Journal of Computer and Communications, 2018, 6, 1-14
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Network Hot Topic Discovery of Fuzzy Clustering Based on Improved Firefly Algorithm

Zhenpeng Liu[1,2], Jing Dong[1], Bin Zhang[2*], Mengjie He[1], Jianmin Xu[3]

[1]School of Electronic Information Engineering, Hebei University, Baoding, China
[2]Center for Information Technology, Hebei University, Baoding, China
[3]School of Cyber Security and Computer, Hebei University, Baoding, China
Email: *zb@hbu.edu.cn

How to cite this paper: Liu, Z.P., Dong, J., Zhang, B., He, M.J. and Xu, J.M. (2018) Network Hot Topic Discovery of Fuzzy Clustering Based on Improved Firefly Algorithm. *Journal of Computer and Communications*, **6**, 1-14.
https://doi.org/10.4236/jcc.2018.68001

Received: July 5, 2018
Accepted: July 31, 2018
Published: August 2, 2018

Abstract

The existing fuzzy clustering algorithm (FCM) is sensitive to the initial center point. And simple clustering of distance can neither discovery hot topics on the Network accurately nor solve the problem of semantic diversity in Chinese. Aiming at these problems, an improved fuzzy clustering method based on dynamic adaptive step firefly algorithm (FA) was proposed. The clustering center was optimized by improved FA, and the FCM was used to complete the final clustering. First, the step length was adjusted adaptively in the current iteration, and the relationship between fireflies was established according to text similarity, then the topic influence value was applied to fuzzy clustering algorithm to improve fitness function optimization. In this process the topic was categorized into the closest class to the cluster center, which can reduce the impact of topic variation. Finally, according to the level of influence value got hot topics. By collecting real data from Sina micro-blog, the effectiveness of the algorithm was verified by experiments, and the accuracy of topic discovery was improved greatly.

Keywords

Topic Discovery, Firefly Algorithm, Dynamic Adaptive Step Size, FCM, Micro-Blog

1. Introduction

The rapid development of social media has brought great challenges to the research of complex networks. Users can make views according to their own moods; it can form various themes through forwarding, comment and so on, with the development of this, generating certain social trends. This social trend can reveal the relevant things that are happening at the moment [1]. If we can

find these in a timely and accurately manner, then we can provide various countermeasures related to it. Hot topics are more representative of the public's recent concerns, and it has a deeper influence on the development of society. Untrue or negative information has triggered a series of network public opinion problems. Therefore, the research on hot topic discovery has received extensive attention from scholars at home and abroad [2].

The research on the network hot topic discovery text is mainly the two directions of feature extraction and clustering algorithm. The clustering methods of traditional topic discovery include hierarchical clustering, cure algorithm, single-pass, DBSCAN etc [3]. All of these algorithms are hard clustering algorithms. These algorithms have their own limitations for the fast change of the network language and the variant of the Chinese semantic diversity, which leads that users cannot discover hot topics in the network timely and accurately. Feng Liguang *et al.* [4] used the FCM parallel algorithm to discover hot topics; however, the fuzzy clustering algorithm has the disadvantage of the common mean algorithm because it is a local search algorithm. For such an algorithm, if the initial value is improperly selected, it is easy to converge to the local minimum point. But group intelligence algorithms have strong global parallelism. Fiho *et al.* [5] combined the improved particle swarm algorithm with two kinds of mixed fuzzy clustering (FCM-IDPSO and FCM2-IDPSO) and made the problem of FCM trapping into local optimum and sensitivity to initial value of clustering center improved; Wu *et al.* [6] combined simulated annealing algorithm and particle swarm optimization algorithm, and proposed an enhanced adaptive weight fuzzy clustering; Jiang *et al.* [7] proposed a data set classification algorithm combining genetic algorithm and FCM; Yang Fei *et al.* [8] combined genetic algorithm with clustering and applied it to topic discovery to improve the efficiency of topic discovery. Babak *et al.* [9] used multi-target enhanced firefly algorithm to discover associations in complex networks, multi-targets and adaptive probability variation increases the accuracy of topic discovery.

In summary, the hot topic discovery method which based on word frequency is difficult to deal with the challenges of derived variant naming entities and new words such as heteromorphic words and polysemous words; topic discovery methods based on the heat is difficult to find hidden topics with low heat but sudden strongness. Aiming at the shortcomings of the FCM algorithm in the hot topic discovery process, the dynamic adaptive step firefly algorithm is used to optimize the FCM algorithm. The combination of text influence with the FCM algorithm is applied to hot topic discovery, which can identify hot topics with low influence but sudden strongness. And it can solve some problems caused by topic variation, thereby improving the accuracy of hot topic discovery.

2. Related Knowledge

2.1. Hot Topic Discovery Process

The hot topic discovery algorithm may include the following steps: data collec-

tion; it is mainly use crawler technology to obtain information; preprocessing is the initial operation of word segmentation, removal of stop words, etc for the text we have got; vector space model is established after TF-IDF feature extraction, this operate is to make it easier to compare the similarity. Clustering is the most important part of this paper. The flowchart of the algorithm for topic discovery is shown in Figure 1.

