Hindawi
Complexity
Volume 2019, Article ID 3475458, 13 pages
https://doi.org/10.1155/2019/3475458

Research Article

A Semantic Community Detection Algorithm Based on Quantizing Progress

Xu Han,[1] **Deyun Chen**,[1,2] **and Hailu Yang**[1,2]

[1]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China*
[2]*Postdoctoral Research Station of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China*

Correspondence should be addressed to Deyun Chen; chendeyun@hrbust.edu.cn and Hailu Yang; yanghailu@hrbust.edu.cn

Received 26 July 2018; Revised 25 November 2018; Accepted 11 December 2018; Published 9 January 2019

Academic Editor: Pasquale De Meo

The semantic social network is a kind of network that contains enormous nodes and complex semantic information, and the traditional community detection algorithms could not give the ideal cogent communities instead. To solve the issue of detecting semantic social network, we present a clustering community detection algorithm based on the PSO-LDA model. As the semantic model is LDA model, we use the Gibbs sampling method that can make quantitative parameters map from semantic information to semantic space. Then, we present a PSO strategy with the semantic relation to solve the overlapping community detection. Finally, we establish semantic modularity (SimQ) for evaluating the detected semantic communities. The validity and feasibility of the PSO-LDA model and the semantic modularity are verified by experimental analysis.

1. Introduction

With the development of society and the improvement of science and technology, semantic social networks are rapidly developed and many semantic networks, like Twitter and Weibo, have made an insignificant impact in our life so far. In these networks, different individuals have different small social "worlds" which are called communities [1]. Thus, researchers focus attention on community detection not only to divide networks into modules but also to make a deep insight into understanding interesting properties within the semantic social network. In practical application, semantic communities have a great promotion on intelligent information retrieval, marketing management, individual service, and other information management domains [2]. Heretofore, the research on community detection mainly reflects on the following three categories: topological community detection [3], community detection on overlapping construction [4], and semantic community detection.

The topological community detection represents the pioneer work, the goal of which is studying the topological constructions and dividing the social networks into several separate networks. The representative algorithms contain Modular Optimization [5], GN [6], and FN [7]. Then, researchers gradually focus on overlapping communities which can be more real than previous research networks. Therefore, CPM [8] was proposed to detect the overlapping communities. Soon afterwards, community detection on overlapping construction received more attention in social networks and many representative algorithms were proposed, including LFM [9], EAGLE [10], COPRA [11], DEMON [12], and so forth. Neuman and Yair [13] proposed an agglomerative spectral clustering method with conductance and edge weights. In their method, the most similar nodes are agglomerated based on eigenvector space and edge weights. But this method only is suitable for the nonsemantic social networks. Then, with the big interest in semantic network, semantic community detection came into researchers' eyes. Yang and McAuley [14] proposed the CESNA model to develop communities by using edge structure and node attributes. This method leads to more accurate community detection as well as improved robustness in the presence of noise in the network structure. But when this method applies into semantic network, it performs instable. Reihanian and

Ali [15] proposed a generic framework for overlapping community detection in social networks with special focus on rating-based social networks. This framework considers the information shared by the users in order to find meaningful communities. The most important feature of semantic communities is that the nodes in these communities not only have topological relationships, but also own semantic context. For the semantic data mining must be considered on the text analysis, and many semantic community detection algorithms applied the Latent Dirichlet Allocation (LDA) [16] model as the core model.

In the last few years, the analysis in semantic social network has become popular. Most of these algorithms utilize LDA model as the basic model. The SVM-DTW method proposed by Solera, Calderara, and Cucchiara [17] can work on the hierachical networks. This method makes simple structure and needs less input parameters, but the semantic context is not considered and the detected community has less connection with the real semantic network. Li and Ming and She [18] proposed the GRTM model which not only simulates users' interests as latent variables through their information, but also considers the connections between users as a result of their information. This method combines the context analysis with topological analysis and the similarity of the detected community is nearly close to the real semantic social network, but it is lack in the feature of sampling that would make some fuzzy irrelevant community. Xiao and Liu [19] proposed the GLDA-FP model which can be extended using the prediscretizing method which can help LDA model detect the topic evolution automatically, but the calculation required is large. As for the LCTA model proposed by Yin, Cao, and Gu [20] which makes the different topic distributions in different communities to make the model reasonable, this method has high accuracy in the result, but the number of communities needs to be preset and some hidden parameters need to be set up with experience.

In this paper, we propose a novel community detection algorithm for the objective of dividing nodes into clusters. The main characteristic of communities detected by this algorithm is that members of the same community have common or similar interests. We take into account the topic and keywords information in text from individuals' words through LDA model, then quantize semantic nodes, and map them into semantic space. Then, we get ideal virtual social communities after using Particle Swarm Optimization algorithm. Last but not least, we build a novel modular model and use the new function *SimQ* to evaluate the virtual social communities we make.

Compared with other models in semantic social network, such as lovain method model [21] and stochastic block model [22], the LDA model provides the probabilistic method so as to promote the foundation of mathematics. Then considering the following sampling, the Gibbs sampling can give an accurate and powerful mathematical proof for the convergence and solution of the LDA model, which is impossible to happen in the other semantic models. Combined with the PSO algorithm, the probability function compiled by LDA model can be closely integrated with the inertia weight and the constriction factor of the particles [23]. In performance measure, we propose a new module detecting evaluation model based on semantic information using the cosine function, which enriches the classic semantic detecting evaluation model.

The rest of the paper is organized as follows: Section 2 introduces LDA model in semantic network. Section 3 shows gibbs sampling and the proposed algorithm. In order to verify our approach, we conducted extensive experiments on a real data set. Performance evaluation and experimental results are shown and discussed in Sections 4 and 5. Finally, in Section 6 we make conclusions and envision further work.

2. Preliminaries

2.1. Community Detection Process. The problem of community detection belongs to NP-hard areas [24] which need initialize solutions at the beginning and optimize solutions constantly in the way of getting the best satisfying solution. The main goal of detecting semantic community is to form communities that individuals share common interests and probably they have similar characteristic [25]. So we show a novel idea that we focus on textual data of individuals' words. According to the complexity of community detection, we utilize the probabilistic graphical model–LDA to design network. This model has a most clearly hierarchical structure [26], and the scale of parameter spatial has no connection with the number of training documents.

First, we select topics and words from individuals' semantic information through LDA model. Then, we map semantic nodes into semantic space via Gibbs sampling method [27]. Last, in order to get more accurate communities, we use Particle Swarm Optimization (PSO) algorithm to form semantic communities. The proposed community detection algorithm is clearly explained in the following steps.

2.1.1. Similar Semantic Information Discovery. Every individual says different words as each node has its own information contents in semantic social network [28]. So we abstract semantic context into topic, and then we extract keywords from topic. Through semantic information, we convey some distributions to constrain our mess context [29]. In this way, dividing communities in semantic social network based on similar documents, topics, and keywords from social semantic contents make communities real [30]. The LDA probability model is shown in Figure 1.

In this section, we research LDA model on information contents. The relevant mathematical symbols for illustrating the LDA model are given in Table 1. LDA model assumes the following generative process for each node:

(1) $\theta \sim Dirichlet(\alpha)$. The parameter θ, which pertains to topic distribution, is subject to the Dirichlet distribution over a priori parameter α.

