ICIC Express Letters
Volume **11**, Number **2**, February **2017**
 ISSN 1881-803X
pp. **437–444**

RESEARCH ON COMMUNITY DISCOVERY IN SOCIAL NETWORK BASED ON CONNECTION DEGREE

Xiao Chen[1,2,3], Jingfeng Guo[1,3] and Junli Yu[2]

[1]College of Information Science and Engineering
[3]The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province
Yanshan University
No. 438, Hebei Ave., Qinhuangdao 066004, P. R. China
jfguo@ysu.edu.cn

[2]College of Qian'an
North China University of Science and Technology
No. 5096, Yanshan Road, Qian'an 064000, P. R. China
chenxiao0604@163.com

Received July 2016; accepted October 2016

Abstract. *With set pair analysis theory, considering social network as a complex IDC (Identical-Different-Contrary) system, a new community discovery algorithm CDCD (Community Discovery based on Connection Degree) is proposed. Firstly, considering the local features and the topological structure attributes, based on set pair connection degree to describe the IDC relation between vertices, a new similarity index is defined, which is the connection degree taking account of weight and clustering coefficient. The measurement can better describe network structure characteristics in lower complexity. Secondly, we propose a new community discovery algorithm CDCD based on the measurement, which considers the network connection degree as the stop threshold of community clustering. Finally, the correctness and effectiveness of the algorithm CDCD are terrified through the experiments.*
Keywords: Set pair, Connection degree, IDC, Community discovery

1. **Introduction.** Community discovery [1] as the basis of social networks analysis has important theoretical significance and practical value to the community evolution and prediction. Thus the related research has become one of the common focuses and aroused much concern of many scholars both at home and abroad. At present, many methods have been put forward to detect communities, such as, the GN algorithm [2] based on betweenness, the CNM algorithm [3] based on modularity, the LPA algorithm [4] based on label propagation and the SC algorithm [5] based on spectral optimization. The main purpose of community discovery is dividing the vertices of tight relation into the same community and those who have sparse relation into different ones. Therefore, the relation (similarity) between vertices has the crucial influence to the effectiveness of community discovery. Now, the similarity between vertices is mainly divided into three categories [6]: global similarity indices, local similarity indices and semi-local similarity indices. Among them, the global indices need to consider the overall network structure, although they have a relatively higher accuracy, they have higher complexity, which is infeasible for large-scale networks. Compared to the global indices, the local ones only consider the nearest neighbor vertices and have lower time complexity, but underestimate the similarity between vertices. It is the main research content that how to combine the vertex's local features and the network topological structure to improve the accuracy of the similarity in the lower complexity.

Set Pair Analysis (short for SPA) [7], as a new theory and method, is proposed by Chinese scholar Keqin Zhao in 1989. It is a core idea that any system (thing) is constituted

by the certainty and uncertainty information, random uncertainty and fuzzy uncertainty which can transform into certainty under some conditions. Hence, we can use set pair theory to solve the social network with incompleteness and uncertainty. In 2011, the SPA theory was firstly applied to social network analysis, and the social network analysis model based on set pair and its property are proposed in [8]. The social network analysis research based on set pair theory opened the prelude. In 2013, a connection degree between vertices based on SPA and the common neighbors is proposed in [9]. In 2014, a new connection degree between vertices based on attribute graph and set pair situation is proposed to achieve community discovery in [10]. However, the existing similarity indices based on connection degree, only consider the influence from the number of uncertain attributes and the number of uncertain common neighbors in network to community formation and network analysis. And these methods ignore the influence of presence paths, direct or indirect paths, the number and length of paths, vertex degree, network density (clustering coefficient) and so on.

Therefore, based on the idea of set pair analysis theory, considering the social network as a complex IDC (Identical-Different-Contrary) system, we propose a new algorithm CDCD (Community Discovery based on Connection Degree). Firstly, redefine the IDC relation between vertices under social network structure; the similarity between vertices is proposed based on weighted clustering coefficient connection degree. Secondly, we propose a new algorithm CDCD of community discovery based on this measurement. Finally, compared with the typical algorithms, the correctness and effectiveness of the algorithm CDCD are terrified through the experiments.

The main contributions of this paper are summarized as follows.