2.2. Firefly Algorithm

Firefly Algorithm is put up by a Cambridge scholar Xin-She Yang based on the glow behavior of fireflies in nature [10], it has a good use prospect in the optimization of continuous space [11]. This article assumes that there are n fireflies, the corresponding text number is n, fireflies i and j move position according to their mutual attraction. If the firefly's brightness is high, it will attract the lower brightness fireflies move to it, and then complete the position optimization. The following are the relative brightness fluorescence formula, the firefly mutual attraction formula, and the position update formula. Among them, α represent the step size, β_0 represent the maximum attraction, I_0 represent brightest firefly brightness, r_{ij} represent the distance between fireflies i and j.

$$I = I_0 \mathrm{e}^{-\gamma r_{ij}} \tag{1}$$

$$\beta_{(r)} = \beta_0 \mathrm{e}^{-\gamma r_{ij}^2} \tag{2}$$

$$\begin{aligned} &X_i(t+1) \\ &= X_i(t) + \beta\left(X_j(t) - X_i(t)\right) + \alpha\left(rand - 1/2\right) \end{aligned} \tag{3}$$

2.3. FCM Algorithm

The FCM algorithm is a popular fuzzy clustering algorithm proposed by Dunn in 1973 [12]. Assuming that n is the number of elements in the data set, we divide it into c classes, that is, there are c class centers. And define the minimum objective function as follows:

$$J_{\mathrm{FCM}} = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} d_{ij}^{2} \tag{4}$$

$d_{ij} = \left\| x_i - c_j \right\|$ express the Euclidean distance of the sample x_i to the cluster center c_j, $u_{ij} \in [0,1]$ represent the degree of membership of sample i belonging

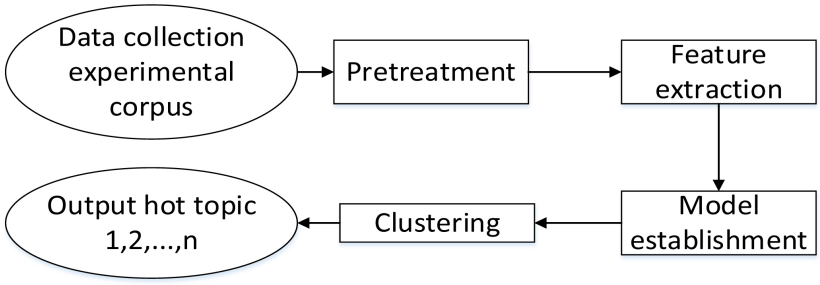

Figure 1. Topic discovery process.

to cluster center j, we define that the larger the value of u_{ij} the probability of belonging to this class is higher. To minimize the value of the target function, the objective function value is satisfied $\sum_{j=1}^{c} u_{ij} = 1$. According to Dunn [12], by using the Lagrange method, the membership degree and class center formula are as follows:

$$u_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\left(\frac{2}{m-1}\right)} \tag{5}$$

$$c_j = \sum_{i=1}^{n} \left(x_i u_{ij}^m \right) \Big/ \sum_{i=1}^{n} u_{ij}^m \tag{6}$$

3. Improved FCM Network Hot Topic Discovery Algorithm Based on Firefly Algorithm

The traditional FCM algorithm is sensitive to noise and outliers, it lead to the FCM algorithm is easy to fall into local minimum, and the selection of the initial center has a strong influence on the final clustering effect, which makes the discover result of hot topics are not ideal. The firefly algorithm has the advantage of not relying on the initial clustering center, and can overcome the shortcomings of FCM. But the FA has its own limitations, it is easy to fall into the local optimum, resulting in the solution accuracy is not high. So we propose the DASFA-FCM algorithm.

3.1. Optimized Firefly Algorithm

With the increase of iterations, the firefly swarm will gather near the optimal value of standard FA. In this way, the distance between the optimal value and other fireflies is small. When approaching the optimal value, it is likely that the distance of the firefly's movement is greater than the distance from the optimal value, so that the firefly will skip the optimal value when updating its position, which leads to a decrease in the optimal solution discovery rate.