(2) $\varphi \sim Dirichlet(\beta)$. The parameter φ, which pertains to keyword distribution, is subject to the Dirichlet distribution over a priori parameter β.

(3) $z_i \mid \theta^{(d_i)} \sim Multinomial(\theta^{(d_i)})$. The topic z_i is subject to the multinomial distribution in case of topic distribution probability $\theta^{(d_i)}$.

Table 1: The symbol description.

SYMBOL	DESCRIPTION
N	Number of keywords in semantic social network
ω	Set of keywords in semantic social network, ω_i is the $i-th$ keyword in ω
d	Node set corresponding to keywords set ω, d_i is the $i-th$ node in the semantic social network
z	Topic set corresponding to keywords set ω, z_i is the $i-th$ topic in semantic social network
$\theta^{(d_i)}$	Topic distribution probability vector θ over node d_i
$\varphi^{(y)}$	Keyword distribution probability vector of topic y, $\varphi_{\omega_i}^{(y)}$ meaning the probability of keyword ω_i specific to topic y, $\varphi_{\omega_i}^{(y)} = P(\omega_i \mid z_i = y)$
α	A priori parameter over topic distribution probability specific to each node
β	A priori parameter over keyword distribution probability specific to a special topic

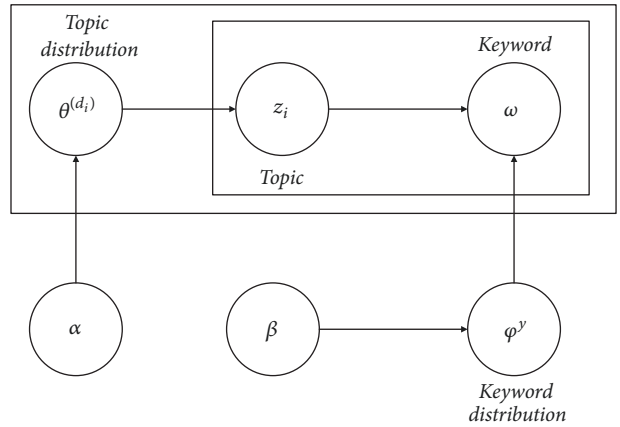


Figure 1: LDA probability model.

(4) $\omega_i \mid z_i, \varphi^{(z_i)} \sim Multinominal(\varphi^{(z_i)})$. The keyword ω_i is subject to the multinomial distribution in case of keyword distribution probability $\varphi^{(z_i)}$ over topic z_i.

The process of forming LDA model is shown in Algorithm 1. And M means the number of documents in the process.

3. Gibbs Sampling and PSO Strategy

3.1. Gibbs Sampling. Gibbs sampling [31] is a simple case of Markov-chain Monte Carlo (MCMC) [32] and aims at extracting a set of approximate samples from Markov-chain that is targeted to make a suitable probability distribution for converging to optimal solutions in high-dimensional models [33] such as LDA. According to the feature of Markov-chain, the probability-distribution function becomes the key to Gibbs sampling [34]. As for LDA in this text, we only sample topics in semantic social network; that is, we only need to consider hidden variety z_i. We denote $z_{\neg i}$ (topic set besides z_i) and $\omega_{\neg i}$ (set of keywords besides ω_i) to draw a posterior probability $P(z_i = y \mid z_{\neg i}, \omega_i)$. As for i, we can find the corresponding keyword ω_i. So the probability can be described as in the following equation.

$$P\left(z_i = y \mid z_{\neg i}, \omega_i\right) \propto P\left(z_i = y, \omega_i = t \mid \omega_{\neg i}, z_{\neg i}\right) \quad (1)$$

When $z_i = y$ and $\omega_i = t$ (t is one of the keywords in ω; y, which corresponds to t, is one of the topics in z), the probability $P(z_i = y, \omega_i = t \mid \omega_{\neg i}, z_{\neg i})$ only involves conjugate distribution of $d-th$ the document and $k-th$ topic under the Dirichlet-multinomial model.

We make $n_p^{[k]}$ as the number of $k-th$ topics in $d-th$ document, and the multinomial distribution can be described as

$$n_p = \left(n_p^{[1]}, n_p^{[2]}, \ldots n_p^{[K]}\right) \quad (2)$$

The number of $m-th$ keywords in $k-th$ topic, named $n_q^{[m]}$, can be shown as follows under multinomial distribution.

$$n_q = \left(n_q^{[1]}, n_q^{[2]}, \ldots n_q^{[M]}\right) \quad (3)$$

The posterior distribution of $\theta^{(d_i)}$ and $\varphi^{(z_i)}$ can be obtained in the following equations.

$$P\left(\theta^{(d_i)} \mid \omega_{\neg i}, z_{\neg i}\right) = Dirchlet\left(\theta^{(d_i)} \mid n_{p,\neg i} + \alpha\right) \quad (4)$$

$$P\left(\varphi^{(z_i)} \mid \omega_{\neg i}, z_{\neg i}\right) = Dirichlet\left(\varphi^{(z_i)} \mid n_{q,\neg i} + \beta\right) \quad (5)$$

$n_{p,\neg i}$ is the number of topics and $n_{q,\neg i}$ is the number of keywords.

(1) Extract the keyword distribution, and $\varphi \sim Dirichlet(\beta)$;
(2) **for** each $m \in [1, M]$ **do**
(3) extract N keywords, and $N \sim Poisson(\varphi)$;
(4) Extract topic distribution, and $\theta \sim Dirichlet(\alpha)$;
(5) **for** each $n \in [1, N]$ **do**
(6) Extract a topic, and this topic obeys $z_i \sim Multinomial(\theta^{(d_i)})$;
(7) Extract a keyword, and this keyword obeys $\omega_i \sim Multino\ \min\ al\ (\varphi^{(z_i)})$;
(8) **end for**
(9) **end for**

Algorithm 1: The generative process of LDA.

The distribution probability $P(z_i = y, \omega_i = t \mid \omega_{\neg i}, z_{\neg i})$ can be calculated by (6)~(11).

$$P\left(z_i = y, \omega_i = t \mid \omega_{\neg i}, z_{\neg i}\right) = \int P\left(z_i = y, \omega_i = t, \varphi^{(z_i)}, \theta^{(d_i)} \mid \omega_{\neg i}, z_{\neg i}\right) d\theta^{(d_i)} d\varphi^{(z_i)} \tag{6}$$

$$= \int P\left(z_i = y, \theta^{(d_i)} \mid \omega_{\neg i}, z_{\neg i}\right) P\left(\omega_i = t, \varphi^{(z_i)} \mid \omega_{\neg i}, z_{\neg i}\right) d\theta^{(d_i)} d\varphi^{(z_i)} \tag{7}$$

$$= \int P\left(z_i = y \mid \theta^{(d_i)}\right) Dirchlet\left(\theta^{(d_i)} \mid n_{p,\neg i} + \alpha\right) d\theta^{(d_i)} \tag{8}$$

$$\bullet \int P\left(\omega_i = t \mid \varphi^{(z_i)}\right) Dirichlet\left(\varphi^{(z_i)} \mid n_{q,\neg i} + \beta\right) d\varphi^{(z_i)} \tag{9}$$

$$= \frac{n^{y}_{p,\neg i} + \alpha}{\sum_{f=1}^{K} n^{f}_{p,\neg i} + \alpha} \frac{n^{t}_{q,\neg i} + \beta}{\sum_{g=1}^{V} n^{g}_{q,\neg i} + \beta} \tag{10}$$

$$\Longrightarrow P\left(z_i = y \mid z_{\neg i}, \omega_i\right) \propto \frac{n^{y}_{p,\neg i} + \alpha}{\sum_{f=1}^{K} n^{f}_{p,\neg i} + \alpha} \cdot \frac{n^{t}_{q,\neg i} + \beta}{\sum_{g=1}^{V} n^{g}_{q,\neg i} + \beta} \tag{11}$$

