IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2019, accepted July 1, 2019, date of publication July 3, 2019, date of current version August 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926816

GPOGC: Gaussian Pigeon-Oriented Graph Clustering Algorithm for Social Networks Cluster

YANG SUN[1], SHOULIN YIN[1,2], HANG LI[1], LIN TENG[1], AND SHAHID KARIM[2]
[1]Software College, Shenyang Normal University, Shenyang 110034, China
[2]School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150000, China

Corresponding authors: Shoulin Yin (yslinhit@163.com) and Hang Li (lihangsoft@163.com)

This work was supported by the Natural Science Fund Project Guidance Plan in Liaoning Province of China under Grant 20180520024.

ABSTRACT As the continuous development of mobile social networks, the structure of the mobile social network increasingly becomes complex. It not only speeds up information transmission between people but also expands the scope of information exchange, which has become an essential and important social media in people's social life. How to effectively identify and classify these online communities has important practical significance for the study of social networks. Correctly detecting the community structure of mobile social networks can not only improve the accuracy of friend recommendation, link prediction, service user positioning, product marketing, and other aspects but also provide an important basis for the monitoring of online public opinion. But the traditional social network cluster method based on the trust degree mainly calculates the user trust by analyzing the interactive feedback information between users. This method cannot effectively solve the ''cold start'' problem in the trust calculation process, that is, for the new network node, the trust value of this node cannot be accurately measured due to the lack of interaction with other nodes. Focusing on this problem, we propose a Gaussian pigeon-oriented graph clustering algorithm for social networks' cluster in this paper. A graph model is first built. Then, an efficient K-medoid algorithm is utilized to search the user center in all groups. The Gaussian pigeon algorithm is used to search the similarity between each user and the central user. Users that meet the similarity threshold are divided into the same user group. Finally, the simulation results show that the proposed method has better cluster effect than other state-of-the-art social networks' clustering approaches.

INDEX TERMS Social network, Gaussian pigeon algorithm, graph clustering, K-medoid algorithm.

I. INTRODUCTION

In social networks, users are not only information acquirers, but also information publishers and transmitters. The emergence of this information transmission mechanism not only greatly reduces the social cost among network users, but enables users to establish social relations with some people with common characteristics through online activities, which forms a network structure similar to that of real social communities in mobile social networks [1], [2].

With the popularity of various applications such as weibo, WeChat, douban movie and netease cloud music, the social network in different fields develops rapidly. At present, there are a variety of social relationships in social networks, such as friend relationship, concern relationship, having the same preferences, etc,. Nodes and connections of social networks also have diverse attributes. Traditional network clustering methods mainly consider the density of links, but not the diversity of social networks. In addition, the presence of low-activity users in social networks also brings adverse effects on the clustering effect of social networks [3], [4].

In addition to the social network clustering algorithm based on the density of links, there are also clustering algorithms that take node diversity, strong and weak social relations and various hidden information into consideration. Hänninen and Kujala [5] mainly considered the similarity of users' interests, and calculated the similarity of users' interests based on the Bayesian probability model. In the current diversified social networks, interest had become a weakly related information. Additionally, trust communication, comment information, rating information and so on should also be considered. Zhang *et al.* [6] proposed

The associate editor coordinating the review of this manuscript and approving it for publication was Basit Shahzad.

a directed network clustering algorithm based on structural similarity. For directed interactivity of social networks, the algorithm considered the arrival neighbor of nodes and used directed edges to define direct structural reachable. Cai *et al.* [7] adopted particle swarm optimization algorithm to optimize the social network, took the network structure as the objective function of particle swarm, and guided the evolution process of particle swarm through greedy strategy. References [8], [9] all took social network structure as the basis of clustering, but only direct social relationship was considered in the process of network construction.

In the current social networks, there are various correlations. In addition to strong relationships, all kinds of weak relationships should also be considered, including attention relationships, trust communication, comments, information [10], [11], rating information, etc,. In addition, there are active users and low active users in social networks. Low active users will lead to the problem of sparsity, thus affecting the accuracy and coverage rate of clustering. In order to solve the above problems, GPOGC: Gaussian Pigeon-oriented graph clustering algorithm for social networks clustering is proposed. In order to ensure the balance between coverage rate and clustering accuracy, a two-dimensional (2D) graph is established under the constraint of coverage rate. In the process of constructing the graph, the direct trust relationship, trust propagation, comment information and other diversified information are considered. In order to solve the sparsity problem, the prediction mechanism of low-activity users is designed by combining Pearson similarity and diversified social relations. In the clustering stage, Gaussian Pigeon algorithm is used to search the users with the highest similarity with the central user, so as to improve the accuracy of clustering. This paper is organized as follows. In section 2, we introduce the related works about social networks. A two-dimensional graph model is constructed for our proposed cluster method in section 3. Gaussian Pigeon optimized algorithm and proposed social networks cluster method are detailed illustrated in section 4 and section 5, respectively. Experiments are conducted in section 6. A conclusion is given in section 7.