The optimal solution and step size largely determine the convergence performance of the algorithm. In the standard FA, the step value is fixed, and all fireflies have a fixed step size during the iteration. It is easy to get the algorithm into local optimal and premature convergence [13]. In mitigate this state, according to the variable step size firefly algorithm of Yu *et al.* (VSSFA) [14], we propose an improved location update method: dynamic adaptive step firefly algorithm (DASFA). Use dynamic step size instead of the fixed step, the step size is automatically changed according to the current number of iterations. In the early stage of the iteration, the firefly has a larger step size and a larger search space, thus ensuring that global search optimization can be achieved. As the number of iterations increases, the step value decreases gradually, and each firefly searches for its own range until they find the most suitable solution. In the later iteration, in order to prevent it from skipping the op-

timal solution step, it does not need to move a lot. At the beginning, DASFA has a better global search capability, and is positioned at a faster speed near the global optimum solution. When the optimal solution is found, the current iteration and the change of step length are stopped to prevent falling into local optimum. The improved adaptive step size calculation formula using nonlinear equations is as follows:

$$\alpha(t)=\begin{cases}\alpha_{(0)}*\left(1-(t-1)/T_{\max}\right)^{\pi} & t\geq 1\\ \alpha_{(0)} & t=0\end{cases} \tag{7}$$

$\alpha_{(0)}$ is the maximum step size, which is also the initial step size at $t = 0$, t is the current number of iterations, and $T_{\max}$ is the maximum number of iterations. The improved position update formula is:

$$X_i(t+1)=X_i(t)+\beta\left(X_j(t)-X_i(t)\right)+\alpha(t)\left(rand-1/2\right) \tag{8}$$

X_i and X_j represent the spatial position of two fireflies i and j, and rand is a random factor on [0, 1], it obeys uniform distribution.

3.2. Fitness Function

The spread of topics is based on the relationship between the number and time between the publishers and forwarders, reviewers, and readers. The higher the value of a topic's attention is, the higher the influence of the topic is, and it is most likely to become a hot topic. Thus, according to Qiu Jiangnan *et al.* [15], we use forwarding amount, number of comments, and number of praises as the influence factors of the topic, the formula of the influence of each text X_i is:

$$\text{Attraction}\left(X_i\right)=\log_2^{\left(2+sf(X_i)+pc(X_i)+lz(X_i)\right)}\left(m=1,2,\cdots,n\right) \tag{9}$$

Among them: $f(X_i)$, $z(X_i)$, $c(X_i)$ respectively indicate the number of forwards, comments, and praises of the i^{th} text. Where s, p, l is the weighting factor, their sum is 1. Because in the propagation of the topic, users have a higher probability of praising content than they comment the topic of the text, so the probability of l is given a higher weight value.

Generally, the trend of hot topics is: germination period, outbreak period, stationary period, turning period, and decline period. During the germination period, as the influence of the topic increases, the attention continues to increase. When the outbreak period is reached, the degree of attention reaches the maximum; during the stationary period, the degree of attention changes little and after that the level of attention continues diminish. It can be seen from the fluorescence brightness formula that the fluorescence brightness of fireflies in nature will decrease with the distance increase and the propagation of intermediate medium, which is similar to the trend of the topic of change over time in the topic propagation process. Based on this, the influence of hot topic is corresponding to the brightness of firefly. When the influence of a topic is high, it can be regarded as the most bright firefly position. By com-

paring with the influence of surrounding topics, it can update the position iteratively and finally find the optimal solution. Due to the relative brightness of the firefly is related to the objective function value $I(X_i) \propto F(X_i)$. If the value of the objective function is smaller, it indicates that the spatial location is better, and the topic influence is greater. So the topic is more likely become the cluster center. In this way, even if the topics are far apart, we can discover the topic as well if it has a highly influence. Thereby, it can reduce the outliers in space. Similarly, if the distance is close to the optimal position but the influence of the topic is low, this means that the topic does not have a certain representative and can be ignored or not clustered into an optimal center position. The updated fitness function is:

$$F(X_i) = \frac{1}{A(X_i)} \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} d_{ij}^{2} \tag{10}$$

3.3. Similarity Calculation

The clustering requires ensure low similarity between classes and high intraclass similarity [16]. Each text is represented as a vector, the similarity formula between text *i* and text *j* is expressed as follows:

$$sim(i,j) = \frac{\sum_{k=1}^{n} w_k(i) \times w_k(j)}{\sqrt{\sum_{k=1}^{n} w_k^2(i)} \times \sqrt{\sum_{k=1}^{n} w_k^2(j)}} \tag{11}$$

$w_k(i)$, $w_k(j)$ represent the weights of the k^{th} feature words of text *i* and text *j* respectively, $sim(i,j) \in [0,1]$.

3.4. Fuzzy Clustering Based on Improved Firefly Algorithm

This paper combines the DASFA with FCM to correspond each of the space sample points to each firefly. The text influence value corresponds to the firefly brightness; the similarity corresponds to the membership value and the maximum attractiveness. If the similarity is within the threshold value, it indicates that the two fireflies are attracted, and there is a membership relationship between the two texts, otherwise there is no related link between the text. We found the similar topics according to the similarity of the text, and clustering is achieved through the attraction between fireflies. In this process, the optimization area is adjusted adaptively through the step length, according to the comparison of fitness values, we can find out the initial clustering center and then using FCM for the last clustering. In this process, according to the distance from the text to cluster center classify topics into the closest class. We can get the hot topic until the end of the termination condition is reached. Specific steps are as follows:

Input: preprocessed micro-blog text;

Output: hot topics after clustering.

