基于多目标进化的复杂网络社区检测

王 聪,柴争义
(天津工业大学 计算机科学与技术学院,天津 300387)

摘 要:为了准确地发现复杂社区结构,提出一种改进的多目标进化的复杂网络社区检测算法。通过在某一范围内等间距产生多个 p 参数,再将其代入AP聚类算法通过半监督聚类方式确定聚类个数以及产生初始种群,克服传统的通过随机方式产生的初始解聚类效果不稳定的缺点,且用模拟退火(SA)算法对多目标进化算法进行改进提高种群搜索能力,防止寻优过程陷入局部最优解。分别在不同 μ 值下仿真40次,以Footbal足球社交网络、Karate-Club网络和Dolphins网络作为测试案例,与传统多目标进化算法以及基于近邻传播(AP)的多目标算法进行实验对比,结果表明文中提出的多目标进化算法在总体上MNI数值更大,即改进效果明显,因此可应用该算法对复杂网络社区进行更加精确的检测。

关键词:复杂网络社区;多目标进化;近邻传播(AP)聚类;模拟退火(SA)算法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)06-0044-05
doi:10.3969/j.issn.1673-629X.2020.06.009

Complex Network Community Detection Based on Multi-objective Evolution

WANG Cong, CHAI Zheng-yi
(School of Computer Science and Technology, Tianjin University of Technology, Tianjin 300387, China)

Abstract: In order to accurately discover the complex community structure, we propose an improved multi-objective evolutionary complex network community detection algorithm. By generating multiple p-parameters at equal intervals in a certain range, and then substituting them into AP clustering algorithm, the number of clusters is determined and the initial population is generated by semi-supervised clustering, so as to overcome the disadvantages of unstable clustering effect of the initial solution of traditional random method. At the same time, the multi-objective evolutionary algorithm is improved by simulated annealing (SA) algorithm to improve the population searching ability and prevent the optimization process from falling into the local optimal solution. Simulating 40 times under different μ values respectively, and using Footbal, a football social network, Karate-Club network and Dolphins network as test cases, we compare the proposed algorithm with the traditional multi-objective evolutionary algorithm and the neighbor-based propagation multi-objective evolutionary algorithm. It is concluded that the improved multi-objective evolutionary algorithm has a larger MNI value in the whole, that is, the improvement effect is obvious. Therefore, it can be used to detect the complex network community more accurately.

Key words: complex network community; multi-objective evolution; neighbor propagation (AP) clustering; simulated annealing (SA) algorithm

0 引 言

在复杂网络中,被广泛关注的一个重要的拓扑属性就是网络的社区结构[1]。所谓社区结构,就是指整个网络由若干“簇”或“组”构成,这些“簇”或“组”叫做社区,位于每个社区内部的节点相互之间的连接非常稠密,而位于不同社区内的节点相互之间的连接比较稀疏[2-3]。

近几年,复杂网络社区发现作为一个研究热点,已经受到越来越多的关注,也有越来越多的学者参与其中,现如今已经涌现出了许多优秀的社区发现算法,主要有以下几类:

(1)基于节点相似度。

为了准确计算节点的相似性,文献[4]通过分析二分网络中两类节点的结构特征及其对社区密度的影

收稿日期:2019-04-21 **修回日期**:2019-08-22
基金项目:国家自然科学基金(U1504613)
作者简介:王 聪(1995-),女,硕士,研究方向为智能计算、社区检测;柴争义,教授,硕导,研究方向为物联网大数据分析、人工智能(机器学习、智能计算等)、网络空间安全(数据异常检测)。

响,结合了 Salton 指数和改进的 Logistic 函数,定义了一种新的相似度计算方法;文献[5]基于网络节点相似度矩阵,结合改进的 K-means 算法对网络节点进行相似性聚类,实现网络的社区发现;文献[6]提出一种基于多层节点的节点相似度计算方法,该方法既可以有效地计算节点之间的相似度,又可以解决节点相似度相同时的节点合并选择问题。

(2)基于层次化聚类。

文献[7]则引进节点相似度并加以适当改进,然后重新定义模块度函数,提出基于节点相似度的层次化社区发现算法,提高了层次化社区发现的准确性;为了处理大规模复杂网络中的社区结构,文献[8]提出一种基于 SCAN 算法的密度聚类算法——QSCAN 算法。

(3)基于 AP 算法。

当前很多学者将 AP 聚类算法应用于社区检测[9],文献[10]利用网络潜在几何概念对网络的相似性矩阵进行改进,能有效提高 AP 聚类算法的聚类效果;对于复杂度较高的社区网络,则需要利用并行计算处理对数据进行 AP 聚类[11]。

