A Method of Retrieval Target Information of Social Network Platform Based on Overlapping Communities

Kai Zhang¹,*

¹Shandong Zhengyun Information Technology Company Limited, Shandong, China

*Corresponding author e-mail: zhangk@sdas.org

Abstract. With the advent of the 5G era, the social network platform has become the most important way for people to obtain information daily. How to retrieve the desired target information from a large amount of information has been the research direction of researchers. This paper proposes a fast overlapping community detection method based on local neighbor information to detect the target information of the social network platform. This method preprocesses the social network platform data within a certain period of time, constructs edges between highly relevant keywords to obtain a complex network. Through the fast overlapping community detection algorithm of local neighbor information, find overlapping communities, and treat each community as a target information. This method can solve the problem that the detection accuracy of one or more keywords belonging to multiple target information is low. Experiments on a social platform verify the reasonableness and effectiveness of the method.

1. Introduction

With the advancement and development of science and technology, China's 5G technology is now leading the world. Social network platforms such as WeChat, Weibo, and QQ have also become important windows for people to communicate daily. As a new era media, they have been recognized by the majority of people and institutions. By analyzing the content published by these social network platforms and detecting target information, it can play an important role in event detection, opinion mining, public opinion analysis, and real-time observation. Therefore, it is of great research significance to detect the target information of each social network platform.

At present, domestic and foreign researchers have done related research on the detection of target information on social network platform [1,2]. Literature[3] proposed a new emerging topic tracking method to detect the target information of blogs, based on local weighted linear regression(LWLR) to estimate word novelty and fading. Literature[4] proposed an MB-SinglePass topic detection algorithm for Weibo based on the structured information such as followers' followers and followers in Weibo,
and the internal associations between reposted comments. Literature[5] proposed a topic detection method for Weibo user community. Firstly, on the microblog texts published by users, using the LDA topic model to analyze the user's interest distribution, a user community-oriented microblog topic discovery method that integrates the importance of words and neighborhood graphs is presented. Literature[6] analyzes and discusses methods that affect emerging topic detection from the aspects of social network platform content burst characteristics and information dissemination models, and looks forward to the prospect of emerging topic detection.

Traditional social network platform target information retrieval algorithms based on keyword clustering often tend to divide keywords into relatively short keyword sets, that is, keywords belong to only one target information. However, in the real social network platform text, different target information often contains the same keywords. Aiming at the problem that different target retrieval information may contain the same keywords and the accuracy of the retrieved information is low, this paper proposes a fast overlapping community detection algorithm based on local neighbor information of complex networks to retrieve social network platform target information. The community is rapidly expanded based on the internal and external degrees of the nodes in the complex network. The local neighbor information area of a single node is used to calculate the membership of this node to the community, and the nodes with low membership are filtered. The experimental results prove that (this article uses Weibo as the research social network platform and Weibo topic as the retrieval target information), the algorithm in this article can identify the situation where one or more keywords belong to multiple retrieval information and can be more accurate detected target information.

2. Retrieval of target information about social network platform

The main steps of the fast overlapping community detection algorithm based on local neighbor information of complex networks to retrieve Weibo topics include five steps.

2.1. Weibo data acquisition

In keyword detection, the time distribution of candidate keywords is used to obtain a keyword set. Therefore, when obtaining data information on Weibo, two fields of time and keywords of Weibo need to be retained.

2.2. Weibo Information Text Preprocessing

Since the information text of Weibo contains irrelevant data and noise data, filtering out this information can improve the detection quality of subsequent Weibo topics. Therefore, it is necessary to filter the irrelevant data and noise data of Weibo information text before keyword detection. The main method of filtering is rules. The filtered content is: Data in the format "@username"; Emoticons in Weibo; Other useless information, including website information, specific words in Weibo, etc.

