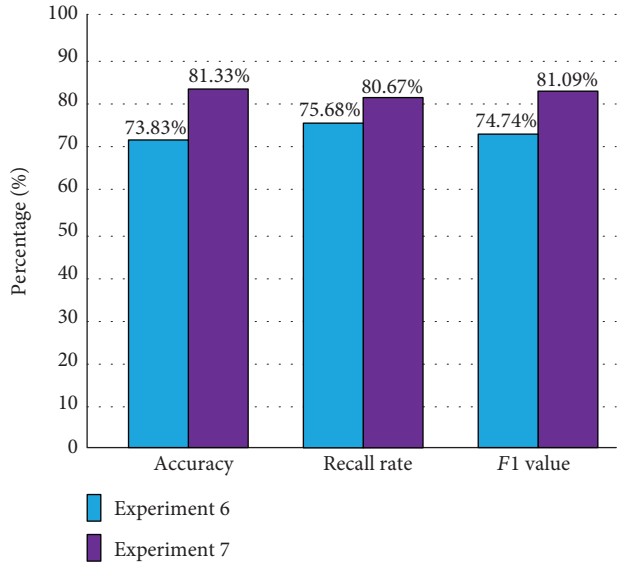

Figure 7: Comparison of Experiment 6 and Experiment 7.

Figures 6 and 7 are comparison graphs of the result of semantic analysis experiments from Experiment 5 to Experiment 7.

The experimental results in Figure 6 show that considering the users' semantic information can significantly improve the accuracy of group division in online social networks. The behaviors of online social network users express their semantic information, which means their concept, attitude, and emotion. Due to various reasons, in the real social network, the same ethnic groups tend to have similar ideas and attitudes, using the similarity between users' semantic information can effectively discover the user groups with similar semantic information, and these groups usually have a certain intimacy degree in real life.

The results of Figure 7 show that the proposed group division algorithm based on similarity aggregation is effective. Because the social structures excavated in this article have real corresponding social groups, it is reasonable to explain the topological structure of the user network by grouping them according to the semantic similarity.

6. Conclusion

This paper presents in the era of big data the user group division method based on semantic information analysis

and semantic similarity of online social networks. Firstly, the LDA topic model is improved so that it can effectively extract semantic information at the user level and microblog text level. Second, because the same ethnic groups often express similar concepts and attitudes in social networks, we analyze the semantic similarities between the semantic information expressed by users and the user semantics and construct a network structure that can describe the intimate relationships between users. Finally, the goal of optimizing the similarity between groups in the network is to realize the discovery of Tibetan user groups in microblog networks. The experimental results show that the proposed method is effective. The topological network constructed by user semantic information can reasonably represent intimacy between users. The clustering algorithm based on similarity aggregation can give out reasonable explanation by combining the actual segmentation results from real social life.

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 is no conflict of interest regarding the publication of this paper.

Acknowledgments

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

References

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] J. Golbeck and M. Rothstein, "Linking social networks on the web with FOAF: a semantic web case study," in *AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence*, vol. 2, pp. 1138–1143, Chicago, IL, USA, July 2008.

[3] X. Yu, X. Zhi-Qiang, and Y. Jing, "Semantic community detection research based on topic probability models," *Acta Automatica Sinica*, vol. 41, no. 10, pp. 1693–1710, 2015.

[4] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315, Seattle, WA, USA, August 2004.

[5] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–273, Toronto, Canada, July-August 2003.

[6] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, "Community discovery using nonnegative matrix factorization," *Data Mining and Knowledge Discovery*, vol. 22, no. 3, pp. 493–521, 2011.

[7] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 487–494, Amsterdam, The Netherlands, July 2007.

[8] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 927–936, Paris, France, June-July 2009.

[9] J. Zhuang, T. Mei, S. C. H. Hoi, X. S. Hua, and Y. Zhang, "Community discovery from social media by low-rank matrix recovery," *Acm Transactions on Intelligent Systems and Technology*, vol. 5, no. 4, pp. 1–19, 2015.

[10] S. A. Rios and R. Munoz, "Dark web portal overlapping community detection based on topic models," in *ISI-KDD '12 Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, Beijing, China, August 2012.

[11] M. Mattingly, A. Barnas, S. Ellis-Felege, R. Newman, D. Iles, and T. Desell, "Developing a citizen science web portal for manual and automated ecological image detection," in *2016 IEEE 12th International Conference on e-Science (e-Science)*, pp. 223–232, Baltimore, MD, USA, October 2016.

[12] Q. Wang and X. Fang, "Improvement of overlapping community detection based on label propagation algorithm," *Modern Computer*, vol. 12, 2017.

[13] S. Kianian, M. R. Khayyambashi, and N. Movahhedinia, "Semantic community detection using label propagation algorithm," *Journal of Information Science*, vol. 42, no. 2, pp. 166–178, 2016.

[14] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "Modeling individual topic-specific behavior and influence backbone networks in social media," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–16, 2014.

[15] A. Nocaj, M. Ortmann, and U. Brandes, "Untangling the hairballs of multi-centered, small-world online social media networks," *Journal of Graph Algorithms and Applications*, vol. 19, no. 2, pp. 595–618, 2015.

[16] B. Hu, Z. Song, and M. Ester, "User features and social networks for topic modeling in online social media," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 202–209, Istanbul, Turkey, August 2012.

[17] N. Natarajan, P. Sen, and V. Chaoji, "Community detection in content-sharing social networks," in *ASONAM '13 Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 82–89, Niagara, ON, Canada, August 2013.

[18] T. S. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," *Physical Review E*, vol. 80, no. 1, article 016105, 2009.

[19] A. Kakisim and I. Sogukpinar, "Community detection in social networks using content and link analysis," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1521–1524, Malatya, Turkey, May 2015.

[20] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *SOMA '10 Proceedings of the First Workshop on Social Media Analytics*, pp. 80–88, Washington D.C., USA, July 2010.