$n^{y}_{p,\neg i}$ is the amount of topics while $z_i = y$, $\sum_{f=1}^{K} n^{f}_{p,\neg i}$ is the amount of topics, $n^{t}_{q,\neg i}$ is the amount of keywords while $\omega_i = t$, and $\sum_{g=1}^{V} n^{g}_{q,\neg i}$ is the amount of keywords.

3.2. PSO Class Dependent LDA (PSO-LDA). Particle Swarm Optimization (PSO) is an intelligent optimization algorithm. It was first proposed by J.Kennedy and R.C.Eberhart [35]. PSO algorithm has the advantages of simplified, rather quick convergence [36] speed and less controlling parameter, and so forth.

Compared with other optimization algorithms, such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Simulate Anneal (SA), PSO algorithm has two attractive features: firstly, PSO optimizes the solution from the local optimum first and runs fast, which makes the algorithm more adaptable to the evolution of networks; secondly, particles in PSO can be mapped to nodes in semantic network; the process of finding the optimal solution in PSO is consistent with the birth process of the semantic community.

PSO puts a set of random solutions at system startup time and uses iterative search to find out optimal solutions [37]. In PSO, a solution of each optimization problem is called "particle". Each particle owns fitness value of itself. So we design a heuristic method to detect communities based on PSO. Each particle searches for the optimal solution by sharing social information between individuals.

In PSO-LDA, some LDA semantic feature is put into PSO. We use nodes in semantic social network mapping to "particle" in PSO and utilize semantic information vector of each node mapping to velocity of each particle in PSO. As for fitness value, we use information similar function instead. In PSO, we normalize that the nodes in semantic social network simulate the behavior of a "bird flock", where social sharing of information takes place, individuals' gains from the discoveries and previous experience of all other nodes during the search for food [38]. Thus, each node, called particle, in semantic social network which is called swarm, is assumed to "fly" over the search place looking for promising regions on the landscape.

First, we assume the search place is D – dimension space; and the $i - th$ particle position of the swarm is denoted as D – dimension, the vector $W_i = (w_{i1}, w_{i2}, \cdots, w_{id}, \cdots, w_{iD})$. Each particle has two pieces of message in the process: its "best" position with the smallest value (i.e., its personal best position) $P_i = (p_{i1}, p_{i2}, \cdots, p_{id}, \cdots, p_{iD})$ and the best function value of global particles in swarm (i.e., the global best position of all particles) $P_g = (p_{g1}, p_{g2}, \cdots, p_{gd}, \cdots, p_{gD})$. At each iteration, $i - th$ particle of the swarm updates its position and the velocity $V_i = (v_{i1}, v_{i2}, \cdots, v_{id}, \cdots, v_{iD})$ according to the following equation:

$$v^{s+1}_{id} = \eta v^{s}_{id} + \lambda_1 r^{s}_1 \left(p^{s}_{id} - w^{s}_{id}\right) + \lambda_2 r^{s}_2 \left(p^{s}_{gd} - w^{s}_{id}\right) \tag{12}$$

s is the current iteration, $j \in [1, 2, \cdots, D]$, $i \in [1, 2, \cdots, N]$, N represents the size of population, D is the dimension of the search place, η is the inertia weight, and λ_1 and λ_2 are two positive constants. r_1 and r_2 are study factors, that is, two random numbers extracted from the range [0, 1] for each dimension.

Input:
The semantic social network gragh disposed by LDA;
Output:
Useful transformable probability matrix;
Step 0. Initialize proper parameters, inertia weight $\eta = 0.632$, constriction factor $\xi = 0.729$, study factors $r_1 = 2.8$, $r_2 = 1.3$, population size(the size of network) $M = 200$, particle size (the number of nodes in semantic social network) $N = 1000$ and maximum iteration $MI = 200$.
Step 1. Initialize all particles and let $s = 0$;
Step 2. Evaluate fitness of each particle;
Step 3. Judge whether the ultimate criteria is satisfied. If $s > MI$, stop and jump to **Final**.; otherwise refresh variables according to the following steps;
Step 4. Refresh p_{id} by comparing the current fitness of each particle with its own historical best position p_{id}, if p_{id} gets smaller, then change it with the current position;
Step 5. Refresh p_{gd} by comparing the current best fitness of all particles with the historical best position p_{gd} of the whole swarm, if p_{gd} gets smaller, then change it with the current best position;
Step 6. Refresh v_{id}^{s+1} and w_{id}^{s+1} using Eq (12) and Eq (13);
Step 7. $s = s + 1$, return **Step 2**;
Final.

Algorithm 2: Optimization algorithm by PSO.

In the search place, once velocity v_{id}^{s+1} updated, the $i-th$ particle position w_{id} is changed as in the following equation.

$$w_{id}^{s+1} = w_{id}^{s} + \xi v_{id}^{s+1} \tag{13}$$

ξ is a constriction factor which manages and regulates the velocity's magnitude to maintain a balance between exploration and exploitation and it can be calculated as follows:

$$\xi = \frac{2}{\left|2 - \lambda - \sqrt{\lambda^2 - 4\lambda}\right|} \tag{14}$$

$\lambda = \lambda_1 + \lambda_2$, $\lambda > 4$. The constriction factor has influence on the proposed algorithm; we discuss the issue in part 4. The pseudocode for PSO is described in Algorithm 2 [39].