(1) The similarity between vertices based on connection degree with weight and clustering coefficient is proposed. Combining the network topology characteristics, such as the clustering coefficient, the vertex degree and the paths, applying set pair connection degree $u = a + bi + cj$, we propose a new measurement of similarity between vertices in this paper. According to the contribution made to the similarity between vertices when micro uncertain (different) relation i and the macro uncertain (different) relation b transform into the certain identical relation a, considering the influence of link density between vertices, this measurement takes the vertex clustering coefficient as the evaluation method of i. Considering the contribution of vertex degree and paths to the similarity, it gives weight to IDC relation between vertices. It can better reflect the characteristics of network topology structure and avoids the incompleteness of only considering the single number of the IDC relation, and avoids the computational complexity of the path between vertices. It dose not only reflect the characteristics of the network topology, but also improves the accuracy of the similarity index.

(2) Aggregation community discovery algorithm based on connection degree between vertices is proposed. Traditional aggregation clustering algorithm based on average distance has a lot of updating average distance operations, and there are some reasonable phenomena of community clustering. In order to avoid the frequent update operations, to ensure the accuracy of community partition, the algorithm first takes priority to corporate vertices of high similarity into small community, and then remerge the small communities through the higher mean of similarity between communities; finally, take network connection degree as the stop threshold of communities cluster.

2. **Social Network Analysis Method Based on Set Pair.** Given set pair $H = (A, B)$, under a specific background, the attributes of two related sets A and B are analyzed, and N attributes are obtained. Among them, S represents the identical attributes, P represents the contrary attributes, the rest $F = N - S - P$ represents the different attributes in set pair. The two sets of connection degree u [7] is shown in Formula (1).

$$u = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \tag{1}$$

where $i \in [-1, 1]$ is the different marker; and $j = -1$ only plays the marked role.

If let $\frac{S}{N} = a$, $\frac{F}{N} = b$ and $\frac{P}{N} = c$, then u can be abbreviated as shown in Formula (2).

$$u = a + bi + cj \ (a + b + c = 1) \tag{2}$$

In [8,9], the connection degree is applied to the field of social network analysis, taking network structure as the foundation, given v_k and v_s as the two study objects, the adjacency relation between vertices as vertex's attributes, N represents the total attributes, and then the connection degree is shown in Formula (3).

$$\begin{aligned} u(v_k, v_s) = &\frac{|N(v_k)_1 \cap N(v_s)_1|}{N} + \frac{|V| - |N(v_k)_1 \cup N(v_s)_1|}{N}i \\ &+ \frac{|N(v_k)_1 \cup N(v_s)_1| - |N(v_k)_1 \cap N(v_s)_1|}{N}j \end{aligned} \tag{3}$$

In [10], the improved connection degree is shown in Formula (4).

$$\begin{aligned} u(v_k, v_s) = &\frac{|N(v_k)_1 \cap N(v_s)_1|}{N} + \frac{|N(v_k)_2 \cap N(v_s)_2|}{N}i \\ &+ \frac{|V| - |N(v_k)_1 \cap N(v_s)_1| - |N(v_k)_2 \cap N(v_s)_2|}{N}j \end{aligned} \tag{4}$$

Instance social network is shown in Figure 1. In Figure 1, let v_1 and v_3 as the two study objects, based on Formula (3), the identical attributes between v_1 and v_3 are v_2 and v_4, their contrary attributes are v_9 and v_{10}, and their different attributes are v_5, v_6, v_7 and v_8. Based on the probability that a friend's friend is more likely to become friends, we can infer that the contrary attributes, compared with the different attributes, can be more easily converted into the identical attributes. Therefore, this method is not reasonable. Based on Formula (4), the identical attributes between v_1 and v_3 are v_2 and v_4, their different attribute is v_5, and their contrary attributes are v_6, v_7, v_8, v_9 and v_{10}. The above problems still exist. Meanwhile, those methods only consider the effect of the number of common neighbors to similarity without considering the effects of different network density, clustering coefficient and vertex degree. So it cannot better reflect the network structure.