DASFA-FCM proceed as follows:

① Initialization parameters: γ, $T_{\max}$, m, generate initial population $X_i (i = 1, 2, \cdots, n)$, n indicates all micro-blog texts, k represents the number of initial cluster centers, initializing the position of each firefly.

② Calculating the influence value $A(X_i)$ of each firefly according to Formula (9).

③ Calculating similarity between two texts(comparison of each micro-blog text and class center). when $sim(i, j) < \varepsilon$, the value of β_0, u_{ij} are 0; when $sim(i, j) \geq \varepsilon$, all are 1. In this moment, the mutual attraction between fireflies is calculated according to Formula (2).

④ According to Formula (7), calculating the dynamic adaptive step length under the current iteration.

⑤ Calculating fitness function $F(X_i)$, $F(X_j)$; if $F(X_i) < F(X_j)$, it shows that the firefly i influence is bigger than j, firefly i is in a better position than j, so firefly j moves to i, update each firefly position according to Formula (8).

⑥ Repeating steps ③ to ⑤ until the maximum number of iteration is reached. We can get the center of the cluster with the most influential fireflies. The number of the cluster center is C.

⑦ Based on the initial class centers found above, calculating the cluster center and membership matrix.

⑧ Calculating the distance $d_{ic} = \|x_i - x_c\|$ from the micro-blog text i to the cluster c, and classifying topics into the nearest cluster center.

⑨ Repeating steps ⑦ and steps ⑧. If the termination condition is reached, the location and influence of the most influential firefly will be output, and the result after clustering, otherwise continue.

⑩ We get hot topics based on the arrangement of influence values, output the top 50% topics.

The algorithm flowchart is shown in Figure 2.

One research shows that the topic similarity published in the near time is relatively higher, but due to the limitation of Chinese semantic diversity and topic frequency-based topic discovery, the general similarity comparison can not accurately identify the topic of variation. Therefore, by spatial distance, we divide the topic into the class closest to the cluster center. So that, for the topic of variation, if the similarity comparison cannot be classified accurately, the partial topic can also be classified into the correct cluster. By doing this, we can solve the problems caused by a small number of variation topics and improve the accuracy of topic discovery. In addition, due to the rule of topic evolution, it is considered that the influence value which is larger can represent the hot topics, so outputting the previous topic.

4. Experimental Results and Analysis

4.1. Data Preprocessing

This experiment uses the method of web crawler to extract real experimental data. From the Sina Weibo website, we get 8126 pieces of micro-blog data from December 17 to 28, 2017, and randomly select 6 topics, a total of 4967 micro-blog datas. We labeled it as experimental data set M, the data including: Topic 1, Xijia (978 strips); Topic 2, 396 Mathematics (349 strips); Topic 3, Discovery of the second solar system (152 strips); Topic 4, The Imperial Palace Response (263 strips); Topic 5, Jiangge Incident (1900 strips); Topic 6, Liu Yifei and Huang