4. Performance Measure

Generally speaking, the performance measure of semantic social network is mostly based on the topological construction. And the EQ model proposed by Shen et al. [40] is widely used in evaluating overlapping communities, which is described in the following equation:

$$EQ = \frac{1}{R}\sum_{i}\sum_{v\in C_i, w\in C_i}\frac{1}{O_v O_w}\left[A_{vw} - \frac{k_v k_w}{R}\right] \tag{15}$$

k_v is the degree of node v and k_w is the degree of node w, $R = \sum_{vw} A_{vw}$ is the total degree of the network, A_{vw} is the element of adjacency matrix of the network, O_v is the number of communities which the node v belongs to and O_w is the number of communities which the node w belongs to, and C_i is the $i-th$ community in the network. For we use both topological construction and semantic context to detect communities, a novel evaluation model named $SimQ$, which we add information similarity into topological evaluation index, is given by the following equation.

$$SimQ = \frac{1}{R_1}\sum_{i,j}\sum_{d_i\in C_i, d_j\in C_j}\frac{Sim\left(d_i, d_j\right)}{O_{d_i} O_{d_j}}\left[A_{d_i d_j} - \frac{k_{d_i} k_{d_j}}{R_1}\right] \tag{16}$$

d_i is the $i-th$ node and d_j is the $j-th$ node, O_{d_i} is the number of communities that the node d_i pertains and O_{d_j} is the number of communities that the node d_j pertains, $R_1 = \sum_{d_i d_j} A_{d_i d_j}$ is the total degree of the network, $A_{d_i d_j}$ is the element of adjacency matrix of the network, and the range of value for $SimQ$ is $(0, 1)$. As for the information similarity $Sim(d_i, d_j)$, we give a normal social graph $G = (D, E, X_{d_i/d_j}^K, Sim(d_i, d_j))$, where D is a set of nodes in the network and d_i/d_j is the $i/j-th$ node; E is the set of edges linking to graph nodes. The actual point of $Sim(d_i, d_j)$ is to measure the structural correlation of nodes and add semantic correlation components at the same time. This is more suitable for the basic characteristics of the semantic communities. Each node d_i has connection with an information vector $X_{d_i}^K = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$; $Sim(d_i, d_j)$ is the information similarity of two neighbor nodes i and j which is calculated as

$$Sim\left(d_i, d_j\right) = \frac{\sum_{i,j=1}^{n}\left(X_{d_i}^K X_{d_j}^K\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(X_{d_i}^K\right)^2\right)\left(\sum_{j=1}^{n}\left(X_{d_j}^K\right)^2\right)}} \tag{17}$$

K is the dimension of the social network. In our method, if the semantic components of two nodes are close, the projection angles of these two nodes in two-dimensional space will be relatively small. On the contrary, the projection vectors are in contradictory situation.

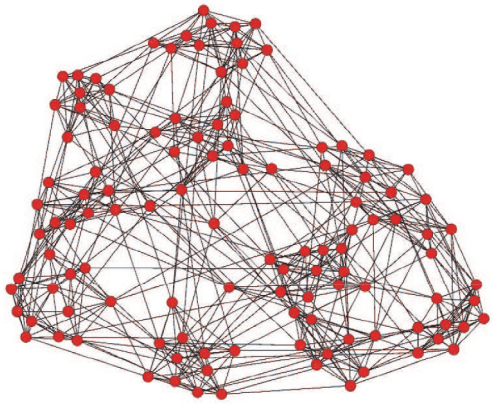


Figure 2: The graph of football network.

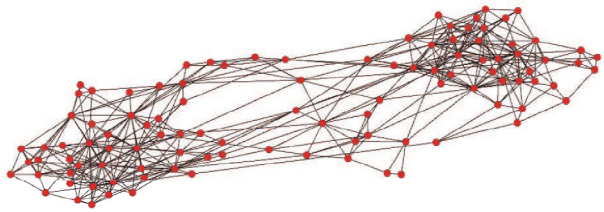


Figure 3: The graph of polbooks network.

5. Experimental Results

In this part, we would present and discuss the experiments with topics number analysis, evaluation criterion, real datasets, and different community detection algorithms, based on three datasets (the American College Football network dataset, the Krebs polbooks network dataset, and the dolphins network dataset).

5.1. The Analysis on Topics Number. The number of topics T, which is one of the input parameters in PSO-LDA model, can influence the compactedness of communities. So we choose the following three datasets to verify the effect of topics T over the result: (1) The American College Football network is shown in Figure 2. This network, created by Newman, is a complex social network about American College Football league. Nodes are regarded as football teams and one edge, between two neighbor nodes, represents that two football teams have played a match. It contains 115 nodes and 616 edges. (2) The Krebs polbooks network established by V.Kreb is shown in Figure 3. The nodes represent the politics books sold on Amazon. Generally, the books on political tendency are approximately divided into three classes. So in order to get topic distribution, Newman collected the political tendency in 3 steps away around each node. (3) The dolphins network collected by Newman is shown in Figure 4. The dolphins network is made up of two families, including 62 nodes and 159 edges. We simulate each node with the semantic information to fit on Dirichlet distribution.

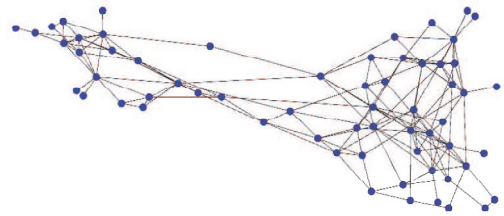


Figure 4: The dolphins network.

In this section, we use the topic number to experimentalize on three datasets (football, polbooks, and dolphins). Figure 5 shows the comparison of EQ and $SimQ$ on the three datasets with $T = (1, 2, \cdots, 20)$. While the topic number T grows bigger and the topic distribution rises higher, the number of detected communities gets bigger as T rises. In Figure 5, when the topic number gets larger to a certain degree, the topic distribution tends to be stable, resulting in the increment of communities. From the comparison of EQ and $SimQ$, these two performance measure models tend to decrease as T increases, since the topic number T arrives at an optimal point. The optimal value of T is 6 in Figure 5.

For the sake of getting communities more intuitive, Figure 6 shows the detected communities of three datasets when T is 6, 12, and 18.

5.2. The Comparison on Different Optimization Algorithms. In this section, we do the comparison on different optimization algorithms with three network datasets above (dolphins, polbooks, and football). We compare the number of communities, the size of communities, runtime, and semantic concentration with PSO algorithm, Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Simulate Anneal (SA). The result is shown in Figure 7. From Figure 7, we can see PSO algorithm makes more numbers of communities and smaller size of communities than others. As for runtime in PSO algorithm, it runs a little better than ACO and SA. The semantic concentration (SC) [41] is a function for measuring and testing degree of coagulation on specific topic and SC is shown in the following equation:

$$SC = \frac{\sum_{ij} Sim\left(d_i, d_j\right) \cdot \delta_{ij}}{\sum_{ij} Sim\left(d_i, d_j\right)} \tag{18}$$

δ_{ij} is the performance measure of communities links, while $\delta_{ij} = 1$ and only if i and j belong to the same community, there is a link between i and j. Compared with similarity function $SimQ$, SC makes focus on the stability of social groups in local environment. But what needs to be noted is that higher $SimQ$ does not mean higher SC in communities and higher SC does not mean we can get the best divisions; this is because the overlapping part of communities can effect the semantic cohesion. So the ideal community construction should be suitable with $SimQ$ and SC, and this also fits the performance measure of overall optimization and local optimization. Compared with GA, ACO, and SA

