2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) | 978-1-6654-0004-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISP-BMEI53629.2021.9624388

A Dynamic Clustering Method of Hot Topics Based on User Interaction and Text Similarity

Shan Liu, Xiaoqing Wu, Jianping Chai
Faculty of Information Science and Technology
Communication University of China
Beijing, China
Email: liushan@cuc.edu.cn

Abstract—**This paper proposes a dynamic clustering method for hot topics based on user interaction and text similarity. It focuses on the analysis of the clustering process from the perspective of movement and combines the two aspects of text similarity and user interaction to comprehensively consider the topic clustering of microblogs, improve the accuracy of clustering. The simulation results demonstrate that the clustering process is dynamic and can be displayed intuitively. Moreover, the model has strong extensibility, which parameters can be added, deleted and changed according to individual needs, and can be personalized for various applications.**

Keywords- Clustering; Agent; Complex network

I. Introduction

Online social networks accelerate the spread of information, represented abroad by Twitter, Facebook and Instagram (Facebook's daily active volume in 2019 is 1.2 billion), and domestically represented by QQ, WeChat, and Sina Weibo (in 2019, the average daily active volume of Sina Weibo is 203 million). The rapid development of domestic social software provides an appropriate platform tool for People's Daily and work information exchange. According to statistics, the Internet penetration rate in China has reached 61.2%, and the total number of Internet users has reached 854 million, among which the percentage of Internet users using mobile phones is as high as 99.1% [1]. People's traditional ways of communication have completely changed and they have become accustomed to uploading their lives, moods and opinions on news on different social software. At the same time, the White Paper on Protection of Consumer Rights and Interests in 2019 pointed out that Weibo has become the main field of public opinion for the production and fermentation of various hot topics, covering more users and spreading efforts [2].

The social network contains a large amount of information. In order to help people quickly get the information they are interested in, some social software classifies the information according to the topic effectively. At present, the Weibo platform and many portal websites have also launched their own topic ranking list and various hot recommendation functions, but the effect is not good, most of them just simply list the hot topics according to the number of searches and clicks. For individuals, the research on the generation mechanism of microblog hot topics is helpful to help users quickly, accurately and comprehensively understand the hot topics they want to pay attention to. For enterprises, timely understanding of the dynamics of this field, in order to grasp the demand changes in this field to make the corresponding product promotion, quickly seize the market to gain profits; For government departments, they can keep track of the hot spots in various fields of society at any time and give correct guidance to public opinions. To sum up, the research on the generation mechanism of hot topics based on microblog data can better serve the dissemination of public opinion, facilitate people's lives, and have high social value and practical significance [3].

For the microblog topic generation mechanism, the core key point is microblog clustering. The microblogs must be classified first, and then keywords are sorted by various means to generate topics according to the Chinese grammatical structure. In this paper, a dynamic clustering method for hot topics based on user interaction and text similarity is proposed to solve the problem of microblog clustering.

II. Related Work

Jiri Kubalik [4] et al. proposed a dynamic model of agent clustering, and verified the applicability of the method through experimental results. Its main contribution is the ability to discover and visualize the agent’s communication neighborhood at runtime. This is A new clustering method that has not been tried so far.

On the basis of the model proposed by Jiri Kubalik, this paper introduces the concept of agent into microblog clustering. Agent represents both microblogs and users who post microblogs. The interaction relationship between users is defined as edges, which constructs the complex network of Weibo.

According to the social characteristics and content characteristics of Weibo, the parameters and formulas of the dynamic clustering model are redefined, and user interaction and text similarity are added as factors affecting clustering, making the model more suitable for Weibo, and the traditional k-means algorithm is combined to establish a new clustering model.

III. Proposed Clustering Method

A. Agent Definition

Considering the characteristics of the Weibo platform, the main content of the text is the core part, mainly composed of short texts; the main user interaction of Weibo's dissemination

978-1-6654-0004-6/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Communication University of China. Downloaded on November 10,2025 at 02:20:28 UTC from IEEE Xplore. Restrictions apply.

has also played a key role in the topic generation process. Therefore, the agent in this article has two meanings. The first layer is Weibo itself, and the second layer is users. Therefore, this clustering model is also established based on comprehensive consideration of the similarity of the text and the interaction between users.