II. RELATED WORKS

At present, there are many methods for the division of social network communities, which can be divided into two categories based on the relationship between users and users' preferences. Classification method based on the relationship between the users contains graph segmentation method, the module degrees optimization algorithm, G-N algorithm, CPM algorithm and label propagation method. The main idea of these methods are to treat individuals as undifferentiated network nodes in a complex network graph. According to the topology of the graph, the social network is divided into user-centered communities, which has the characteristics of tight internal connections and sparse external connections.

Xia and Cao [12] presented a spectrum split method by building a Laplace matrix. Because in the process of dividing the network, it needed many repeatedly operations. The method is more complicated. Liu *et al.* [13] put forward a new spectrum division method. But this method would be calculated in the process of implementation standard eigenvalue of the matrix. A higher cost was needed for larger network size. The algorithm not only had certain limitation in the implementation process, it was also inadequate. Just considering network topology dividing network formation, the strength of the relationship between users were not reflected. It ignored the similarity of node contents in the social network. Due to the link noise in mobile social networks, if only the link relations between nodes are considered and the similarity between nodes are ignored, it is completely unreasonable to divide nodes with weak link relations or no link relations but obviously similar in content into different communities, Therefore, in the process of mobile social network partitioning, not only the link relationship between nodes should be considered, but the similarity of content is same.

Koyama *et al.* [14] established a new preference prediction model by analyzing user dialogues and online interactions, the experiment confirmed that the model could improve the accuracy of interest mining. Zhou *et al.* [15] used topic model to predict the similarity of users to publish content, and the experiment proved that the model could more accurately infer the contents of the similarity between users. Muchnik *et al.* [16] set the users as the receiver and the sender to construct receiver and sender model. The subject probability distribution was used to divide the community, and finally groups with the same social roles were obtained. The above methods only consider the community characteristics to mine users that are consistent with the community theme characteristics. Although the user preferences within the same community are similar, there are not necessarily close connections between users.

III. TWO-DIMENSIONAL GRAPH MODEL

The sparsity problem of social networks can be effectively alleviated by building two-dimensional graph under coverage constraints. The effect of similarity measurement highly depends on the user's comment information. So improving the reliability of similarity measurement can improve the reliability and accuracy of clustering [17].

Trust-aware social networks improve the accuracy of clustering by predicting the information of low active users. The basic idea is to assume that users are easily influenced by users with high trust, but this mechanism can easily lead to reduced coverage. Many researchers have found that users are not only influenced by the direct trust users, also by the indirect users. However, their influence decreases with the increase of the distance between two users. This theory is also called trust propagation.

A graph model based on trust and similarity is designed. The process of model is shown in algorithm 1. The nodes of the graph represent users. Edges represent double-weighted connections between users. It is expressed as tuple $(W1, W2) = (pcc(u, v), T(u, v))$, where pcc denotes

similarity measurement and T represents trust propagation. The inputs of the algorithm are the direct trust information, the indirect trust information, the Pearson correlation coefficient (*pcc*) and the maximum distance of trust propagation (MP). The output is the graph of the social networks. The social graph is represented by adjacency matrix. The *setdiff* () function cancels an existing new connection. It calculates the trust propagation based on the shortest path between users. The coefficient $\frac{1}{i}$ in line 6 indicates that the longer the distance between two users is, the lower the trust value is.

Algorithm 1 Building Graph in Social Network Algorithm

Input *pcc*, trust graph, *MP*.
Output W_{graph}.
1.Setting N_{user} as the number of users.
2.Initializing the temporary user matrix *tmp*.
3.Initializing the user matrix *mt*.
4.for each $i = 1$ to *MP* do
5.$tmp = tmp \times T$
6.Calculating the differences between trusted users $mt = mt + (1/i)setdiff\,(mt, tmp)$
7.for each (u_i, u_j) do
8.if $pcc(u_i, u_j)$ and $MT(u_i, u_j)$ are existing.
9.W_s, $W_{mt} = (pcc(u_i, u_j), \theta)$
else if
set $h(t) = r(t)$

IV. GAUSSIAN PIGEON OPTIMIZED ALGORITHM

Due to the complexity of data in the actual clustering analysis, it is difficult to obtain the appropriate clustering results by manual calculation. Therefore, this paper adopts Gaussian Pigeon optimized algorithm to solve the clustering results. Compared with the traditional Pigeon-inspired optimized algorithm (PIO) [18,19], the Gaussian Pigeon algorithm improves the convergence speed of the algorithm and the clustering results due to the introduction of gaussian term.

In the nature, when a pigeon returns to its nest, it will perceive the magnetic field through a magnetic object when it is far away from its destination. The height of the sun is as a compass to constantly adjust its position and speed. When it is close to the destination, familiar locations near the pigeon nest will be selected for navigation. Therefore, the whole PIO algorithm is divided into two stages: 1) the geomagnetic and sun-based map and compass stage, and 2) the landmark stage.