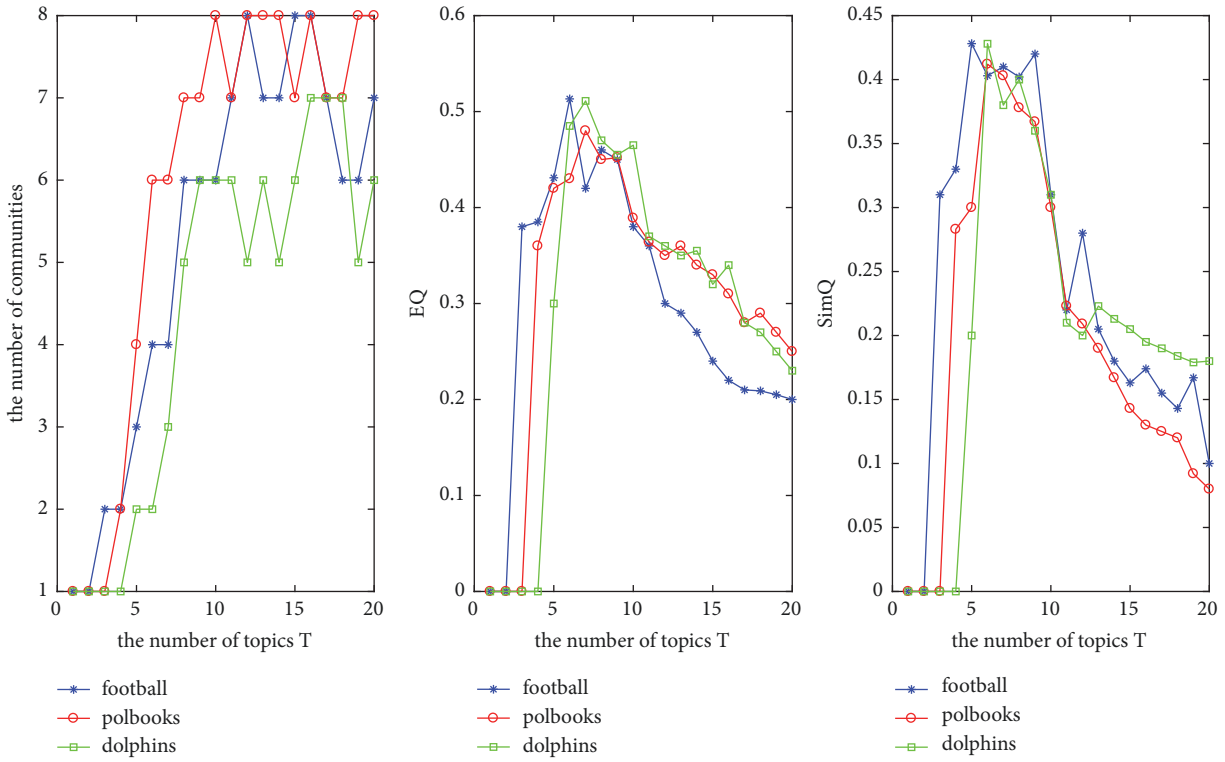


Figure 5: The performance of detected communities with T.

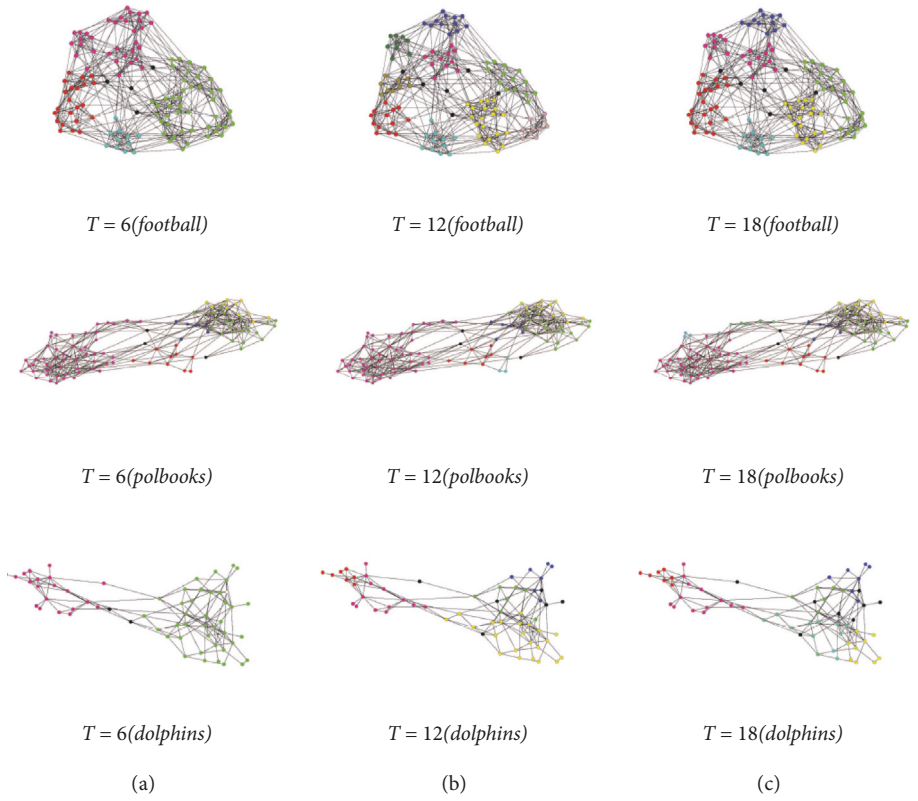


Figure 6: The communities for $T = \{6, 12, 18\}$ (the black nodes are overlapping nodes).

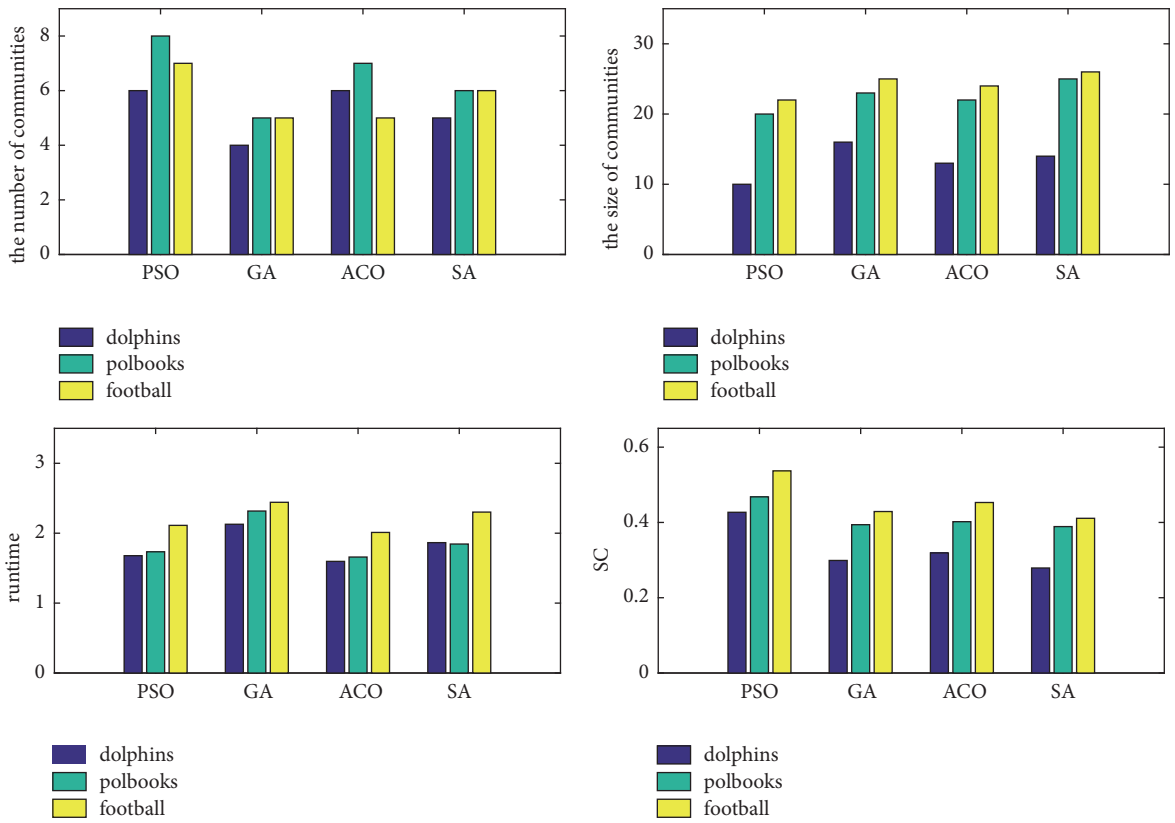


Figure 7: The performance of different optimization algorithms.