Construct a user interaction network $G\ (V, E)$, V is the agent, E is the interaction relationship between users, to reflect the interaction of Weibo users, and provide a calculation basis for the user interaction attractiveness below.

B. Text Similarity

1) Editing Distanc

Editing distance refers to the minimum number of editing operations required to convert two strings from one to the other. The greater the distance between them, the more different they are. Editing operations include replacing one character with another, inserting a character, and deleting a character.

Since the research object is Weibo and the text language is mostly Chinese, there will be a big difference in meaning if there is a single word difference between Chinese words, so the preliminary judgment of editing distance is not applicable to this experiment.

2) Jaccard index

Jaccard index [5], also known as Jaccard similarity coefficient, is used to compare the similarity and difference between finite sample sets. The larger the Jaccard coefficient value, the higher the sample similarity.

In fact, its calculation method is very simple, that is, the value obtained by dividing the intersection of two samples by the union, which is equal to 1 when the two samples are exactly the same, and 0 when the two samples are completely different.

3) TF-IDF & cosine similarity

TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used weighting technique for information retrieval and information exploration. TF-IDF is a statistical method used to evaluate the importance of a word to a document set or one of the documents in a corpus. The importance of a word increases in proportion to the number of times it appears in the document, but at the same time it decreases in inverse proportion to the frequency of its appearance in the corpus.

Term frequency (TF) refers to the number of times a given word appears in the file. The TF calculation formula is:

$$TF_{\omega} = \frac{n_w}{n_{\sum w_i}} \tag{1}$$

where, n_w is the number of occurrences of the term w in the D document, and $n_{\sum w_i}$ is the number of all the terms in the D document.

However, it should be noted that some common words do not have much effect on the topic. On the contrary, some words that appear less frequently can express the topic of the article, so it is not appropriate to use TF alone. Therefore, IDF is introduced here.

Inverse document frequency (IDF), the main idea of IDF is: if there are fewer documents containing the term t, the larger the IDF, indicating that the term has a good ability to distinguish categories. The IDF of a particular word can be obtained by dividing the total number of documents by the number of documents containing the word, and then taking the logarithm of the obtained quotient. The IDF calculation formula is:

$$IDF = log\big(N/\big(n_{w\in D} + 1\big)\big) \tag{2}$$

where, N is the total number of documents in the corpus, and $n_{w\in D}$ is the number of documents containing the term w.

A high word frequency in a particular document and a low document frequency of the word in the entire document collection can produce a high-weight TF-IDF. Therefore, the value of TF-IDF is the product of TF and IDF.

$$TF - IDF = TF \times IDF \tag{3}$$

The literatures [6] [7] extract keywords based on TF-IDF, which shows that the weight value is a mainstream method for calculating word weight.

Cosine distance, also known as cosine similarity, uses the cosine value of the angle between two vectors in a vector space as a measure of the difference between two individuals.

$$cos(\theta)_{i,j} = \frac{(x_i, y_i)\cdot(x_j, y_j)}{\|(x_i, y_i)\|\cdot\|(x_j, y_j)\|} \tag{4}$$

The closer the cosine is to 1, the closer the Angle is to 0 degrees, the more similar the two vectors are, and that's called cosine similarity. Literatures [8] and [9] rely on cosine similarity to measure the similarity between content, indicating that cosine similarity is often used when calculating the similarity between texts.

4) Word2Vec

Word2Vec literally means word to vector, a method from word to vector; its professional interpretation is that Word2Vec uses a layer of neural network to map one-hot (one-hot encoding) word vectors to distributed word vectors, using Hierarchical SoftMax, negative sampling and other techniques to optimize the training speed. It is mainly used for preprocessing or analysis of natural language to help computers understand the natural language used in daily life. Literatures [10] [11] both use the Word2Vec word embedding method to construct the word vector space to measure the similarity between words.