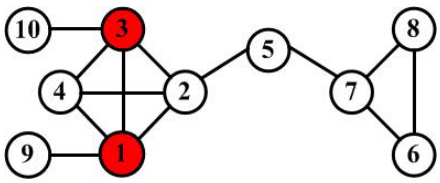


FIGURE 1. Instance social network

3. **Connection Degree Model.**

Definition 3.1. *Given $G = (V, E)$, for any vertex v_k, the 1-level neighbor set of vertex v_k is denoted as $N(v_k)_1$, as shown in Formula (5).*

$$N(v_k)_1 = \{v_q | (v_q, v_k) \in E \cap v_q \in V\} \cup \{v_k\} \tag{5}$$

Definition 3.2. *Given $G = (V, E)$, for any vertex v_k, the 2-level neighbor set of vertex v_k is denoted as $N(v_k)_2$, as shown in Formula (6).*

$$N(v_k)_2 = \{v_q | (v_q, v_p) \in E \cap v_p \in N(v_k)_1, v_q \notin N(v_k)_1\} \tag{6}$$

For any vertex v_k and v_s, the common 1-level neighbor set is denoted as $CN(v_k, v_s)_1 = N(v_k)_1 \cap N(v_s)_1$; the common 2-level neighbor set is denoted as $CN(v_k, v_s)_2 = N(v_k)_2 \cap N(v_s)_2$; the common 1∩2-level neighbor set is denoted as $CN(v_k, v_s)_{1\cap2} = N(v_k)_1 \cap N(v_s)_2$, the common $2 \cap 1$-level neighbor set is denoted as $CN(v_k, v_s)_{2\cap1} = N(v_k)_2 \cap N(v_s)_1$. Note: Any vertex only belongs to a certain class of neighbor set, the ownership order: $CN(v_k, v_s)_1$, $CN(v_k, v_s)_{1\cap2}$, $CN(v_k, v_s)_{2\cap1}$, $CN(v_k, v_s)_2$.

Definition 3.3. *Given $G = (V, E)$, for any vertex v_k and v_s, the identical attributes between v_k and v_s are the common 1-level neighbors, that is $S = |CN(v_k, v_s)_1|$, the different attributes are the common $1\cap2$, $2\cap1$ and 2-level neighbors, that is $F = |CN(v_k, v_s)_{1\cap2}| + |CN(v_k, v_s)_{2\cap1}| + |CN(v_k, v_s)_2|$; the contrary attributes are the rest vertices, that is $P = N - S - F$, where $N = |V|$.*

Definition 3.4. *Given $G = (V, E)$, for any vertex v_k and v_s, the connection degree with weight and clustering coefficient $u(v_k, v_s)$ is shown in Formula (7).*

$$u(v_k, v_s) = \frac{(1)_{1\times S} \times w(v_i)_{S\times1}}{N} + \frac{w(v_i)_{1\times F}}{N} \times i(v_i)_{F\times1} + \frac{(1)_{1\times P} \times w(v_i)_{P\times1}}{N} \times j \tag{7}$$

Among them, N, S, F and P are shown in Definition 3.3. $(1)_{1\times S}$, $(1)_{1\times F}$ and $(1)_{1\times P}$ respectively represent the row vector of the identical, different, contrary attributes, and the vector value is 1; $w(v_i)$ is the weight value; $(i(v_i))_{F\times1}$ is the difference value i for vertex v_i, and j only has the marked effect.

In Formula (7), j is marked effect of the contrary attributes; however, in community discovery, it is expected to divide the vertices with more relation to the same community, so let $j = 0$. i as a different mark on the micro level, consider the density structure characteristics in network, the i value is quantified by clustering coefficient. $w(v_i)$ is described through degree, path reachability and transition probability.

Definition 3.5. *Given $G = (V, E)$, for any vertex v_k, the connection degree of v_k is the average of connection degree between v_k and all vertices, denoted as $u(v_k)$, as shown in Formula (8).*

$$u(v_k) = \sum_{s=1}^{|V|} \frac{u(v_k, v_s)}{|V|} \tag{8}$$

Definition 3.6. *Given $G = (V, E)$, the network connection degree is the average of all vertices' connection degree in network, denoted as $u(G)$, as shown in Formula (9).*

$$u(G) = \frac{\sum_{k=1}^{|V|}\sum_{s=1}^{|V|} u(v_k, v_s)}{|V|} = \frac{\sum_{k=1}^{|V|} u(v_k)}{|V|} \tag{9}$$

Definition 3.7. *Given $G = (V, E)$, for any two communities $C_K = (V_K, E_K)$ and $C_S = (V_S, E_S)$, the connection degree between communities is denoted as $u(C_K, C_S)$, as shown in Formula (10).*

$$u(G) = \frac{\sum_{k=1}^{|C_K|}\sum_{s=1}^{|C_S|} u(v_k, v_s)}{|C_K| \times |C_S|} \tag{10}$$

4. **Algorithm of Community Discovery.** The detail algorithm description is the following.