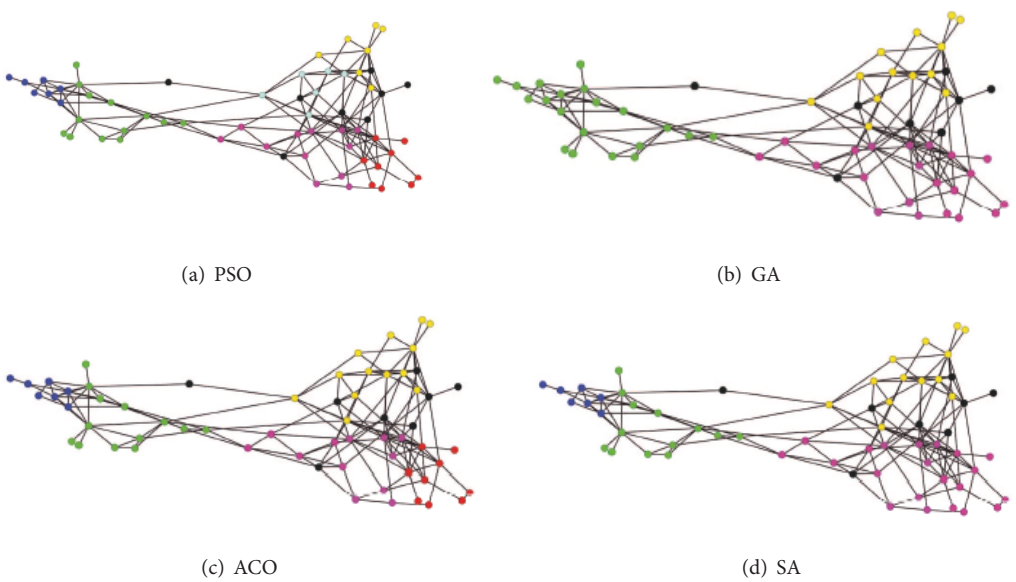


Figure 8: The comparison on different optimization algorithms on dolphins (the black nodes are overlapping nodes).

in Figure 7, the detected communities by PSO have a little small size and a bit more community numbers, which is in accordance with the topic distribution. As for runtime, PSO runs a bit slower than ACO but much better than GA and SA. Figure 8 shows four optimization algorithms run on dolphins network, and as similar as Figure 7, PSO works much better than other algorithms on community detection.

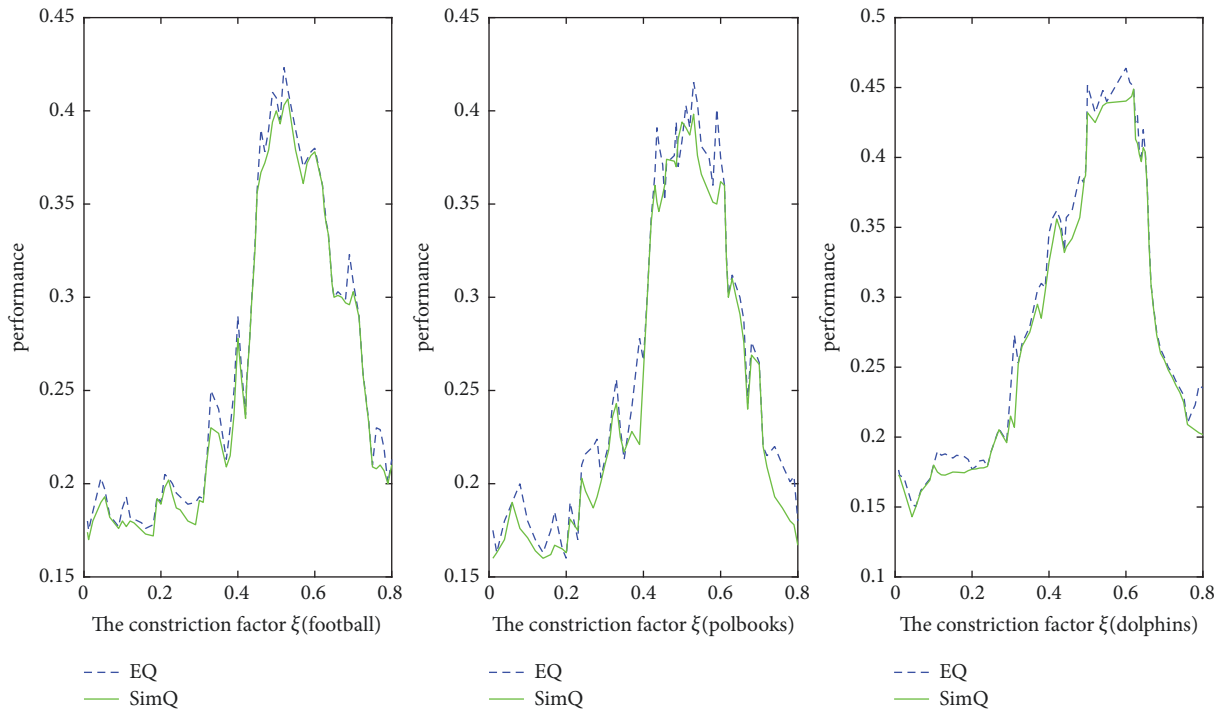


Figure 9: The digrams of comparison on the constriction factor with *EQ* and *SimQ*.

Table 2: The classical nonsemantic algorithms on *EQ*, *SimQ*, and *SC*.