Because the collected data is inaccurate in word segmentation during processing, or there are special words (Internet buzzwords), some words are not included in the existing trained model, and it is necessary to adjust the word segmentation results or train a new model. Therefore, this experiment does not use this similarity acquisition method.

C. User Interaction

Interactions on Weibo are mainly divided into comments, reposts, and likes. Among them, the most influential and easiest to spread interactive behavior is reposting, and this behavior is clearly marked by “@” in the Weibo body content, which is easy to obtain and does not require additional collection. "@" can be understood here as mentioning a certain user, which can mean reminding him to check the user's Weibo or tell him his Weibo is reposted by the user.

Authorized licensed use limited to: Communication University of China. Downloaded on November 10,2025 at 02:20:28 UTC from IEEE Xplore. Restrictions apply.

There are three main situations for the appearance of the "@" logo: the first is that the user wants to share this Weibo with the mentioned user, or the content of the Weibo involves the mentioned user; Second, there will be some "big V" similar to opinion leaders on Weibo, and users will tend to mention the same "big V" when discussing the same topic; third, the reposting behavior on Weibo will also be natural There is the "@" mark, and the user of "@" is the source of forwarding. These three situations can show that the interactive behavior of "@" can reflect the relevance of Weibo. Therefore, there will be an attraction between the same user of "@" or Weibo with "@" relationship.

This paper sets up an adjacency matrix based on user interaction. The value between the microblogs mentioning the same user is set to 1, the direction is two-way, and the value between the microblogs with the mentioned relationship is set to 1, and the direction is the user's Weibo pointing to The Weibo of the mentioned user, that is, the Weibo, will be attracted by the Weibo of the mentioned user.

IV. DYNAMIC CLUSTERING MODEL

A. Model Definition

Based on the agent dynamic clustering model proposed in [4], according to the application scenario of Weibo, some parameters are redefined, and a new model based on Weibo hot topic clustering is obtained.

Literature [4] pointed out that agents are affected by three kinds of forces, so that they move or their trajectories are changed. These three forces are attraction, friction, and repulsion.

The size of the attractiveness is related to the similarity of the text and the degree of user interaction. We use the TF-IDF and cosine similarity mentioned above to measure the similarity between Weibo texts, and the value range is (0, 1). The user interaction degree adopts the adjacency matrix based on user interaction described above, and the value is 0 or 1. Both matrices need to be scaled numerically according to the actual situation, and the scale mainly depends on the position range of the agent.

The quantitative mathematical formula for attractiveness is:

$$\vec{F}_{i,j}^{a} = k_a \cdot S \cdot \vec{u}_{i,j} + k_b \cdot R \cdot \vec{u}_{i,j} \tag{5}$$

where, $\vec{F}_{\mathrm{i},j}^{a}$ is the attraction of agent i from agent j, S is the text similarity matrix of nXn, R is the microblog user interaction matrix of nXn, k_a,k_b are constant coefficients, $\vec{u}_{i,j}$ is the unit vector from agent i to agent j.

The quantitative mathematical formula of friction is:

$$\vec{F}_{i}^{f} = -k_f \cdot \vec{v}_i \vec{u}_{i,j} \tag{6}$$

where, $\vec{F}_{\mathrm{i},j}^{a}$ is the friction force received by agent i, $\vec{v}_i$ is the moving speed of agent i, and k_f is the constant coefficient.
The quantitative mathematical formula of repulsive force is

$$\vec{F}_{i,j}^{r} = -k_r \cdot \frac{m_i \cdot m_j}{{d_{i,j}}^2} \cdot \vec{u}_{i,j} \vec{u}_{i,j} \tag{7}$$

where, $\vec{F}_{\mathrm{i},j}^{r}$ is the repulsive force from agent j received by agent i, k_r is a constant coefficient, $d_{i,j}$ is the distance between the two agents, the mass of agent m_i should be the same as the popularity of blogs (influence or attention) is directly proportional. In this experiment, m_i is set to 1.