A. MAP AND COMPASS STAGE

Assuming the number of pigeons is N and the dimension is D, the maximum iteration number is T_1. The velocity v and position x of the pigeon are initialized. Then rules are made to adjust the velocity and position of pigeons. In the PIO algorithm, the pigeon can be guided to search for the optimal solution according to geomagnetism and the sun. The specific rules are shown in equations (1) and (2).

$$V_i(T) = V_i(T-1)e^{-R\times T} + rand \cdot [X_{best} - X_i(T-1)]. \quad (1)$$

$$X_i(T) = X_i(T-1) + V_i(T). \quad (2)$$

where T represents the number of current iterations. R stands for map and compass factor. *rand* is a random number between 0 and 1. X_{best} represents the global optimal position obtained through comparison in the $t-1$ iteration process. $V_i(T)$ and $V_i(t-1)$ represent the velocity of the $T(0 < T \leq T_1)$ iteration and the $T-1$ iteration, respectively. $X_i(T)$ and $X_i(t-1)$ represent the position of the T iteration and the $T-1$ iteration, respectively.

B. LANDMARK STAGE

In this stage, the pigeon will select landmarks near the pigeon nest to adjust the position, determine the maximum iteration number T_2 and fitness function $F(X)$ to update the pigeon position, as shown in equation (3)-(5).

$$X_c(T-1) = \frac{\sum_{i=1}^{N(T-1)} X_i(T-1)F[X_i(T-1)]}{\sum_{i=1}^{N(T-1)} F[X_i(T-1)]}. \quad (3)$$

$$X_i(T) = X_i(T-1) + r \cdot [X_c(T-1) - X_i(T-1)]. \quad (4)$$

$$N(T) = \frac{N(T-1)}{2}. \quad (5)$$

where $N(t-1)$ represents the number of pigeons in $T-1(0 < T \leq T_2)$ iteration. $X_c(T-1)$ denotes the center position of the remaining pigeons; r represents a random number between 0 and 1 that satisfies uniform distribution. $F(X)$ represents the fitness function, which is determined according to the specific problem.

According to the above equations, after each iteration, the number of pigeons is reduced by half. That is, those pigeons away from the landmark are no longer instructive and have to fly with other pigeons closer to the destination.

C. GAUSSIAN PIGEON OPTIMIZED ALGORITHM

PIO algorithm has a fast convergence speed, but it is still easy to fall into the local optimal. And there is a problem to balance the two stages. In order to improve the efficiency of the algorithm, the gaussian term is introduced into the landmark stage.

Landmark parameter r has good global search ability. In most cases, when the destination has been identified, optimization algorithm should have good focus search ability. But it only satisfies uniform distribution rule, which cannot meet this requirement. So in order to improve the global and local search abilities, in the landmark stage, it appropriately changes the distribution of r to find the global optimal solution.

$$\begin{cases} r = (R_1 - 0.5) \cdot m\sqrt{n}, & R_2 > p \\ r = 2(R_1 - 0.5) \cdot \sqrt{n}, & R_2 \leq p \end{cases} \quad (6)$$

where p is a parameter that balances uniform distribution and gaussian distribution. R_1 and R_2 are two random numbers between 0 and 1.

$$\begin{cases} m = R_n \\ n = 1 - 0.5\dfrac{T}{T_2} \end{cases} \quad (7)$$

where R_n is a random number satisfying the gaussian distribution between 0 and 1. T_2 is the maximum number of iterations in the landmark stage.

As can be seen from equation (6) and (7), the Gaussian pigeon group optimizes the parameter r on the basis of the pigeon group. m and n are constantly changing, while the parameter p is selected according to the optimization objective.

V. GPOGC: GAUSSIAN PIGEON-ORIENTED GRAPH CLUSTERING ALGORITHM

The coverage of social network has no correlation with the number of groups. Therefore, the number of the first group searched can be regarded as the number of groups in the clustering algorithm.

K-medoids algorithm [20] is adopted to search the central users of user groups. The objective function F of the K-medoids algorithm is defined as:

$$F = min \sum_{c_E \in C} \sum_{m,n \in C_E} dist(m, n). \tag{8}$$

where C is the set of classes. $dist(m, n)$ denotes the distance between user m and n in the two-dimensional graph. Because each edge in the graph is double-weighted, the distance between users is calculated as:

$$dist^2(u, v) = d_S^2(u, v) + d_T^2(u, v). \tag{9}$$

where u and v are two target users. d_S and d_T are similarity distance and trust distance respectively, calculated by:

$$d_S(u, v) = 1 - W_S^{2D-graph}(u, v). \tag{10}$$

$$d_T(u, v) = 1 - W_{MT}^{2D-graph}(u, v). \tag{11}$$

Then we look for user groups that are highly similar to the central user. The process mainly includes three steps: permutation processing, weight processing and prediction processing.