2.3. Keyword Search

Keyword retrieval mainly includes five parts: divided words, word feature filtering, word frequency statistics of time period, burst coefficient calculation, and keyword retrieval. This part first uses an open source word segmentation system to mark the characteristics of words and words on Weibo text information, selectively save verb nouns that have meaning to the target retrieval information, and count the frequency of these words in different time periods. Extensive features filter
out low-frequency words, obtain candidate keyword sets, and then obtain formal keyword sets from candidate keyword sets.

Based on the burst characteristics of Weibo keywords, the frequency of keywords in each period is calculated. The burst coefficient $G_i$ is introduced to quantify the degree to which a candidate keyword is a keyword, and the burst coefficient $G_i$ is a ratio of the occurrence probability of a candidate keyword $i$ to the occurrence probability of other unit times.

$$G_i = \frac{\max_i \left( \frac{F_{i_j}}{F_j} \right)}{\left( \sum_j \frac{F_{i_j}}{F_j} - \max_i \left( \frac{F_{i_j}}{F_j} \right) \right) / (k - l)}$$

$F_{i_j}$ represents the number of Weibo containing the word $i$ in $j$ period, $F_j$ represents the total number of Weibo in $j$ period, and $k$ represents $k$ periods. Dividing the number of Weibo appearing in the word $i$ by the total number of Weibo during a certain period of time can avoid the problem of low accuracy of keyword extraction caused by the inconsistency of the number of Weibo in different periods of time. The burst coefficient $G_i$ is used to describe the degree to which a candidate keyword is a formal keyword. A candidate keyword with a value of $G_i$ greater than or equal to a threshold value $G$ is selected to obtain a formal keyword.

2.4. Construction of complex networks

The keywords on the same topic are closely related and highly related. After the keyword set is obtained, the relevance between the keywords is evaluated, and the keywords with high relevance are constructed to obtain a complex network with keywords as nodes and close relationships between the keywords as edges.

The co-occurrence word vector $V_i$ of the keyword $i$ is formalized as follows:

$$V_i \sim \left( f_{i_1}, f_{i_2}, \ldots, f_{i_j}, \ldots, f_{i_k} \right)$$

In formula (2), $k$ represents the size of the word space, and $f_{i_j}$ represents the number of Weibo in which the word $i$ and the word $j$ appear together. The keyword similarity calculation formula based on the keyword's co-occurrence word vector is as follows:

$$\text{Sim}(i, j) = \frac{\sum_{i=1}^{k} f_{i_i} \times f_{i_j}}{\sqrt{\left( \sum_{i=1}^{k} f_{i_i}^2 \right) \left( \sum_{i=1}^{k} f_{i_j}^2 \right)}}$$

In the formula (3), $f_{i_i}$ and $f_{i_j}$ represent the number of Weibo that appear in the word $i$, the word $j$ and the word $t$, and $k$ represents the size of the word space.

The correlation between keywords describes the relationship between keywords, and the higher the correlation between keywords on the same topic. Choose a proper relevance threshold value $S$. The probability that two keywords with a correlation greater than or equal to the threshold value belong to the same topic is high, otherwise the probability is low. Construct edges between keywords with a correlation threshold value greater than or equal to $S$, and get a complex network $G$ that reflects the keywords.
2.5. Fast overlapping community detection algorithm based on local neighbor information of complex networks

After obtaining the complex network G, a fast overlapping community detection algorithm of local neighbor information can be used to detect Weibo topics, and each keyword is regarded as a node, and a topic composed of multiple keywords is regarded as a community. The basic idea of the algorithm is to use the local topology information of the network to rapidly expand the community through the seed expansion strategy. Table 1 lists the algorithm flow.