The resultant force received by agent i is:

$$\vec{F}_i = \sum_{\substack{j=1 \\ i \neq j}}^{n} \vec{F}_{i,j}^{a} + \vec{F}_{i}^{f} + \sum_{\substack{j=1 \\ i \neq j}}^{n} \vec{F}_{i,j}^{r} \vec{u}_{i,j} \tag{8}$$

As long as the time interval is set, the force that the agent receives at each time interval can be calculated, so that the speed and position of the agent can be calculated at that moment (set a time period to keep the state of motion unchanged). The motion equation of agent i is:

$$\vec{v}_i\big((n+1)\Delta T\big) = \vec{v}_i(n\Delta T) + \frac{\vec{F}_i(n\Delta T)}{m_i} \Delta T \vec{u}_{i,j} \tag{9}$$

$$\vec{p}_i\big((n+1)\Delta T\big) = \vec{p}_i(n\Delta T) + \frac{1}{2}\frac{\vec{F}_i(n\Delta T)}{m_i} \Delta T^2 + \vec{v}_i(n\Delta T)\Delta T \vec{u}_{i,j} \tag{10}$$

After getting the final position coordinates, you can roughly see that there are several types of topics, get the number of clusters k, and use the k-means algorithm to determine the cluster to which each point belongs.

B. Simulation experiment

The data used in this experiment comes from some microblogs about the "epidemic" from December 8, 2020 to January 28, 2021, with a total of 1886 articles. The data mainly consists of tweets discussing three topics, “Covid-19 in Chengdu”, “Covid-19 in Beijing”, “Covid-19 in Tonghua”. The dimensions of the data include “user name”, “content”, “topic”, “@ user”, “release time”, “release location”, “release tool”, “like”, etc.

The space we set is a two-dimensional plane, the position coordinates of agent i are $p = [p_i, p_j]$, p_i is the value of the abscissa of agent i, and p_j is the value of the ordinate of agent i. Similarly, we divide the force as well. Divided into two directions, x-axis and y-axis.

The initial location setting conditions of the agents are: the x-axis is the time axis, from left to right, the microblogs are arranged in order from the farthest to the nearest according to the release time. The consideration for this is that the microblogs with similar time are more likely to belong to the same topic. Therefore, distributing their initial positions according to time will help the accuracy of clustering.

Generate a one-dimensional array consisting of incremental random numbers between 0 and 1, as the x-axis coordinates of the point, and the y-axis coordinates as the random number between (0, 1), and then enlarge them according to a certain ratio. In this simulation they are magnified 5000 times. According to the coordinates of the agent, the values in the text similarity matrix and the user interaction matrix are both magnified 1000 times. k_a， k_b are set to 0.5, k_f, k_r are set to 0.1. The initial distribution of Agent is shown in Fig. 1.

Authorized licensed use limited to: Communication University of China. Downloaded on November 10,2025 at 02:20:28 UTC from IEEE Xplore. Restrictions apply.

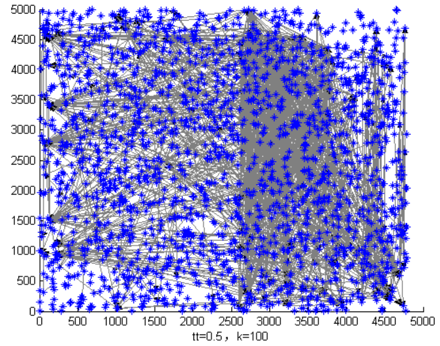
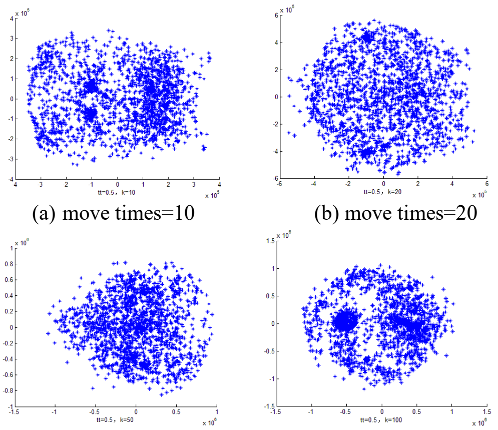


Figure 1. The initial distribution of Agent

The results in different times are shown in Fig. 2 (a) (b) (c) (d) when the time interval is 0.5 and the number of moves is 10, 20, 50, and 100. Due to the existence of the edges of the complex network, it is not convenient to see the effect of clustering at this time, so we hide the edges.