Algorithm 1 CDCD (Community Discovery based on Connection Degree)
Input: Social Network $G = (V, E)$
Output: Sub-community set C_k $(k = 1, \ldots, K)$
BEGIN
(1) Initialize each vertex into a community, that is $C_k = v_k$ $(k = 1, \ldots, |V|)$.
(2) According to Formula (7), the $u(v_k, v_s)$ is calculated, and the matrix R is obtained.

(3) According to Formula (8), the $u(v_k)$ is calculated.

(4) $Max\{u(v_k)\}$ is selected. If $u(v_k, v_s) \geq 4 \times u(G)$, then v_s is clustered into v_k's community.

(5) Repeat (4), until all vertices could not be further clustered.

(6) If v_s is not clustered, and find $Max\{u(v_k, v_s)\}$, then make v_s into v_k's community.

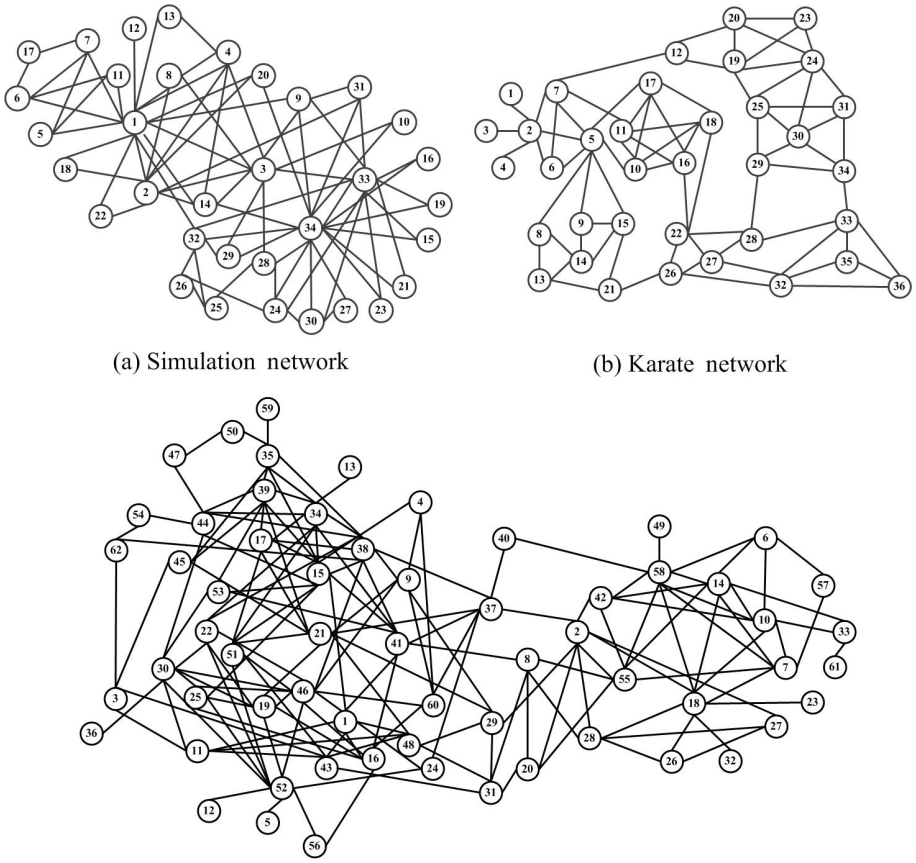
(7) According to Formula (10), the $u(C_k, C_s)$ is calculated, and find $Max\{u(C_k, C_s)\}$. If $Max\{u(C_k, C_s)\} \geq u(G)$, then C_k and C_s are clustered, that is C_{new}.

(8) Update the connection degree between C_{new} and other communities.