Algorithms	*EQ*	*SimQ*	*SC*
GN	0.4615	0.3573	0.3873
FN	0.4061	0.3174	0.4012
LFM	0.3255	0.2331	0.3625
COPRA	0.5407	0.4115	0.3902
PSO-LDA	0.5132	0.4258	0.4842

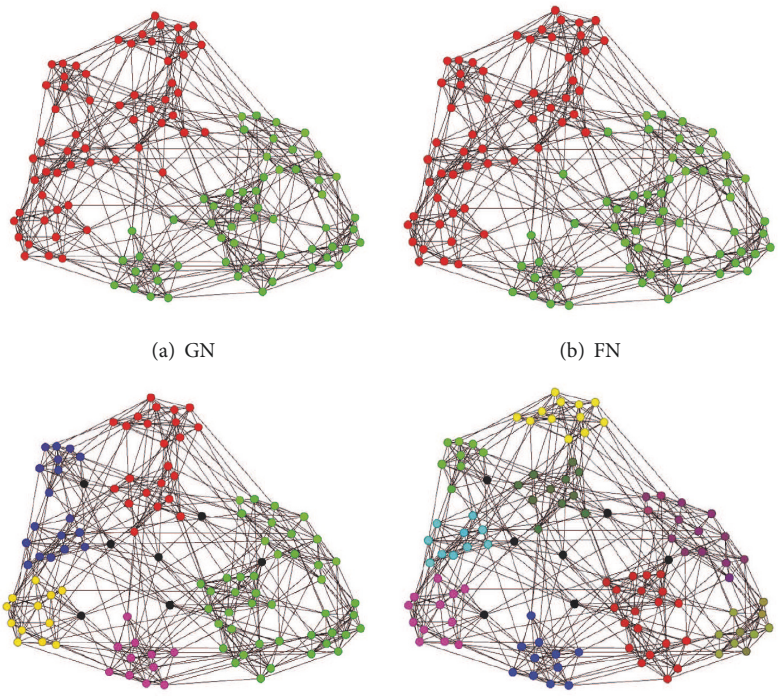
5.3. The Comparison on the Constriction Factor with EQ and SimQ. In this section, we compare *EQ* and *SimQ* over three datasets. The run diagrams, which *EQ* and *SimQ* run in three datasets, are shown in Figure 9. From (16), we put the similar function of information $Sim(d_i, d_j)$ into *SimQ* and $Sim(d_i, d_j) < 1$. So generally, the tendency of *EQ* diagram can be higher than *SimQ*. The maximum value of *EQ* in football dataset is 0.4233 (ξ =0.52) and *SimQ* is 0.4064 (ξ=0.53); and there exists bias when ξ =0.53, and the value of *EQ* is 0.4112 (not the maximum one). There is also bias in polbooks dataset and dolphins dataset, and the maximum value of *EQ* is 0.4154 (ξ =0.54) and *SimQ* is 0.3982 (ξ =0.55) in polbooks dataset while the maximum value of *EQ* is 0.4639 (ξ =0.60) and *SimQ* is 0.4489 (ξ =0.62) in dolphins dataset.

5.4. The Comparison on Community Detection Algorithms. Considering the bias in the semantic community detection, we utilize classical nonsemantic algorithms to illuminate the issue with the football dataset, for example.

We choose GN, FN, LFM, COPRA as nonsemantic classical algorithms, where LFM and COPRA are the overlapping community detection algorithms. The *EQ* and *SimQ* of the algorithms above are covered in Table 2 and the detection of communities is shown in Figure 10 with football dataset.

From the result in Table 2, the *EQ* of nonsemantic classical algorithms work higher than that of PSO-LDA (0.5132), but the *SimQ* works lower than PSO-LDA (0.4258). So it suggests that the nonsemantic classical algorithms make a higher *EQ* in the topological construction detection and a lower *SimQ* in the semantic detection. There is a bias in community detection by nonsemantic classical algorithms compared to semantic algorithms in the way of getting the ideal communities. On the one hand, we verify the performance of these algorithms; on the other hand, we use this experiment to verify the relation above *EQ*, *SimQ*, and *SC*. As for *SC* in Table 2, PSO-LDA performs better in *SimQ* and has high *EQ*, and PSO-LDA is higher than other algorithms in *SC*. This means PSO-LDA performs well in overall search (*EQ* and *SimQ*) and works better than others in local search (*SC*).

5.5. The Comparison on Real Datasets. In this section, we compare real different datasets, including Quantifying Link Semantics-Publication (QLSP) dataset (805 nodes), Academic Social Network (ASN) dataset (extract 2500 nodes) (https://www.aminer.cn/aminernetwork), extracting 10000 nodes and 20000 nodes from DBLP (December 31, 2014) dataset (2839219 nodes) (http://dblp.uni-trier.de/ db/) as DBLP(A) and DBLP(B), and Enron email network (Enron)



(a) GN (b) FN

(c) LFM (d) COPRA

Figure 10: The detected communities with nonclassical algorithms on football.

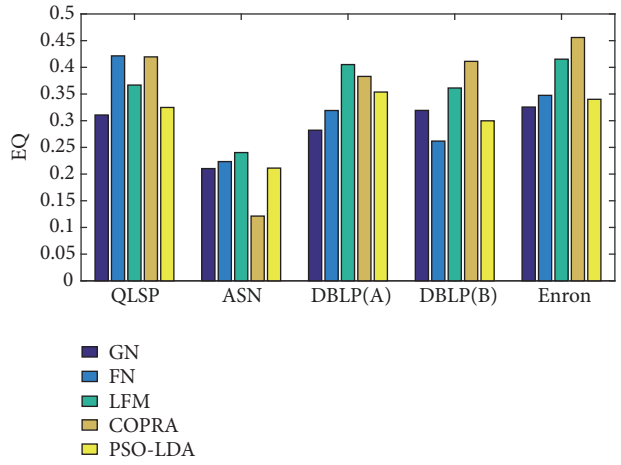


Figure 11: The histogram of *EQ* with various classical algorithms.

dataset (extract 25000 nodes) (http://snap.stanford.edu/data/email-Enron.html). The *EQ*, *SimQ*, and *NC*(the number of detected communities) of datasets above detected by various algorithms are reported in Table 3, as the PSO-LDA for $T = 6$. The histogram of *EQ* is shown in Figure 11 and *SimQ* in Figure 12. From Figures 11 and 12, the PSO-LDA model can be more suitable to solve the semantic community detection than the classical nonsemantic algorithms.

6. Conclusion

In this paper, we presented a novel community detection algorithm PSO-LDA that combines the topological construction with semantic information. It can avoid the number and the size of communities. For the Gibbs sampling solving the hidden parameter in the proposed model, the sampling result approaches to the realistic state. The main contribution of this research focuses on how to use different similarity measure to

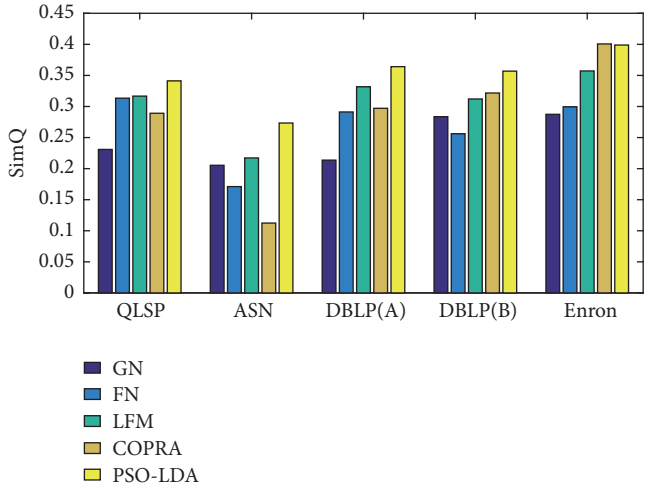


Figure 12: The histogram of *SimQ* with various classical algorithms.

Table 3: The results of classical nonsemantic algorithms under various datasets.