From Fig.2 (d), we can see two obvious aggregation points and the slightly scattered point set below, so we assume there are three clusters. After K-means processing on the obtained position matrix, the classification of each point can be obtained, as shown in Fig.3.



(a) move times=10 (b) move times=20

(c) move times=50 (d) move times=100

Figure 2. Dynamic clustering results at different times.

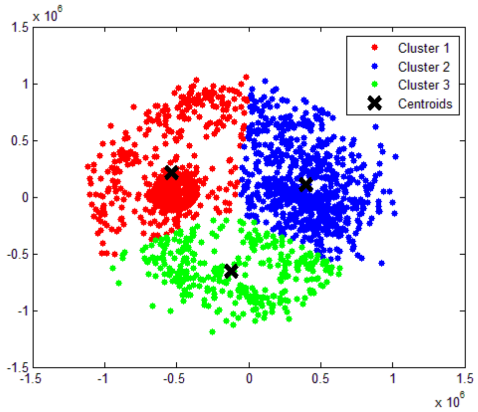


Figure 3. The classification situation after k-means

C. *Effect Evaluation*

Then, we compare it with Fig. 4 which shows the real situation of clustering (the real classification of the data set is known in advance), and different classes are marked with different colors. Among them, “Covid-19 in Chengdu” corresponds to "Cluster 2", “Covid-19 in Beijing” corresponds to "Cluster 1", and “Covid-19 in Tonghua” corresponds to "Cluster 3" in Fig.3.

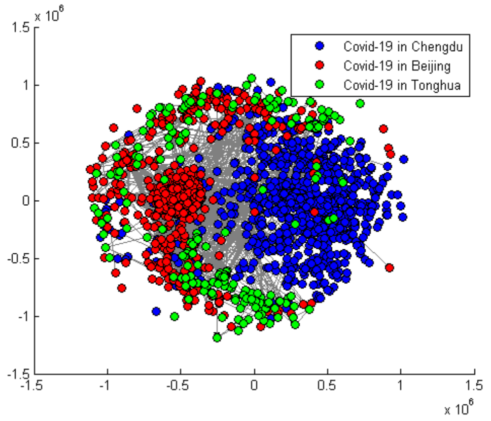


Figure 4. the real situation of clustering

According to a simple comparison of images, we can find that the clustering effect achieved by the dynamic clustering model is consistent with the real situation. Next, we will illustrate the accuracy of the clustering model with more accurate numbers. We have made statistics on the number of correctly classified items in each category, compared with the actual data, and calculated the correct rate of each category and the overall correct rate. What’s more, this paper compares its performance with other commonly used models, shown in table I. The compared models are LDA model and K-means model because the topic model has become a hot spot in the field of Weibo comment analysis [12], and K-means is also classical clustering model.

TABLE I. COMPARISON OF THE EFFECT WITH OTHER MODELS

Cluster	Cluster 1	Cluster 2	Cluster 3	total
Real number	617	1017	252	1886
Correct number (Dynamic Clustering)	484	823	126	1433
Correct number (LDA)	114	898	158	1170
Correct number (K-means)	330	389	252	971
Correct rate (Dynamic Clustering)	0.784440843	0.809242871	0.5	0.759809
Correct rate (LDA)	0.184764992	0.882989184	0.626984127	0.620361
Correct rate (K-means)	0.534846029	0.382497542	1	0.514846

From the data in the table, we can see that the accuracy of clustering is about 80% for the "Cluster 1" and " Cluster 2", and the accuracy of " Cluster 3" is only 50%. However, the actual

Authorized licensed use limited to: Communication University of China. Downloaded on November 10,2025 at 02:20:28 UTC from IEEE Xplore. Restrictions apply.

number of items in the third category is also the least, so it is speculated that this clustering method is more suitable for objects with a large amount of data. Taken together, the overall accuracy of the clustering method is 76%. Compared with the other two clustering models, the dynamic clustering model has a better effect.