1) Initializing permutation processing. The goal of this step is to calculate the similarity value between each user and the target user (central user) based on the trust information and comment information, and extract the top-n similar users. If there is a direct trust relationship between users, such as friend relationship, concern relationship, etc., then the trust value is directly calculated; If there is no direct trust relationship between users, then the hidden trust relationship, such as comment information, rating information and so on, will be extracted. If there is no direct trust relationship between user u and target user a, we use pcc to calculate the trust value of u and a according to the comment information or scoring information. The nodes of the network represent users, and the weight of edges represents the similarity between users. The user similarity calculation based on trust is:

$$\begin{cases} W_{a,u} = \dfrac{2 \cdot sim(a, u) \cdot T_{a,u}}{sim(a, u) + T_{a,u}}, sim(a, u) + T_{a,u} \neq 0. \\ W_{a,u} = T_{a,u}, sim(a, u) = 0, T_{a,u} \neq 0. \\ W_{a,u} = sima, u, sim(a, u) \neq 0, T_{a,u} = 0. \end{cases} \tag{12}$$

where $T_{a,u}$ is the trust value between target user a and user u, the formula is:

$$T_{a,u} = \frac{d_{max} - d_{a,u} + 1}{d_{max}}. \tag{13}$$

In here, $d_{a,u}$ represents the trust propagation distance between a and u. d_{max} is the maximum trust propagation distance, and $dmax$ is set as the average path length in the graph.

$$d_{max} = \frac{\ln(n)}{\ln(k)}. \tag{14}$$

where n is the number of users in the network. k is the average degree of the network. Suppose $sim(a, u)$ represents the similarity between a and u, the similarity based on pcc is calculated as:

$$sim(a, u) = \frac{\sum_{i \in A_{a,u}} (r_i(a) - \bar{r}(a))(r_i(u) - \bar{r}(u))}{\sqrt{\sum_{i \in A_{a,u}} (r_i(a) - \bar{r}(a))} \sqrt{\sum_{i \in A_{a,u}} (r_i(u) - \bar{r}(u))}}. \tag{15}$$

where $r_i(u)$ is the score value of user u for project i. $\bar{r}(u)$ is the average score value of user u. $A_{a,u}$ is the set of items rated by user a and u. Eventually, the user is set as one group whose similarity is higher than threshold value θ.

2) Weight processing for two-dimensional graph model. We adopt gaussian PIO to process top-n users and analyze their importance. First of all, we establish the two-dimensional graph of the user. Then, the gaussian PIO in the graph to adjust the similarity between each user and target user.

First, we select top-n users similar to the target user. Then, a two-dimensional graph is established for the social network, where nodes represent users. Edges and weights represent the similarity between users (as formula (12)), and the range value of weight is [0,1].

The purpose of determining the fitness function of gaussian PIO is to judge the clustering result, because the clustering result requires that all users in the cluster have similar scores. Therefore, in this paper, the distance between users and the clustering center is taken as the fitness function to evaluate the clustering results. The selection way of clustering center is shown in equation (16), and the fitness function is shown in equation (17).

$$S_k(i) = \frac{\sum_{i \in k} d_2(i, j)}{M}. \tag{16}$$

where k represents cluster number. i denotes the number of the selected element. j represents the number of elements except i in cluster k. M represents the number of all elements in cluster k. $S_k(i)$ represents the mean value of element i relative to cluster k. The one with the smallest mean is selected as the clustering center of the cluster.

$$F = \frac{1}{\sum d_2(x_i, x_{kc})}. \tag{17}$$

Here, $\sum d_2(x_i, x_{kc})$ denotes the distance between user and cluster center. x_{kc} is the $k - th$ cluster center. The cluster

is better when F is bigger. Formula (17) can prevent the algorithm from falling into local optimization. In the social network, this function can select the set of users with similar interests and low redundancy.

3) Prediction processing for low active users. For users without comment information, it predicts their comments based on the most similar user comments. The prediction method is:

$$\hat{r}_{u,j} = \frac{\sum_{v \in U} w_v r_{v,i}}{\sum_{v \in U} w_v}. \tag{18}$$

where $\hat{r}_{u,j}$ is the predicted comment of target user u for project i. U is the set of users selected by PIO. $r_{u,j}$ represents the true score of v for project i. w_v is the position of v. The cost of each solution is calculated as the error between the predicted value and the true value.

$$f(u) = \frac{\sum_{i=1}^{I_u} |\hat{r}_{u,i} - r_{u,i}|}{|I_u|}. \tag{19}$$

I_u is the predicted project number.

The goal of this processing is to predict the information of low active users based on the information of active users. This processing is helpful to alleviate the common ''cold start'' problem and sparsity problem in social networks.

Therefore, we can summarize a final GPOGC algorithm for social networks cluster as follows.

1) Step 1. Calculating the similarity between the target user and other users. Selecting users with higher similarity than θ and inputting them into gaussian PIO algorithm.
2) Step 2. Using gaussian PIO to assign weights to users. In each iteration of gaussian PIO, pigeons traverse in the graph and select a set of similar users.
3) Step 3. Updating the position and speed of pigeons. In gaussian PIO algorithm, the position of each pigeon represents a clustering result. Pigeons will constantly change their position and speed to seek the global optimal solution.
4) Step 4. The pigeon's position and speed update consist of two stages, namely the map, compass stage and the landmark stage. In the map and compass stage, the fitness value of each pigeon's experienced position, individual optimal position and group optimal position are compared to obtain the global optimal position X_{best}. Then update position and speed.
5) Step 5. Repeat step 4 until the number of iteration T is greater than the maximum number of iteration t_1. Then it enters the landmarks stage. After each iteration, the pigeons will be cut in half. Those pigeons who no longer have the ability to distinguish the path away from the goal of pigeons must be abandoned. When the conditions satisfy the iteration (the maximum iteration number is T_2), it outputs the best position of pigeons, namely the optimal solution.