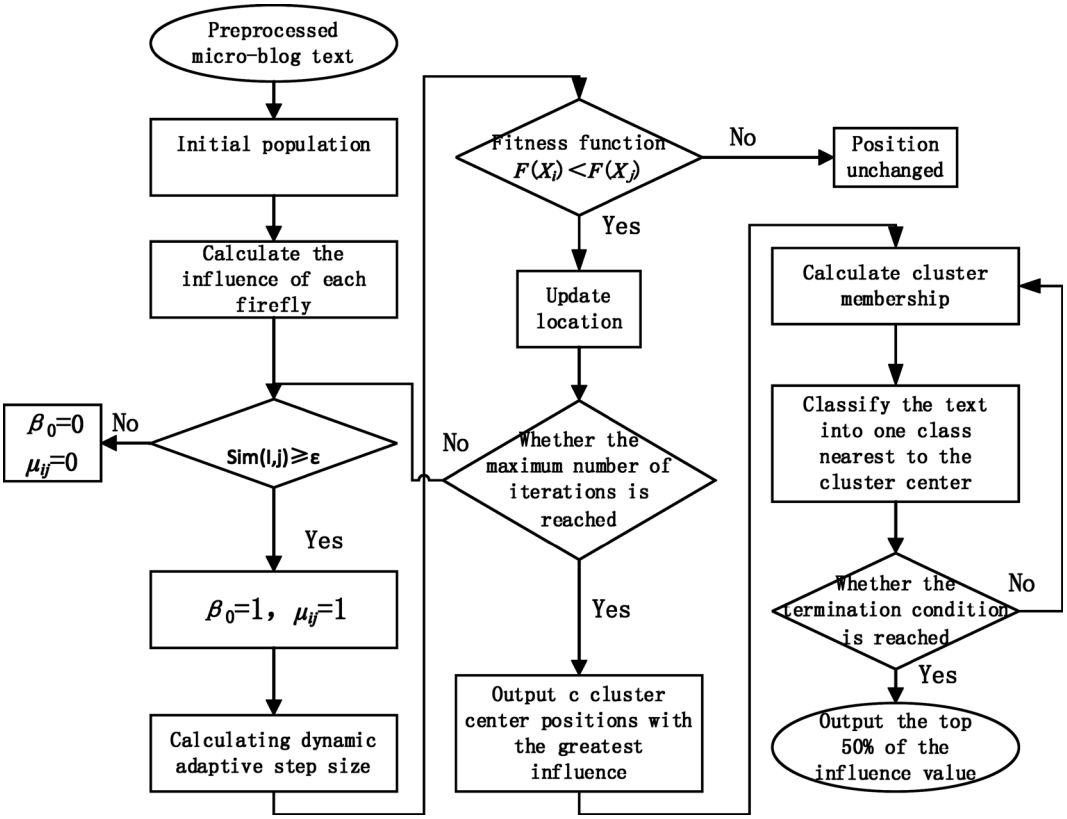

Figure 2. The improved algorithm flow chart.

Xiaoming (1325 strips); we also crawling other data from February 2018 about Finance, Commercial Housing, Winter Olympics, Mobile Communications Conference, Syria, and Automobile. Each topic has 1000 pieces of micro-blog data, marked as experimental data set N; each piece of data contains micro-blog text, user ID, forwarding amount, praises, comments, publishing time, and link URL. We delete words, remove stop words for text segmentation, and remove less than 10 words after this process. The data sets M and N are subtracted by the feature word treaty to obtain 26,996 and 37,941 feature words. The parameter fuzzy index $m = 2$; weight coefficient: $s = 0.3$, $p = 0.2$, $I = 0.5$; $T_{\max} = 200$; the light absorption coefficient $\gamma = 1$.

4.2. Evaluation Standard

In this paper, the clustering effect is evaluated by common topic detection evaluation criteria: precision *P*, recall *R*. We use *F*-measure to comprehensively evaluate clustering performance. The formula is as follows:

$$P = \frac{A}{A+B},\ R = \frac{A}{A+C},\ F = \frac{2 \times P \times R}{P+R}$$

A: is the number of samples in the sample set that are originally belonging to the cluster and are correctly clustered into the cluster; *B*: is a sample that does not belong to the cluster but is clustered into the cluster; *C*: is the number of samples that belong to the cluster but are clustered into other clusters.

4.3. Experimental Results and Analysis

1) Initial value determination

The initial step value setting has a great impact on the clustering results. Figure 3 shows the value of $\alpha_{(0)}$ in the range of 0 to 1 at intervals of 0.1.Comparison of the corresponding values of *P*, *R*, *F* in the range of values. As can be seen from the figure, $\alpha_{(0)}$ reaches a maximum value around 0.8. However, in order to accurately determine the value, we re-select two points which is nearby 0.8, we select 0.75, 0.85 to conduct the experiment again. Finally, the performance values of the two experiments are combined together. The results show that 0.8 is a turning point. When $\alpha_{(0)}$ is 0.8, the performance value of the algorithm reaches the maximum. Therefore, the initial step is 0.8.

The selection of the number of initial class centers also has a certain influence on the final experimental results. The number of cluster centers is adjusted, and the number of final class centers is obtained by comparing FCM and method of this article. The results are shown in Figure 4.

It can be seen from the figure that when the number of cluster cores is close to the number of topics, the *F* value of the algorithm is larger, but if the number of clusters is smaller than the number of topics, the clustering effect is very poor, and with the number of clusters increases, the effect is basically showing a gradual upward trend. However, when it is larger than the number of topics, the clustering effect does not change basically, and it has the same effect for the FCM algorithm. Therefore, the number of initial cluster centers of data sets M and N is 6, and the *F* value of the algorithm is about 10% higher than that of FCM algorithm.

2) Similarity threshold determination

Since the data is obtained from micro-blog, micro-blog has the characteristics of short text, timeliness and randomness, so the similarity of the same topic is low. If the similarity value which selected is larger, the similar topics will be few,

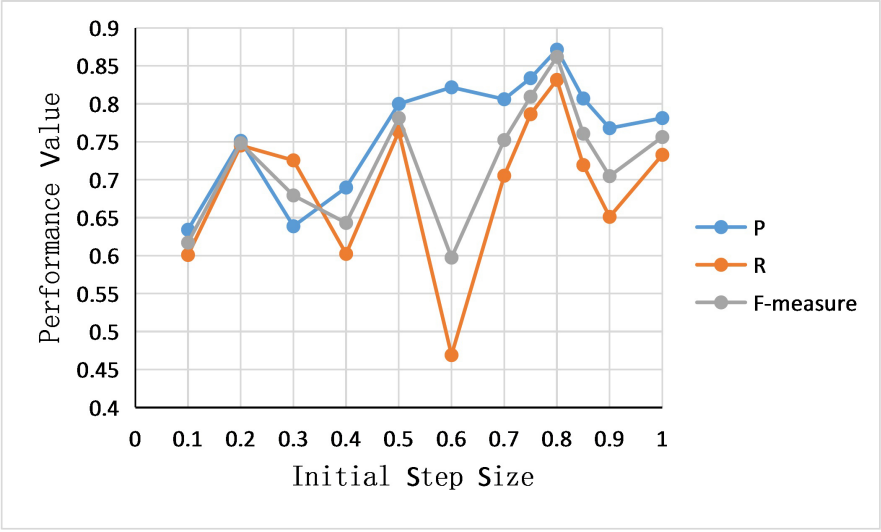

Figure 3. The initial value is selected differently for each indicator value.