(9) Repeat (7) and (8), until $Max\{u(C_k, C_s)\} < u(G)$.

END

5. **Experimental Results and Analysis.** Experiment hardware environment is Intel Core i5-3337U CPU@1.8GHz, with 4GB memory. Software environment is Windows 7 and Matlab R2012a. Experiment datasets: simulated network is shown in Figure 2(a); Zachary karate club network is shown in Figure 2(b); dolphin network is shown in Figure 2(c).



(a) Simulation network

(b) Karate network

(c) Dolphin network

FIGURE 2. Social network data set

This paper experimentally verifies mainly from two aspects. (1) Through the comparison of three kinds of similarity based on connection degree, the correctness and rationality of similarity index proposed by this paper shall be verified. (2) Through the comparison of five community discovery methods, the correctness and effectiveness of the algorithm CDCD proposed by this paper shall be verified. Among them, the advantages and disadvantages of community structure are evaluated by NMI function.

Normalization Mutual Information (NMI) [11] as an index to evaluate the quality of community, has been widely used, its calculation method as shown in Formula (11).

$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{C_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{11}$$

where A is the community structure of real network, B is the community structure by algorithm; C_A and C_B are the community's numbers of A and B network, respectively; $N_{i.}$ and $N_{.j}$ are the number of vertices in the community i and j respectively, N_{ij} is the number of vertices shared by the community i and j; N is the number of vertices in network. The larger NMI value, the better community partition.

5.1. **Experimental analysis of the similarity indices.** Using CDCD algorithm, we respectively perform community mining based on the similarity indices proposed by [8], [10] and this paper in three benchmark networks. The experiment results are shown in Table 1. In Table 1, the 1st column is the real network list; the 2nd column is the basic network statistics: number of vertices/edges; the 3rd to 5th columns are three indices, for which the statistical communities' number and NMI value are listed. The maximum value of NMI in the table is identified in bold font. As can be seen, by comparing the NMI value, using the similarity index proposed in this paper, the maximum value is obtained, and it is more reasonable and also better reflects the results of community discovery.

Table 1. Experiment results of three similarity indices

	Vertices/Edges	Formula (3)	Formula (4)	Formula (11)
Simulation	36/69	6/0.96	6/0.83	6/**1.00**
Karate	34/78	5/0.57	3/0.65	2/**0.73**
Dolphin	62/159	3/0.88	2/**1.00**	2/**1.00**

5.2. **Experimental analysis of community discovery.** In order to verify the accuracy and validity of the community structure of the algorithm CDCD, we compare it with GN, CNM, LPA and SC algorithm in three benchmark networks. The experimental results are shown in Table 2. In Table 2, the communities number, NMI value and the running time for each algorithm are listed statistically. The maximum value of NMI in the table is identified in bold font; the second maximum value is in underlined bold font. As can be seen, the algorithm CDCD can get maximum value of NMI in three benchmark networks, so it also better reflects the results of community discovery.

Table 2. Experiment results of community discovery algorithms

	GN	CNM	LPA	SC	CDCD
Simulation	6/0.87/0.36	5/0.89/0.04	6/0.87/0.02	6/**<u>0.91</u>**/0.01	6/**1.00**/0.05
Karate	5/0.58/0.38	3/0.69/0.04	3/0.57/0.02	2/**0.84**/0.01	2/**<u>0.73</u>**/0.15
Dolphin	5/0.55/1.44	4/0.57/0.07	4/0.43/0.04	2/**<u>0.88</u>**/0.02	2/**1.00**/0.33

Similarly, in a real data concentration, the running time efficiency of algorithm GN, CNM, LPA, SC and CDCD is verified. The experiment result is shown in Figure 3. With the increase of the number of vertices and edges in the network, the running time of each algorithm increases significantly. The greedy CNM algorithm and LPA algorithm run faster, so they are more suitable for processing large network. Since the SC algorithm uses ARPACK to accelerate the calculation method of characteristic root, with the increase of network scale, its running time increases slowly. The running speed of GN algorithm is the