**Table 1.** Fast overlapping community detection algorithm based on local neighbor information of complex networks

| Input: Network G (V, E), threshold value S, α; |
| Output: a community division of the network--Community; |
| First initialize each node in the network as unvisited; |
| V₀=argmaxᵢ(degree(i)); |
| If visited(V₀)==0 |
| Initial Community={V₀ ∪ all neighbor nodes}; |
| Delete the node of kᵢ<sub>in</sub><kᵢ<sub>out</sub> in the Initial Community collection; |
| Single Community={Initial Community}; |
| 2 Using local information of complex networks to select a candidate node set that is currently expanding the community |
| While(newAdd!=null) |
| Candidate=map<int,double>; |
| Single Community=Single Community ∪ Add |
| newAdd=Add; |
| End while |
| 3 Calculating the wearing coefficient of each node in the Single Community through local network information |
| For each vᵢ∈Single Community |
| NeighborTemp=Neighbors(vᵢ)∩Single Community; |
| End for |
| Community=Community ∪ Single Community |
| Mark all nodes in the Single Community as visited |
| End if |

The algorithm uses the local topology of a complex network to quickly select multiple candidate nodes that may be added to the community at one time from the neighbors of the current community, avoiding the repeated calculation of the contribution of a node to the community by using the local benefit function, which greatly improves this improves the efficiency of the algorithm. Finally, the membership coefficient of each node in the community is calculated, and only the nodes with higher membership coefficients are retained, which further improves the accuracy of the algorithm for detecting overlapping communities.
3. Experiments

3.1. Data acquisition and preprocessing

Collected Weibo data from October 1, 2019 to October 7, 2019. After obtaining Weibo data information, remove noise and useless information, and retain the date and filtered Weibo content. Manually mark representative Weibo topics with common keywords in this time period. There are four main topics: "70th anniversary of founding of new China", "Movies of My People, My Country", "70th anniversary of the National Day parade", and "Songs of My People, My Country".

3.2. Keyword detection

According to the previous keyword detection method, analysis and part-of-speech filtering are performed to obtain the distribution of the word in each time period. The burst coefficient of the word and the frequency of the word in the time period with the maximum coefficient are calculated. Threshold value (frequency value in this article is 10), to get the candidate word keyword set. Then analyze the influence of the burst coefficient and threshold value on the keyword accuracy rate and recall rate and the comprehensive evaluation index F1 of the accurate recall rate. Formula (4) is the formula for calculating F1. The candidate keywords with the burst coefficient greater than 3.0 are used to obtain the keyword set. The accuracy, recall, and F1 values change with the burst coefficient, as shown in Fig. 1.

\[
F1 = \frac{\text{Accuracy} \times \text{recall} \times 2}{\text{Accuracy} + \text{recall}}
\]  

(4)

Figure 1. The change of three values with burst coefficient

3.3. Construction of complex networks

After the keyword set is obtained, the correlation between keywords can be calculated according to the keyword correlation calculation formula introduced earlier, and edges can be constructed between keywords whose correlation is greater than or equal to threshold value \(S\). In order to avoid too many keywords leading to overly complicated calculation of correlations, only the correlations between keywords that have a common occurrence relationship are calculated. Then analyze the accuracy and recall of edges and the value of F1 when \(S\) is sequentially taken from 0.1 to 0.9. The accuracy of the edge recall and the change of the F1 value with the threshold value \(S\) are shown in Fig.2. When the
similarity threshold value $S$ is 0.4, the comprehensive index $F1$ value of the accuracy and recall rate of the close relationship between keywords reaches the highest.

![Figure 2](image_url)

**Figure 2.** The change of three values with correlation thresholds

3.4. **Target retrieval information detection effect**

A complex network is constructed based on the close relationship between keywords, and overlapping communities are mined according to the fast overlapping community detection algorithm based on the local neighbor information of the complex network introduced earlier. Each community is regarded as a topic. Fig. 3 is a subgraph in a complex network constructed from keywords, and Fig. 4 is a community result obtained by performing overlapping community discovery on the subgraph in Fig. 3. In Fig. 4, $c$ indicates a community, and the number after $c$ indicates the community id. If a keyword belongs to multiple communities, they are separated by "|". It can be seen from Fig. 4 that Fig. 3 includes two topics, and Table 2 lists keywords of these topics and related Weibo content.