VI. EXPERIMENTS AND ANALYSIS

A. PERFORMANCE ANALYSIS OF THE GAUSSIAN PIO

First, to verify the performance of the gaussian pigeon-inspired algorithm, Rosenbrock function and Rastrigin function are selected as test functions in this paper. Rosenbrock function belongs to an unimodal function, and the minimum value is 0. Rastrigin function belongs to a multimodal function, and the minimum value is 0. The formulas are shown in (20) and (21), respectively.

$$f_1(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]. \tag{20}$$

$$f_2(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + +10]. \tag{21}$$

PIO algorithm, ABP [21] and gaussian PIO algorithm are selected as comparison objects. The population number is set as 10, the maximum iteration times is 150, and the dimension is set as 100. The relationships between fitness value and iterations of the three algorithms are shown in figure 1 and figure 2. The optimal values are displayed in table 1 and table 2.

As can be seen from table 1, in terms of single-peak function, since the function has only one extreme value, it is relatively easy to solve. The three algorithms have a big gap,

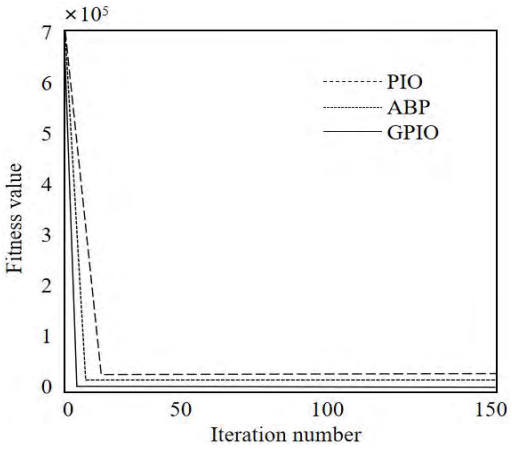

FIGURE 1. The convergence curve of Rosenbrock function.

TABLE 1. Comparison of Rosenbrock function results.

Algorithm	Value
PIO	110.57
ABP	99.65
GPIO	89.32

TABLE 2. Comparison of Rastrigin function results.

Algorithm	Value
PIO	177.61
ABP	52.46
GPIO	41.23

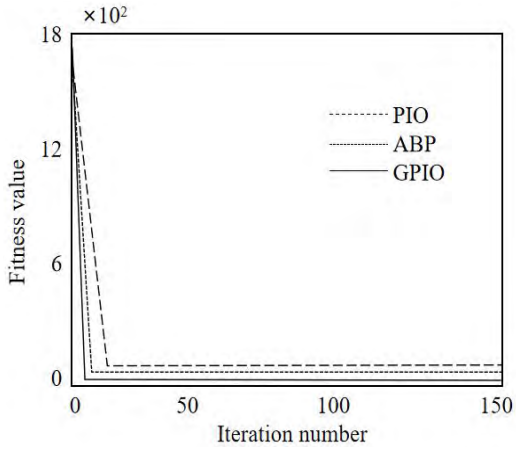

FIGURE 2. The convergence curve of Rastrigin function.

GPIO algorithm has obvious advantages. In terms of bimodal function, the optimal value of GPIO algorithm is improved to some extent compared with the ABP. When solving the optimal value of multi-peak function, GPIO algorithm has great advantages. Figures1-2 also illustrate this advantage of GPIO.

B. PERFORMANCE EVALUATION INDEX

Recommendation system is an important application scene of social network clustering technology. The experimental scheme is adopted to combine clustering technology with collaborative filtering recommendation system. The effect of social network recommendation technology is evaluated through the effect of recommendation system. Three data sets are used to test the clustering performance of GPOGC algorithm. The experimental environment is Windows 10 with 16GB memory and Core i7. Each data set is divided into five subsets by the half-fold cross test scheme. In each iteration, four subsets are randomly selected as the training set and the other one as the test set.

Three classical recommendation system performance indexes including Mean Square Error (MAE), Root mean square error (RMSE) and fraction of coverage (FC) are used to evaluate the accuracy of prediction. MAE calculates the difference between the predicted rating and the real rating.

$$MAE = \frac{1}{Z}\sum_{(u,j)} |\hat{r}_{uj} - r_{uj}|. \tag{22}$$

where Z, $\hat{r}_{uj}$ and r_{uj} are number of scores, the estimated number of scores, and the actual number of scores, respectively. RMSE is also an indicator to evaluate the performance of the recommendation system, which measures the absolute error between the predicted score and the real score.

$$RMSE = \sqrt{\frac{1}{Z}\sum_{(u,j)} (\hat{r}_{uj} - r_{uj})^2}. \tag{23}$$

FC evaluates the performance of the recommendation system from another perspective, and evaluates the ability of the recommendation system to mine long-tail commodities. The calculation method of FC is:

$$FC = \frac{ES}{WS}. \tag{24}$$

ES and WS are the estimated number of scores and the whole number of scores respectively.