(4)结合算法。

有学者提出 MOEA/D 算法的多目标社区检测算法,且经过案例测试,在解决复杂网络社区检测问题上具有一定优势,且通过 AP 聚类算法对初始解进行优化,再用多目标检测算法进行社区发现的社区发现效果更佳[12-13]。

文中将聚类和启发式的算法相结合,提出一种改进的多目标进化算法来进行复杂网络的社区检测,用半监督的 AP 聚类算法确定初始解的聚类数目以及节点所属类别,克服传统的通过随机方式产生的初始解聚类效果不稳定的缺点,进而采用模拟退火算法改进 MOEA/D 算法完成参数寻优过程,防止算法陷入局部最优,提高算法的全局搜索能力。

1　相关技术

1.1　近邻传播聚类

Affinity Propagation (AP)算法是一种半监督聚类算法,该算法无需制定聚类数目以及聚类中心,将所有样本点作为可疑的聚类中心,算法会在每一次迭代过程中对(responsibility)和(availability)的数据进行更新。该算法的终值条件是可以找到 m 个高质量的聚类中心,且其他样本点都能规划到相应的类别。

该算法的步骤如下:

(1)计算数据集的相似度 S,再对 p 赋值。

(2)计算样本点之间的 responsibility 值。

(3)计算样本点间的 availability 值。

(4)responsibility 及 availability 值更新。

其中 λ 是收敛系数,主要用于调节算法的收敛速度以及迭代过程的稳定性。

终止计算的条件必须满足迭代次数已经超过设定的最大值或者经过若干次的迭代计算后聚类中心已经不再发生改变时,计算终止后得出聚类的中心点和每类所包含的数据,反之返回步骤 2,继续计算。

文中将 AP 聚类算法作为初始解的产生算法,用来实现初始解的半监督聚类。

1.2　模拟退火算法

SA(模拟退火)算法是一种模仿系统中的粒子通过运动逐渐趋向平稳的过程,这个过程可以分为加热、等温和冷却三个阶段。随着迭代的进行,系统的温度逐渐下降,对于优化问题来说其当前解逐渐逼近真实解,因而系统趋向于稳定[14-16]。

文中温度 T 的初始温度 $T_0 = 100$,且降温系数为 0.9,即个体下一状态温度为上一状态温度的 90%,如下式:

$$T_{k+1} = 0.9T_k \tag{1}$$

其中,k 为当前状态,$k+1$ 为下一状态;T_k,T_{k+1} 分别为两次相邻状态的温度。

对于个体的变异操作将加入模拟退火的改进,假设当代种群的温度为 T_k,对于个体 i 来说是否接受其变异后的个体取决于该个体变异后的新个体的适应度是否大于变异前个体的适应度。如果变异后的个体的适应度大于原来个体的适应度,则对其进行变异;反之则以概率 p 接受该个体的变异,其中:

$$p = \mathrm{e}^{(f(x_{k,i,\mathrm{mutatio}}) - f(x_{k,i}))/T_k} \tag{2}$$

其中,p 为在变异后个体的适应度小于当前个体的适应度的情况下的变异概率,k 为当前的迭代种群索引,i 为个体索引,$f(x_{k,i,\mathrm{mutatio}})$ 和 $f(x_{k,i})$ 分别为变异后个体的适应度与当前个体的适应度。

变异原则如下式:

$$\begin{cases} x_{k+1,i} = x_{k,i,\mathrm{mutatio}}, f(x_{k,i,\mathrm{mutatio}}) > f(x_{k,i}) \\ x_{k+1,i} = x_{k,i,\mathrm{mutatio}}, f(x_{k,i,\mathrm{mutatio}}) < f(x_{k,i}) \& \\ \qquad p > \mathrm{rand}(0,1) \\ x_{k+1,i} = x_{k,i}, f(x_{k,i,\mathrm{mutatio}}) < f(x_{k,i}) \& \\ \qquad p < \mathrm{rand}(0,1) \end{cases} \tag{3}$$

其中,rand(0,1)为[0,1]之间的一个随机数。变异后个体的适应度 $f(x_{k,i,\mathrm{mutatio}})$ 大于原来个体的适应度 $f(x_{k,i})$,则对其进行变异;反之只有当变异概率 p 大于随机数 rand(0,1),接受该个体的变异,其余情况下保持个体不变异,对于 $k+1$ 代种群的第 i 个个体依旧沿用上一代的第 i 个个体。