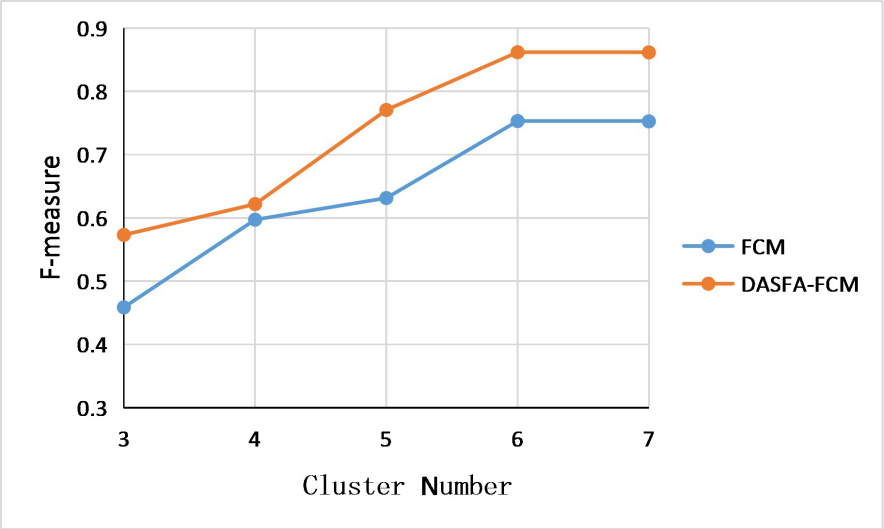

Figure 4. Different values of cluster hearts correspond to *F* values.

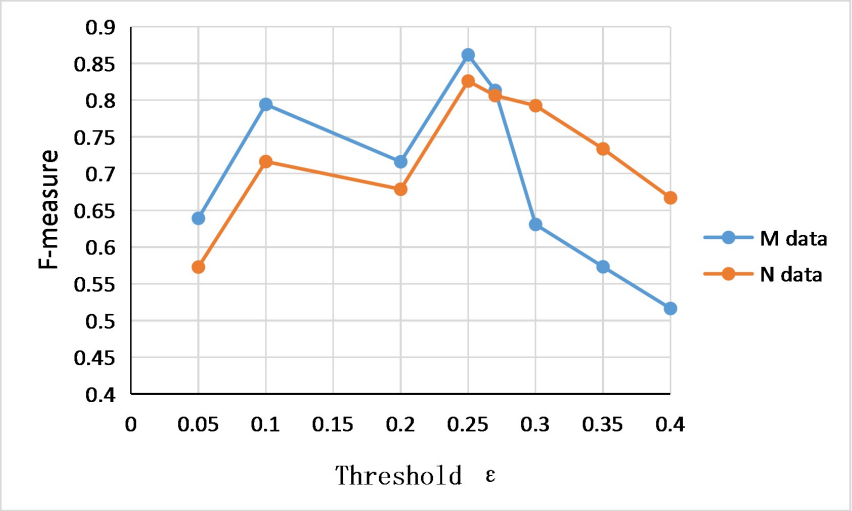

Figure 5. Threshold differences correspond to *F* values.

which makes the accuracy of clustering is reduced. It is assumed that as long as the topics have certain similarities, they will attract each other. Based on this, the experimental results are shown in Figure 5.

From the graph, we can see that the similarity threshold presents a fluctuating situation. When $\varepsilon = 0.1$, it seems that the value of F reaches the highest, but as the threshold increases, the value of F has a decreasing trend. We take the value later, when $\varepsilon = 0.25$, we can see that F has reached an optimal state. Then, with the increase of threshold, the value of F keeps decreasing. In addition, there are obvious differences between the F values of the two data sets. The measure value of the data set M is higher than the data set N in a certain threshold, but when $\varepsilon > 0.27$, the measure value of the data set M is significantly drops and lower than the data set N. This may be due to the large amount of data set M and the relationship with the characteristics of the topic itself. There are more topics of

variation with the development of time which we crawled. With the increase of threshold, if the similarity threshold is higher, and the possibility of the topic to be accurately classified is getting smaller and smaller, it leads to a lower value of *F*. In a comprehensive view, the most suitable ε is 0.25.