V. Conclusion

The dynamic clustering method of hot topics based on user interaction and text similarity proposed in this paper visually and dynamically presents the clustering process, and makes the clustering process more intuitive and easier to control under the premise of ensuring accuracy. In the follow-up work, the influence of environmental factors on clustering will be considered, that is to say, the external force provided by the external environment will be added to make the model more personalized and at the same time improve the accuracy; further adjust the parameters to make the results closer to The real situation; adding more influencing factors, such as considering the user's personal information (interests, age, living area, etc.), so that more powerful support and reference will be obtained from the user level. In the future, research work on topic evolution will also be carried out based on this clustering model, because this model itself is dynamic and will objectively show the fission and fusion of topics after being added to the timeline.

Acknowledgment

The work of this paper was supported by the National Natural Science Fund of China (Project 61901421), the National Social Science Fund of China (Project 19ZDA272), the Fundamental Research Funds for the Central Universities, State Key Laboratory of Media Convergence and Communication, Communication University of China.

References

[1] China Internet Network Information Center. The 44th Statistical Report on Internet Development in China. (2019-08-30)[2020-03-16]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/P0201908303567 87490958.pdf.

[2] Black cat complaints. 2019 Consumer Rights Protection White Paper. (2020-03-15) http://zhongc e.sina.com.cn/article/view/39122.

[3] W. Wei. Research on hot topic discovery method and system design based on Weibo data stream. Beijing Jiaotong University, 2018.

[4] J. Kubalik, P. Tichy, R. Sindelar and R. J. Staron, "Clustering Methods for Agent Distribution Optimization," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 1, pp. 78-86, Jan. 2010, doi: 10.1109/TSMCC.2009.2031093.

[5] Y.K. Men,M.D. Qian,Z. Yu,J.Z. Teng, S.K. Chen, X. Yan.Research on text differentiation based on retrieval reranking model.Electrical Measurement and Instrumentation:1-8[2021-04-18].

[6] X. Li, "A Method for Extracting Keywords from English Literature Based on Location Feature Weighting," 2020 IEEE 20th International Conference on Communication Technology (ICCT), Nanning, China, 2020, pp. 1457-1460, doi: 10.1109/ICCT50939.2020.9295829.

[7] X.H. Wang, J. Cao, Y. Liu, S. Gao and X. Deng, "Text clustering based on the improved TFIDF by the iterative algorithm," 2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM), Kuala Lumpur, Malaysia, 2012, pp. 140-143, doi: 10.1109/EEESym.2012.6258608.

[8] D. Soyusiawaty and Y. Zakaria, "Book Data Content Similarity Detector With Cosine Similarity (Case study on digilib.uad.ac.id)," 2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA), Yogyakarta, Indonesia, 2018, pp. 1-6, doi: 10.1109/TSSA.2018.8708758.

[9] X. Wang, Z. Xu, X. Xia and C. Mao, "Computing User Similarity by Combining SimRank++ and Cosine Similarities to Improve Collaborative Filtering," 2017 14th Web Information Systems and Applications Conference (WISA), Liuzhou, China, 2017, pp. 205-210, doi: 10.1109/WISA.2017.22.

[10] M. Al-Amin, M. S. Islam and S. Das Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 2017, pp. 186-190, doi: 10.1109/ECACE.2017.7912903.

[11] H. Tian and L. Wu, "Microblog Emotional Analysis Based on TF-IWF Weighted Word2vec Model," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018, pp. 893-896, doi: 10.1109/ICSESS.2018.8663837.

[12] D. Wu, R.X. Yang, C. Shen.Research on Weibo Comment Clustering Algorithm Based on Weighted Emotional Topic Feature Words.Modern Electronic Technology,2020,43(23):67-71+75.

Authorized licensed use limited to: Communication University of China. Downloaded on November 10,2025 at 02:20:28 UTC from IEEE Xplore. Restrictions apply.