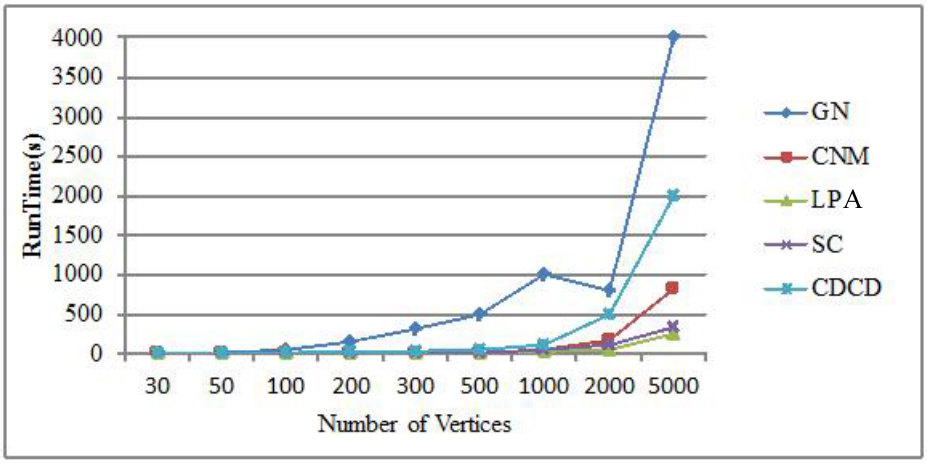


FIGURE 3. Relations between number of vertices and runtime

slowest because it needs to calculate betweenness from a global perspective. The running efficiency of CDCD algorithm is obviously better than GN algorithm. Although it is inferior to CNM, LPA and SC algorithm, the community division effect is obviously better than other algorithms. The speed increases depend on the community discovery effect. Thus, the CDCD algorithm based on connection degree between vertices in community discovery, has a better effect of community division.

6. **Conclusions.** Considering social network as an identical-different-contrary system, in allusion to network topology structure, based on set pair connection degree $u = a + bi + cj$, this paper described and analyzed the measures of similarity between vertices, and proposed the model expression and calculation method of similarity between vertices based on the connect degree with weight and clustering coefficient. This paper further put forward the similarity model between network communities. In order to further investigate the similarity index, this is applied to community discovery in the social network datasets. The experimental results show that the similarity index can realize correct and effective community discovery. Thus how to describe the similarity between vertices in a more complex network, and how to study the dynamic evolution of the network are the further research targets.

Acknowledgment. This work is supported by the National Science Foundation of China, (No. 61472340), and is supported by the National Youth Science Foundation of China (No. 61602401), and is supported by the Nature Science Foundation of Hebei Province, China (No. F2016209344). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, who have improved the presentation.

REFERENCES

[1] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, *Proc. of the National Academy of Sciences of the United States of America*, vol.99, no.12, pp.7821-7826, 2002.

[2] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical Review E*, vol.69, no.2, 2004.

[3] A. Clauset, M. E. J. Newman and C. Moore, Finding community structure in very large networks, *Physical Review E*, vol.70, no.6, 2004.

[4] U. N. Raghavan, R. Albert and S. Kumara, Near linear time algorithm to detect community structures in large-scale net-works, *Physical Review E*, vol.76, no.3, 2007.

[5] M. E. J. Newman, Finding community structure using the eigenvectors of matrices, *Physical Review E*, vol.74, no.3, 2006.

[6] Q. Jiao, Y. Huang and H. Shen, Community mining with new node similarity by incorporating both global and local topological knowledge in a constrained random walk, *Physical A*, vol.424, no.2015, pp.363-371, 2015.
[7] K. Zhao, *Set Pair Analysis and Its Preliminary Application*, Zhejiang Science and Technology Press, Hangzhou, 2000.
[8] C. Zhang, R. Liang and L. Liu, Set pair social network analysis model and its application, *Journal of Hebei Polytechnic University (Natural Science Edition)*, vol.33, no.3, pp.99-103, 2011.
[9] C. Zhang and J. Guo, The α relation communities of set pair social networks and its dynamic mining algorithms, *Chinese Journal of Computers*, vol.36, no.8, pp.1682-1692, 2013.
[10] C. Zhang and J. Guo, *The Attribute Graph Model of Social Networks and Its Application*, Beijing University of Posts and Telecommunications Press, Beijing, 2014.
[11] A. Strehl and J. Ghosh, Cluster ensembles – A knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.*, vol.3, no.2002, pp.583-617, 2002.