3) Performance comparison

In order to verify the effectiveness of the proposed algorithm, the proposed algorithm is compared with traditional FA, traditional FCM algorithm and PSO + FCM algorithm [5]. The effectiveness of the proposed algorithm is verified by comparing the clustering effects of the three algorithms. In order to observe the effect of each topic clustering more clearly, the six topics in the data set M are sequentially listed to compare the performance of different topics. The data set N is the average comprehensive performance value comparison of the six topics.

According to the comparison of the experimental results of the three algorithms, the analysis shows that the proposed algorithm has the best performance. The results of the experiment are shown in Table 1. Firstly, in terms of topic accuracy, the accuracy of the algorithm is high, although it is not ideal for some topics. For example, the accuracy of the algorithm in "topic 4" is lower than that of the FCM algorithm. This is due to the characteristics of the topic itself, but the comprehensive assessment is optimal. This shows that for the FA algorithm, the improvement of the algorithm step makes the search accuracy of the topic discovery improved. For the FCM algorithm, the proposed algorithm uses the FA to obtain the accurate initial class center and then uses the fuzzy

Table 1. Performance comparison.

		M						N
		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
P	FA	0.8172	0.8326	0.8105	0.7314	0.8613	0.835	0.7743
	FCM	0.7876	0.7782	0.6491	0.8805	0.7799	0.7355	0.6886
	Proposed algorithm	**0.8514**	**0.9450**	**0.8644**	**0.8712**	**0.9654**	**0.891**	**0.8508**
	PSO + FCM	0.8394	0.8701	0.7922	0.8163	0.7962	0.7383	0.7832
R	FA	0.7485	0.6588	0.8000	0.6933	0.8344	0.7913	0.631
	FCM	0.728	0.7019	0.6833	0.8589	0.7878	0.7008	0.6668
	Proposed algorithm	**0.7852**	**0.7412**	**0.850**	**0.8712**	**0.9289**	**0.8129**	**0.8038**
	PSO + FCM	0.6935	0.7931	0.8139	0.7756	0.8102	0.7397	0.7154
F	FA	0.7813	0.7356	0.8052	0.7118	0.8477	0.8126	0.6953
	FCM	0.7566	0.7381	0.6658	0.8563	0.7838	0.7177	0.6774
	Proposed algorithm	**0.817**	**0.8308**	**0.8571**	**0.8697**	**0.9468**	**0.8498**	**0.8261**
	PSO + FCM	0.7595	0.8298	0.8029	0.7954	0.8031	0.739	0.7478

clustering to overcome the shortcomings of FCM algorithm that sensitive to the initial value. PSO can also overcome the shortcomings of FCM, but the performance value is still slightly lower than DASFA-FCM. Dividing the topic into the class closest to the cluster center and using the combination of topic influence value and distance reduces the outliers in the space, so that some topics which has low influence but sudden characteristics can be accurately clustered and integrated, it improved the accuracy of the algorithm. In terms of *R* and *F*-measure value, although the performance values of several algorithms are different for different topics, but the algorithm performance values are lower than DASFA-FCM.

In the data set N, the performance of the proposed algorithm is also optimal. Especially for the FCM algorithm, the proposed algorithm is significantly higher than the FCM algorithm in all three indicators. From the results of data sets M and N, it can be seen that the algorithm of text crawling at different time periods is the best, which shows the universal applicability of the proposed algorithm to topic discovery.

4) Convergence performance analysis of fitness value

The experimental research also aims to improve the convergence of the algorithm, taking the data set M as an example, the optimal curve under different algorithms are shown in Figure 6.

Through the comparison of the results of the four algorithms, the FCM algorithm is the most easiest to reach the convergence state with the increase of the number of iterations, but its convergence accuracy is lower and the effect is poorer; FA convergence speed is slower, but because it is high-dimensional data, the convergence speed of PSO + FCM is the slowest. The algorithm achieves the optimality when iterating 82 times, while PSO + FCM achieves the best when iteratively 130 times, and the convergence speed is obviously slower. The convergence accuracy is also the highest for the algorithm of this paper compared with other algorithms, and the search result is higher than other algorithms.

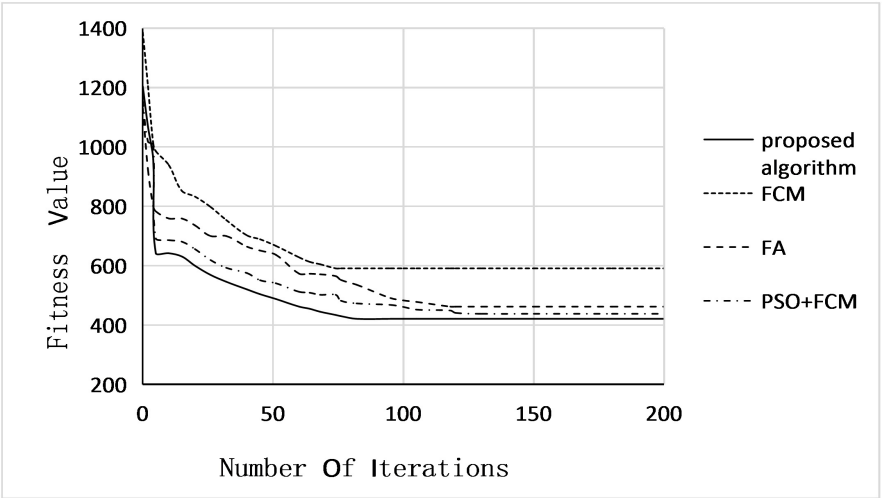

Figure 6. Comparison of fitness values.