**Table 2.** Different topics with the same keyword "My People, My Country"

| Topic             | Topic clustering by algorithm | Relevant Weibo content |
|-------------------|-------------------------------|------------------------|
| Movies of My People, My Country | Chinese, Kaige Chen, Zheng Xu, Movie revenue, the National Day film, the Chinese film dream team | Kaige Chen, Zheng Xu and other seven directors led the Chinese film dream team; From September 30 to October 7, the total Movie revenue of the National Day film has reached 5.042 billion yuan. The movie become the first film of more than 2 billion yuan in the National Day film. |
| Songs of My People, My Country | classical melodies, expressing the motherland, Guyi Li | Those classical melodies expressing the motherland. Express love to the motherland for 30 days. Youth singing for the motherland! Guyi Li Special Edition "My People, My Country" |

The fast overlapping community detection and discovery method based on local neighbor information of complex networks can also identify situations where multiple keywords exist in multiple topics simultaneously. As shown in Fig. 5, according to the community division result in Fig. 5, it can be known that the keywords “Jinping Xi” and “President” exist in topics with ids 7 and 8. Table 3 lists the keywords of these topics and related Weibo content. It can be known from the
contents of Weibo that although different Weibo contain the same keywords, they do not belong to the same topic.

Figure 3. Complex network between keywords

Figure 4. Findings from overlapping communities

Figure 5. Networks with multiple public keywords in different communities

4. Summary

Traditional target information retrieval divides keywords into the most likely target retrieval information, and ignores the situation where multiple target retrieval information may share the same keyword. This paper proposes a community information retrieval method based on local neighbor information for a social network platform to quickly overlap target information, and applies this
method to Weibo topic detection. Experiments show that the method can effectively detect different topics of public keywords in a complex network composed of Weibo topic words, and proves the effectiveness of the method. In the future research direction, this method can be applied to target information retrieval of different social network platforms, which provides new ideas for target information retrieval of social network platforms.

Table 3. Different topics with the same keywords "Jinping Xi" "President"

| Topic | Topic clustering by algorithm | Relevant Weibo content |
|-------|-----------------------------|------------------------|
| 70th anniversary of founding of new China | Communist Party, State leaders, Tiananmen Square, Jinping Xi, China, Celebrating, President | Communist Party and State leaders headed by President Jinping Xi came to the podium of Tiananmen Square. Celebrating the 70th anniversary of the founding of the People's Republic of China. |
| 70th anniversary of the National Day parade | Party flags, national flags, military flags, Jinping Xi, President, Parade | Party flags, national flags, and military flags are displayed in the wind. Jinping Xi drove to the parking stood before the three flags, the flag salute. President Jinping Xi announces: Parade begins! |

References

[1] Abe H. Analyzing User Behaviors Based on Temporal Patterns of Sequential Pattern Evaluation Indices on Twitter[M]// Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2015.

[2] Stilo G, Velardi P. Efficient temporal mining of micro-blog texts and its application to event discovery[J]. Data Mining and Knowledge Discovery, 2015, 30(2).

[3] Huang J, Peng M, Wang H, et al. A probabilistic method for emerging topic tracking in Microblog stream[J]. World Wide Web, 2017, 20(2):325-350.

[4] Gang Zhou, Hongcheng Zou, Xiaobing Xiong, et al. MB-SinglePass: Weibo topic detection based on combined similarity [J]. Computer Science, 2012, 39 (10): 198-202.

[5] Zhixiong Liu, Caiyan Jia. Microblog topic detection method oriented to user interest and community relationship [J]. Journal of Intelligent Systems, 2016, 11 (3): 294-299.

[6] Chengcheng Gou, Pan Du, Yue Liu, et al. A Survey of Emerging Topic Detection Technologies in Online Social Networks [J]. Journal of Chinese Information Processing, 2016 (5): 9-18.