C. EXPERIMENTAL DATA SET

The FilmTrust, Epinions and Ciao data set are used as benchmarks data set [22]. FilmTrust is a real data set of movie recommendation sites where users review and grade films. They can add friends and share opinions. The score of FilmTrust data set is real number, and the range is [0.5,4]. The Epinions data set includes a variety of social relations including the comment on the project, the score and the trust relationship between users. The score is integer from [1,5]. Trust relationships have two values: "1" (stands for trust) and "0" (for distrust). The score of Ciao data set is an integer ranging from 1 to 5. Information about the three benchmark data sets is shown in table 3.

TABLE 3. Information of the three benchmark data sets.

Feature	FilmTrust	Epinions	Ciao
User	1508	40163	30444
Project	2071	139738	72665
Score	35497	664824	1625480
Trustor	609	33960	6792
Trusted	732	49288	7297
Trust number	1853	487183	111781

In order to test the effect of this proposed algorithm on the sparsity problem and the 'cold start' problem, the data sets are further divided according to two conditions. The division conditions are: 1) 'cold start' users. Extracting user set with score number less than five. 2) sparse project. Extracting project with score number less than five. 3) all user sets. Table 4 shows the relevant information of the sub-data set.

D. COMPARISON RESULTS

We make comparison on 'cold start' data set, sparse data set and the whole data set with other three state-of-the-art cluster algorithms including EbD [23], MCC [24], DPC [25]. The results of MAE and RMSE of each group are analyzed. Tables 5-7 show the experimental results of FilmTrust, Epinions and Ciao data sets, respectively. GPOGC algorithm has a

TABLE 4. Information of the division sub-data sets.

Condition	dataset	sample number	score number
'cold start'	FilmTrust	300	600
'cold start'	Epinions	17000	34000
'cold start'	Ciao	12000	21000
sparse	FilmTrust	1700	3200
sparse	Epinions	116000	176000
sparse	Ciao	9500	25000

TABLE 5. Results for FilmTrust.

Data	Index	EbD	MCC	DPC	GPOGC
'cold start'	MAE	0.712	0.708	0.693	0.558
'cold start'	RMSE	0.837	0.826	0.796	0.732
sparse	MAE	0.834	0.827	0.813	0.782
sparse	RMSE	1.135	1.067	0.954	0.921
complete set	MAE	0.706	0.689	0.573	0.481
complete set	RMSE	0.772	0.763	0.694	0.631

TABLE 6. Results for Epinions.

Data	Index	EbD	MCC	DPC	GPOGC
'cold start'	MAE	0.836	0.821	0.794	0.722
'cold start'	RMSE	1.234	1.119	0.983	0.927
sparse	MAE	0.841	0.837	0.818	0.793
sparse	RMSE	1.085	0.961	0.923	0.879
complete set	MAE	0.845	0.826	0.798	0.754
complete set	RMSE	1.138	1.097	0.993	0.901

TABLE 7. Results for Ciao.

Data	Index	EbD	MCC	DPC	GPOGC
'cold start'	MAE	1.093	0.945	0.882	0.658
'cold start'	RMSE	1.031	0.987	0.946	0.893
sparse	MAE	0.557	0.784	0.697	0.576
sparse	RMSE	0.774	0.765	0.698	0.654
complete set	MAE	0.723	0.708	0.637	0.653
complete set	RMSE	0.938	0.721	0.875	0.746

better accuracy rate for FilmTrust and Epinions than the other three methods. The Ciao dataset also achieves good results, but its recommendation accuracy for the complete dataset is slightly lower than that for the EbD, and the recommendation accuracy for the sparse dataset is slightly lower than that of the MCC. In general, GPOGC has achieved good recommendation effect and better mitigation effect for both 'cold start' problem and sparsity problem.

Fraction of coverage (FC) index is an important indicator of recommendation system and social network. FC results of four methods are shown in figure 3. As can be seen from the figure, the four algorithms all achieve high coverage, and the EbD, MCC, DPC all achieve FC above 0.9. While the FC of GPOGC is slightly higher than that of the MCC and DPC algorithms.

E. CONVERGENCE ANALYSIS

In this subsection, we conduct convergence experiments. As can be seen from the figures 4-6, the convergence speed

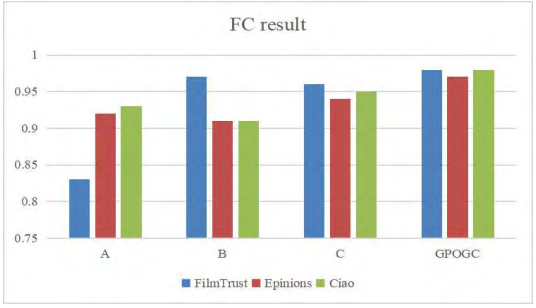

FIGURE 3. FC comparison.