Algorithms	*EQ*/*SimQ*/*NC*	QLSP	ASN	DBLP(A)	DBLP(B)	Enron
GN	*EQ*	0.3107	0.2103	0.2822	0.3193	0.3256
	SimQ	0.2309	0.2054	0.2137	0.2863	0.2874
	NC	10	35	17	16	27
FN	*EQ*	0.4215	0.2234	0.3191	0.2618	0.3475
	SimQ	0.3134	0.1711	0.2912	0.2561	0.2994
	NC	10	33	19	16	26
LFM	*EQ*	0.3668	0.2403	0.4052	0.3613	0.4153
	SimQ	0.3167	0.2172	0.3317	0.3121	0.3572
	NC	12	29	21	12	30
COPRA	*EQ*	0.4196	0.1213	0.383	0.4112	0.4559
	SimQ	0.2891	0.1124	0.2971	0.3217	0.4007
	NC	13	31	21	13	26
PSO-LDA	*EQ*	0.3248	0.2112	0.3537	0.2998	0.3401
	SimQ	0.3412	0.2734	0.3641	0.3569	0.3989
	NC	14	30	23	15	27

measure similarity between nodes based on topological construction and their semantic information. As for future work, we would apply the model in some fields such as privacy protection and worm containment in semantic social network.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is sponsored by National Natural Science Foundation of China (61402126), Nature Science Foundation of Heilongjiang province of China (F2016024), Heilongjiang Postdoctoral Science Foundation (LBH-Z15095), University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (UNPYSCT-2017094), Heilongjiang Province Foundation for Returned Scholars (LC2018030), and National Training Programs of Innovation and Entrepreneurship for Undergraduates (201810214020). The paper is also supported by China Natural Science Fund.

References

[1] S. Fortunato and D. Hric, "Community detection in networks: a user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[2] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the International Conference on World Wide Web*, pp. 1089–1098, 2013.

[3] U.-U. Narantsatsralt and S. Kang, "Social network community detection using agglomerative spectral clustering," *Complexity*, vol. 2017, Article ID 3719428, 10 pages, 2017.

[4] C.-D. Wang, J.-H. Lai, and P. S. Yu, "NEIWalk: Community discovery in dynamic content-based networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1734–1748, 2014.

[5] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 70, no. 2, Article ID 066111, 2004.

[6] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, article 066133, 2004.

[7] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 69, no. 2, article 026113, 2004.

[8] G. Palla, I. Dere Nyi, I. S. Farkas, and T. S. Vicsek, "Uncovering the overlapping community structure," *Nature*, vol. 435, no. 7043, pp. 398–406, 2005.

[9] A. Lancichinetti, S. Fortunato, and J. Kertesz, "Detecting the overlapping and hierarchical community structure of complex networks," *New Journal of Physics*, vol. 11, no. 3, pp. 19–44, 2012.

[10] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. 155–168, 2008.

[11] H. A. Deylami and M. Asadpour, "Link prediction in social networks using hierarchical community detection," in *Information and Knowledge Technology*, pp. 1–5, 2015.

[12] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 905–918, 2014.

[13] Y. Neuman, Y. Neuman, and Y. Cohen, "A novel procedure for measuring semantic synergy," *Complexity*, vol. 2017, Article ID 5785617, 8 pages, 2017.

[14] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM '13)*, pp. 1151–1156, 2013.

[15] A. Reihanian, M. R. Feizi-Derakhshi, and H. S. Aghdasi, "Overlapping community detection in rating-based social networks through analyzing topics, ratings and links," *Pattern Recognition*, 2018.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[17] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Computer Society*, 2016.

[18] X. Li, C. Ming, and J. She, "Connection discovery using shared images by gaussian relational topic model," in *Proceedings of the IEEE International Conference on Big Data*, pp. 931–936, 2017.

[19] Y. Xiao, L. Liu, M. Xu, H. Wang, and Y. Liu, "Glda-fp: Gaussian lda model for forward prediction," in *Proceedings of the International Conference on Big Data*, pp. 124–139, 2018.

[20] X. Yu, J. Yang, and Z. Q. Xie, "A semantic overlapping community detection algorithm based on field sampling," *Expert Systems with Applications*, vol. 42, no. 1, pp. 366–375, 2015.

[21] S. Gupta and P. Kumar, "Community detection in heterogenous networks using incremental seed expansion," in *Proceedings of the 2016 International Conference on Data Science and Engineering (ICDSE)*, pp. 1–5, 2017.

[22] Y. Zhao, E. Levina, and J. Zhu, "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012.

[23] W. B. Towne, C. P. Rosé, and J. D. Herbsleb, "Measuring similarity similarly: Lda and human perception," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 1, 2016.

[24] S. Cavallari, V. W. Zheng, H. Cai, K. C.-C. Chang, and E. Cambria, "Learning community embedding with community detection and node embedding on graphs," in *Proceedings of the 26th ACM International Conference on Information and Knowledge Management, (CIKM '17)*, pp. 377–386, 2017.

[25] Z. Yin, L. Cao, Q. Gu, and J. Han, "Latent community topic analysis: Integration of community discovery with topic modeling," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, pp. 1–21, 2012.

[26] Z. Xia and Z. Bu, "Community detection based on a semantic network," *Knowledge-Based Systems*, vol. 26, pp. 30–39, 2012.

[27] F. Zhao, Y. Zhu, H. Jin, and L. T. Yang, "A personalized hashtag recommendation approach using LDA-based topic model in microblog environment," *Future Generation Computer Systems*, vol. 65, pp. 196–206, 2016.

[28] S. Ahajjam, M. El Haddad, and H. Badir, "A new scalable leader-community detection approach for community detection in social networks," *Social Networks*, vol. 54, pp. 41–49, 2018.

[29] X. Yang and J. Cao, "A Fast and accurate way for API network construction based on semantic similarity and community detection," in *Proceedings of the IFIP International Conference on Network and Parallel Computing*, pp. 75–86, 2017.

[30] C. X. Zhai, "Probabilistic topic models for text data retrieval and analysis," in *Proceedings of the International ACM SIGIR Conference*, pp. 1399–1401, 2017.

[31] G. Heinrich, "Parameter estimation for text analysis," Technical Report, 2008.

[32] W. K. Hastings, "Monte carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[33] M. Sachan, D. Contractor, T. A. Faruquie, and V. L. Subramaniam, "Using content and interactions for discovering communities in social networks," in *Proceedings of the International Conference on World Wide Web*, pp. 331–340, 2012.

[34] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Community detection with edge content in social media networks," in *Proceedings of the IEEE 28th International Conference on Data Engineering, (ICDE '12)*, pp. 534–545, 2012.

[35] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995.

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

[37] H. Abadlia, N. Smairi, and K. Ghedira, "Particle swarm optimization based on island models," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 49-50, 2017.

[38] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence, (AAAI '16)*, pp. 265–271, 2016.

[39] N. A. Helal, R. M. Ismail, N. L. Badr, and M. G. Mostafa, "An efficient algorithm for community detection in attributed social networks," in *Proceedings of the International Conference on Informatics and Systems*, pp. 180–184, 2016.

[40] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, 2009.

[41] A. Clauset, "Finding local community structure in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 72, no. 2, Article ID 026132, 2005.