2 算法思想

通过 AP 聚类算法可实现对节点网络的半监督聚类,不用再去人为确定聚类数目以及聚类中心;继而采取模拟退火算法对遗传算法进行改进,因为模拟退火算法的独特之处就在于其在个体进化过程中会满足一定的概率来决定是否接受新个体的产生,这个准则也被称为 Metropolis 准则,这也是模拟退火算法相较于遗传算法的优势,通过一定概率接受新个体来提高算法的全局解空间的搜索能力,防止陷入局部最优。

3 MOEA/D 多目标进化社区发现算法

MOEA/D 是由 Zhang 和 Li 在 2007 年提出的一种基于分解的进化多目标优化算法,该算法将一个多目标问题根据权值矢量分解成个单目标子问题,并通过进化计算方法同时优化这些子问题。

文中采用模块度密度 D 作为指标对 Pareto 解进行优秀个体的选取。

3.1 目标函数及评价指标

为了使网络社区划分的好坏可以有一个明确的评价指标来衡量,根据对复杂网络的研究提出了模块度(modularity)的概念,但利用模块度来优化会导致分辨率限制的问题,于是有研究者引入了模块度密度 D 来避免这个问题。在大量的重复验证中,以模块度密度 D 作为指标要比以模块度 Q 作为指标在复杂网络社区检测的效果更好。

给定图的一个划分 $S=\{V_1,V_2,\cdots,V_k\}$,子图 G_i 的节点集用 V_i 来表示。在这之中,定义 V_1 和 V_2 是两个互相不相交的子集,定义 $L(V_1,V_2)=\sum_{i\in V_1,j\in V_2}A_{ij}$, $L(V_1,\overline{V_2})=\sum_{i\in V_1,j\in\overline{V_2}}A_{ij}$,模块度密度的定义如下:

$$D=\sum_{i=1}^{k}\frac{L(V_i,V_i)-L(V_i,\overline{V_i})}{|V_i|} \tag{4}$$

其中, $L(V_1,V_2)$ 相当于 Ratio Association (RA)[12],表示子图 G_i 节点的内度之和与 G_i 节点数的比值; $L(V_1,\overline{V_2})$ 相当于 Ratio Cut (RC),表示子图 G_i 节点的外度之和与 G_i 节点数的比值。

文中把 RA 改成了 Kernel K-means (KKM),将上述问题转换成了关于最小值的优化问题,则目标函数定义为下式:

$$\min\begin{cases}\mathrm{KKM}=(n-k)\sigma-\sum_{i=1}^{k}\frac{L(V_i,V_i)}{|V_i|}\\ \mathrm{RC}=\sum_{i=1}^{k}\frac{L(V_i,\overline{V_i})}{|V_i|}\end{cases} \tag{5}$$

其中, $n=|V|$, σ 为一实数。这样一来,社团内部的紧密性就可以用 KKM 的值来表示,而内部节点和外部节点连接的稀疏性则可以用 RC 的值来表示。KKM 的值越小,社区内节点间的连接越紧密;RC 的值越大,社区间节点间的连接越稀疏。

3.2 算法构建

文中算法共有三个部分,分别是初始解优化、多目标进化以及结果可视化。

步骤 1 ~ 6 为 AP 算法初始解优化部分,步骤 7 ~ 13 为 AP 算法多目标进化部分。

步骤 1:设置 AP 聚类算法的 p 参数的上下限 $[p_{\min},p_{\max}]$,并且对该范围内进行 N 等分,令 $p=p_{\min}$, $\Delta p=\frac{p_{\max}-p_{\min}}{N-1}$ 。

步骤 2:计算数据集的相似度 S 。

步骤 3:计算样本点之间的 responsibility 值。

$$r(i,k)\leftarrow s(i,k)-\max_{j\neq k}(s(i,j)+a(i,j)) \tag{6}$$

其中, $a(i,j)$ 表示节点 j 对于节点 i 的 availability 值, $s(i,k)$ 表示节点 i 和节点 j 的相似度, $r(i,k)$ 表示节点 k 对于节点 i 的 responsibility 值。