5. Conclusion

Due to the diversity of Chinese semantic topics in the process of communication, it is difficult to deal with the challenges of heterography, polysemous words and new words. It is easy to produce mutations, data sparsity and other issues, which lead to the problem of accuracy in topic discovery. Based on this, fuzzy clustering is used. A topic can belong to multiple classifications for clustering hot topics, but the cluster has the disadvantages of relying on initial values and is sensitive to outliers. Based on this, a dynamic adaptive step size firefly algorithm is proposed to improve the fuzzy clustering algorithm. The text is classified by similarity. The topic forwarding number, the number of comments, and the number of praises are used as the influence factor to improve the fitness function to reduce the outliers in the space. The algorithm improves the shortcomings of the traditional clustering algorithm, divides topics into classes closest to the cluster heart, and reduces the impact of topic variation. After the improvement, the accuracy of hot topic clustering is improved, but the noise points are not processed. The next step will consider the influence of noise points on topic clustering and the improvement of text feature extraction.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Zubiaga, A., Spina, D. and Martínez, R. (2015) Real-Time Classification of Twitter Trends. *Journal of the Association for Information Science & Technology*, **66**, 462-473. https://doi.org/10.1002/asi.23186

[2] Cao, J.X., Xu, S. and Chen, G.J. (2017) Regional Topic Discovery in Online Social Networks. *Computer Journal*, **40**, 1530-1542.

[3] Li, X., Wang, L.X. and Lu, M.X. (2016) Network Hot Topic Recognition and Description Algorithm Based on Compound Words. *Library and Information Work*, **23**, 128-134.

[4] Feng, L.G. and Liu, Q.C. (2015) Micro-Blog Hotspot Based on FCM Parallel Algorithm Finds. *Computer Application and Software*, **32**, 232-237.

[5] Filho, T.M.S., Pimentel, B.A. and Souza, R.M.C.R. (2015) Hybrid Methods for Fuzzy Clustering Based on Fuzzy C-Means and Improved Particle Swarm Optimization. *Expert Systems with Applications*, **42**, 6315-6328. https://doi.org/10.1016/j.eswa.2015.04.032

[6] Wu, Z. and Zhang, J. (2017) An Improved FCM Algorithm with Adaptive Weights Based on SA-PSO. *Neural Computing & Applications*, **28**, 3113-3118.

[7] Jiang, H., Gu, J. and Liu, Y. (2013) Study of Clustering Algorithm Based on Fuzzy C-Means and Immunological Partheno Genetic. *Journal of Software*, **8**, 134-141. https://doi.org/10.4304/jsw.8.1.134-141

[8] Yang, F. and Huang, B.X. (2013) Application of Genetic Clustering of Word Co-Occurrence Network in Topic Discovery. *Computer Engineering and Application*, **49**, 126-129.

[9] Amiri, B., Hossain, L., Crawford, J.W. and Wigand, R.T. (2013) Community Detection in Complex Networks: Multi-Objective Enhanced Firefly Algorithm. *Knowledge-Based Systems*, **46**, 1-11. https://doi.org/10.1016/j.knosys.2013.01.004

[10] Yang, X.S. (2008) Nature-Inspired Metaheuristic Algorithms. Luniver Press, 1-147.

[11] Wang, C. and Lei, X.J. (2014) New Niching Firefly Partitioning Clustering Algorithm. *Computer Engineering*, **40**, 173-177.

[12] Bezdeck, J.C., Ehrlich, R. and Full, W. (1984) FCM: Fuzzy C-Means Algorithm. *Computers & Geoscience*, **36**, 691-698.

[13] Wang, H., Cui, Z. and Sun, H. (2017) Randomly Attracted Firefly Algorithm with Neighbor Hood Search and Dynamic Parameter Adjustment Mechanism. *Soft Computing*, **21**, 1-15.

[14] Yu, S., Zhu, S. and Ma, Y. (2015) A Variable Step Size Firefly Algorithm for Numerical Optimization. *Applied Mathematics & Computation*, **263**, 214-220. https://doi.org/10.1016/j.amc.2015.04.065

[15] Qiu, J.N. and Gu, W.J.Z. (2017) Research on Hot Topic Detection Method Based on User Influence. *Information Magazine*, **36**, 156-161.

[16] Senthilnath, J., Omkar, S.N. and Mani, V. (2011) Clustering Using Firefly Algorithm: Performance Study. *Swarm & Evolutionary Computation*, **1**, 164-171. https://doi.org/10.1016/j.swevo.2011.06.003