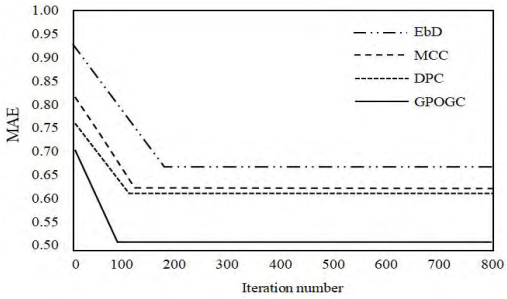

FIGURE 4. Convergence of FilmTrust.

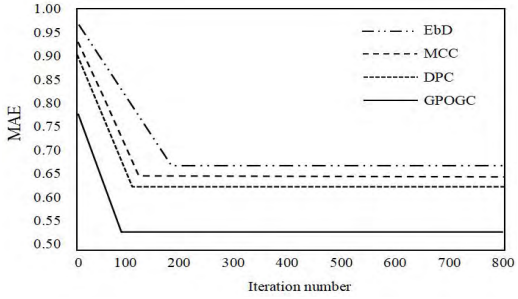

FIGURE 5. Convergence of Epinions.

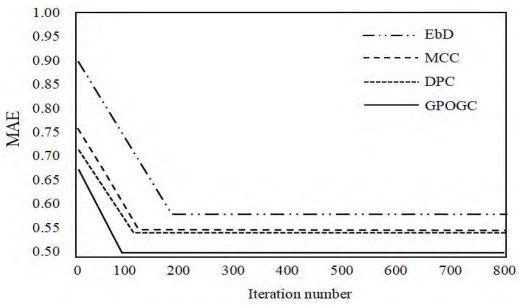

FIGURE 6. Convergence of Ciao.

and accuracy of GPOGC are better than other algorithms. The GPOGC has strong global and local search ability with fast convergence speed. In addition, GPOGC adopts a rich relationship between direct trust relations and indirect trust relations to establish the weights in the graph. This mechanism enables the pigeons to traverse the graph quickly and efficiently at the beginning of iteration. Therefore, GPOGC achieves better exploitation ability and convergence speed.

VII. CONCLUSION

In this paper, we consider the explicit and implicit relationships between users in social networks. This paper proposes a social network clustering algorithm based on graph clustering and gaussian pigeon group algorithm. In order to ensure the balance between coverage rate and clustering accuracy, a two-dimensional graph is established under the constraint of coverage rate. In the process of constructing the graph, the direct trust relationship, trust propagation, comment information

and other diversified information are considered. The proposed algorithm achieves better recommendation effect. And it can alleviate both the 'cold start' problem and the sparsity problem. This new algorithm utilizes the direct and indirect trust relations to establish the weights in the graph. This mechanism enables the pigeons to traverse the graph quickly and efficiently at the beginning of iteration. Therefore, this proposed algorithm achieves better excavation ability and convergence speed. In the future, more hidden social information and external information will be considered to enhance the judgment basis of social network, such as user profile, comment context and behavior track.

REFERENCES

[1] O. Bandiera and I. Rasul, "Social networks and technology adoption in northern Mozambique," *Econ. J.*, vol. 116, no. 514, pp. 869–902, 2010.

[2] C. M. K. Cheung, P.-Y. Chiu, and M. K. O. Lee, "Online social networks: Why do students use Facebook?" *Comput. Hum. Behav.*, vol. 27, no. 4, pp. 1337–1343, 2011.

[3] J. Gao, P. Li, and Z. Chen, "A canonical polyadic deep convolutional computation model for big data feature learning in Internet of Things," *Future Gener. Comput. Syst.*, vol. 99, pp. 508–516, Oct. 2019. doi: 10.1016/j.future.2019.04.048.

[4] T. Liu and S. Yin, "An improved particle swarm optimization algorithm used for BP neural network and multimedia course-ware evaluation," *Multimedia Tools Appl.*, vol. 76, no. 9, pp. 11961–11974, 2017.

[5] M. Hänninen and P. Kujala, "Influences of variables on ship collision probability in a Bayesian belief network model," *Rel. Eng. Syst. Saf.*, vol. 102, pp. 27–40, Jun. 2012.

[6] M. Zhang, Z. He, H. Hu, and W. Wang, "E-rank: A structural-based similarity measure in social networks," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Dec. 2012, pp. 415–422.

[7] Q. Cai, M. Gong, L. Ma, S. Ruan, F. Yuan, and L. Jiao, "Greedy discrete particle swarm optimization for large-scale social network clustering," *Inf. Sci.*, vol. 316, pp. 503–516, Sep. 2015.

[8] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, "Model-based clustering for social networks," *J. Roy. Stat. Soc.*, vol. 170, no. 2, pp. 301–354, 2010.

[9] L. Teng and H. Li, "A high-efficiency discrete logarithm-based multi-proxy blind signature scheme via elliptic curve and bilinear mapping," *Int. J. Netw. Secur.*, vol. 20, no. 6, pp. 1200–1205, Nov. 2018.

[10] W. W. L. Chan and W. W. K. Ma, "Exploring the influence of social ties and perceived privacy on trust in a social media learning community," in *Proc. Int. Conf. Hybrid Learn. Continuing Edu.*, 2013, pp. 134–144.