步骤 4:计算样本点间的 availability 值。

$$a(i,k)\leftarrow\min\{0,r(k,k)\sum_{j=i,k}\max(0,r(j,k))\} \tag{7}$$

$$a(k,k)\leftarrow\sum_{j\neq k}\max(0,r(j,k)) \tag{8}$$

其中, $a(i,k)$ 表示节点 k 对于节点 i 的 availability 值, $r(k,k)$ 表示节点 k 的 responsibility 值。

步骤 5:responsibility 及 availability 值更新。

$$\begin{cases}r_{i+1}(i,k)=\lambda r_i(i,k)+(1-\lambda)r_{i+1}^{\mathrm{ok1}},\lambda\in[0,1)\\ r_{i+1}(i,k)=\lambda a_i(i,k)+(1-\lambda)a_{i+1}^{\mathrm{ok1}},\lambda\in[0,1)\end{cases} \tag{9}$$

其中, λ 是收敛系数,主要用于调节算法的收敛速度以及迭代过程的稳定性。

步骤 6:计算个体的适应度并记录。

步骤 7:判断 p 是否等于 $p_{\max}$,如果成立则进入下一步的多目标进化算法,初始寻优过程结束;反之, $p=p+\Delta p$ 返回步骤 2。

步骤 8:个体编码采用直接编码,其中每个个体都包含 m 个十进制的染色体,也就是每个节点所属的社团的类别号,这些类别号都是整数,可以是 1-m 之间的随机整数。对于一个具有 m 个节点的网络图最多可以分为 n 类社区,则在这种情况下个体编码 $[1,2,\cdots,m]$,则该个体具有 m 个染色体,每个染色体代表一个节点的所属社团的类别。

步骤 9:选择。 $f(X_i^1),f(X_i^2),\cdots,f(X_i^m)$ 分别为 X_i^1 , $X_i^2,\cdots,X_i^m$ 的适应度;总适应度 $F=\sum_{j=1}^{m}f(X_i^j)$;则个体 j 被选为当代优秀个体的概率 $P_j=\frac{f(X_i^j)}{F}$;采用轮盘赌

的形式选择优秀个体。也就是说适应度越大则其被选择的概率越大。

步骤 10:交叉。对相邻的两个父代个体采取单点交叉方式,通过产生一个随机的位置点,将该点以后的个体的相应位置的编码进行交换,以此产生两个新的相邻的子代个体。对于相邻两个个体的交叉过程如图 1 所示。

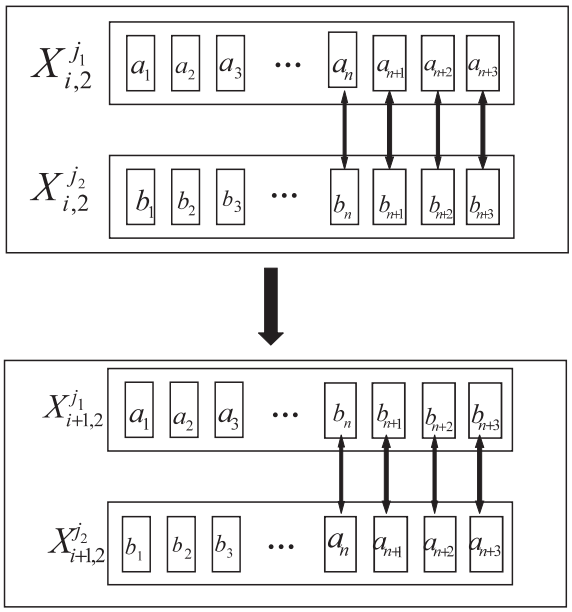

图 1　单点交叉示意图

图中,$X_i^{j_1}$ 和 $X_i^{j_2}$ 分别是第 i 次种群的两个个体 j_1 和 j_2,首先随机选中一个变异位点 n ,然后对位点以后的位置点都进行染色体互换,也就是将两个染色体的 n 节点以后的社区类别进行等位点交换;且以概率 pc 接受是否交叉,进而产生新的下一代相邻的两个个体,也就是两个新的社区划分的方案。

步骤 11:变异。就是对某一节点所属的社团分类以概率 p_m 进行随机改变,但是社团分类小于节点的数目 m ,也就是对于第 i 个种群的第 j 个个体的第 k 个染色体是否变异应满足一下原则:

$$X_{i,2}^{j,k}=\begin{cases}1-\text{randi}(1,m),\text{rand}(0,1)<p_m\\ \text{randi}(1,m),\text{rand}(0,1)>p_m\end{cases}\tag{10}$$

其中,$X_i^{j,k}$ 为第 i 个种群的第 j 个个体的第 k 个染色体的数值,也就是第 k 个节点所属社团类别; randi(1, m) 为 1 到 m 之间的随机整数,但是不等于 $X_i^{j,k}$,该数值也是变异后的节点的归属类别, rand(0,1) 为[0,1]区间内的一个随机数; p_m 为变异概率。