[11] Y. He, C. Liang, R. Yu, and Z. Han, "Trust-based social networks with computing, caching and communications: A deep reinforcement learning approach," *IEEE Trans. Netw. Sci. Eng.*, to be published.

[12] W. Xia and M. Cao, "Analysis and applications of spectral properties of grounded Laplacian matrices for directed networks," *Automatica*, vol. 80, pp. 10–16, Jun. 2017.

[13] H. Liu, X. Xu, J.-A. Lu, G. Chen, and Z. Zeng, "Optimizing pinning control of complex dynamical networks based on spectral properties of grounded Laplacian matrices," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[14] Y. Koyama, Y. Sawamoto, Y. Hirano, S. Kajita, K. Mase, T. Suzuki, K. Katsuyama, and K. Yamauch, "A multi-modal dialogue analysis method for medical interviews based on design of interaction corpus," *Pers. Ubiquitous Comput.*, vol. 14, no. 8, pp. 767–778, 2010.

[15] Y. Zhou, X. Guan, Z. Zhang, and B. Zhang, "Predicting the tendency of topic discussion on the online social networks using a dynamic probability model," in *Proc. Hypertext Workshop Collaboration Collective Intell.*, 2008, pp. 7–11.

[16] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. Andrade, Jr., S. Havlin, and H. A. Makse, "Origins of power-law degree distribution in the heterogeneity of human activity in social networks," *Sci. Rep.*, vol. 3, no. 19, p. 1783, May 2013.

[17] H.-T. Chang, Y.-W. Li, and N. Mishra, "mCAF: A multi-dimensional clustering algorithm for friends of social network services," *Springerplus*, vol. 5, no. 1, p. 757, 2016.

[18] H. Duan and P. Qiao, "Pigeon-inspired optimization: A new swarm intelligence optimizer for air robot path planning," *Int. J. Intell. Comput. Cybern.*, vol. 7, no. 1, pp. 24–37, Mar. 2014.

[19] S. Zhang and H. Duan, "Gaussian pigeon-inspired optimization approach to orbital spacecraft formation reconfiguration," *Chin. J. Aeronaut.*, vol. 28, no. 1, pp. 200–205, Feb. 2015.

[20] P. Arora, D. Deepali, and S. Varshney "Analysis of k-means and *k*-medoids algorithm for big data," *Procedia Comput. Sci.*, vol. 78, pp. 507–512, 2016.

[21] T. Li, C. Zhou, B. Wang, B. Xiao, and X. Zheng, "A hybrid algorithm based on artificial bee colony and pigeon inspired optimization for 3D protein structure prediction," *J. Bionanosci.*, vol. 12, no. 1, pp. 100–108, 2018.

[22] X. Ye *et al.*, "Clustering algorithm of social networks based on graph clustering and ant colony optimization algorithm," *Appl. Res. Comput.*, vol. 37, no. 6, pp. 1–7, 2019.

[23] X. Qi, H. Song, J. Wu, E. Fuller, R. Luo, and C.-Q. Zhang, "Eb&D: A new clustering approach for signed social networks based on both edge-betweenness centrality and density of subgraphs," *Phys. A, Stat. Mech. Appl.*, vol. 482, pp. 147–157, Sep. 2017.

[24] M. A. Wani and S. Jabin, "Mutual clustering coefficient-based suspicious-link detection approach for online social networks," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published.

[25] D. Wu, J. Shi, and N. Mamoulis, "Density-based place clustering using geo-social network data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 838–851, May 2017.

YANG SUN received the master's degree in information science and engineering from Northeastern University. He is a Full Associate Professor with the Kexin Software College, Shenyang Normal University. He is also a Department Head of network engineering. His research interests include wireless networks, mobile computing, cloud computing, social networks, and network security. He had published more than 30 international journal and conference papers on the above research fields.

SHOULIN YIN was born in Puyang, China, in 1990. He received the B.S. and M.S. degrees from the Software College, Shenyang Normal University, Shenyang, China, in 2013 and 2015, respectively.

He is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Harbin Institute of Technology, Harbin, China. His research interests include image fusion, target detection, and image recognition.

HANG LI received the Ph.D. degree in information science and engineering from Northeastern University. He is a Full Professor with the Kexin Software College, Shenyang Normal University. He is also a master's Supervisor. His research interests include wireless networks, mobile computing, cloud computing, social networks, network security, and quantum cryptography. He had published more than 30 international journal and international conference papers on the above research fields.

LIN TENG received the B.Eng. degree from Shenyang Normal University, Shenyang, China, in 2016, where she is currently a Laboratory Assistant with the Software College. Her research interests include multimedia security, network security, filter algorithm, and data mining.

SHAHID KARIM received the B.S. degree in electronics from the COMSATS Institute of Information Technology, Abbottabad, Pakistan, in 2010, and the M.S. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, School of Electronics and Information Engineering, Harbin Institute of Technology, China. His current research interests include image processing, object detection, and classification toward remote sensing imagery.