步骤 12:对当代种群进行解码;计算当代各个个体所对应的两个目标函数,KKM 以及 RC 数值;且计算个体的模块度密度 D 对 Pareto 解进行优秀个体的选取。

步骤 13:是否达到预设最大迭代次数,如果是,则输出最后一代种群,算法结束;反之则返回步骤(2)。

结果可视化部分则是通过将算法在不同的 μ 值之下仿真 40 次,分别求均值以此逼近不同 μ 值之下的复杂网络的 NMI 数值,通过画图进行算法效果的对比。

4　实验分析

文中的分析案例为 Footbal 足球社交网络、Karate-Club 网络和 Dolphins 网络,编程平台为 MATLAB2016B,机器配置为 CPUi5-4200h,12 GB 内存。首先,设置 AP 聚类算法的 p 参数的上下限为[-20,0],且对其进行 400 等分,作为初始种群;然后设置 MOEA/D 算法的迭代次数为 40,种群大小为 400,交叉概率为 0.8;接着通过改变 μ 参数从 0.1 到 0.9,间隔为 0.1 分别进行 40 次实验得到实验结果;最后绘制出不同 μ 值下所对应的 NMI 曲线,且以 MOEA/D 算法[12]和 AP-MOEA/D 算法作为对比[13]。

选用 NMI 指标来评价社区检测效果,归一化互信息函数(NMI)用于评价两个社区划分之间的相似性。互信息在信息论中是一种重要的炮测量方法,用来衡量两个事件之间的相关性,其启发式原则是,如果两个社区是相似的,那么通过其中一个社区的类内信息即可得到另外一个社区的结构。使用几何均值的方法归一化互信息的表达式为:

$$\text{NMI}=\frac{I(X;Y)}{\sqrt{H(X)\cdot H(Y)}}\tag{11}$$

利用 NMI 来评价两个社区之间的相似性,当其值接近 1 时,表示两个社区之间结构很相似;当 NMI 的值接近 0 时,表示两个社区之间结构很不相似。

经过 MATLAB 编程运算可以得到 AP-SA-MOEA/D 社区发现算法在不同 μ 值之下的 40 次实验经过剔除异常值的 NMI 平均值曲线,且以 AP-MOEA/D 算法、MOEA/D 算法作为对比算法,对比曲线如图 2 ~ 图 4 所示,对比结果如表 1 ~ 表 3 所示。

表 1　Footbal 网络三种算法 NMI 均值对比

μ	AP-SA-MOEA	AP-MOEA	MOEA
0.10	0.95	0.96	0.96
0.20	0.95	0.96	0.94
0.30	0.95	0.95	0.92
0.40	0.94	0.94	0.90
0.50	0.94	0.93	0.83
0.60	0.94	0.73	0.73
0.70	0.66	0.51	0.58
0.80	0.58	0.44	0.47
0.90	0.33	0.14	0.38

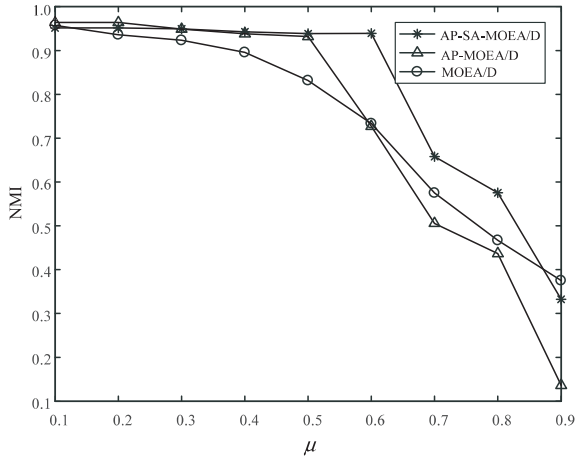

图 2 Footbal 网络三种算法 NMI 对比曲线

表 2 Karate-Club 网络三种算法 NMI 均值对比

μ	AP-SA-MOEA	AP-MOEA	MOEA
0.10	0.91	0.92	0.51
0.20	0.93	0.92	0.50
0.30	0.92	0.91	0.44
0.40	0.95	0.92	0.41
0.50	0.94	0.91	0.36
0.60	0.94	0.90	0.32
0.70	0.59	0.56	0.29
0.80	0.49	0.37	0.27
0.90	0.06	0.07	0.25

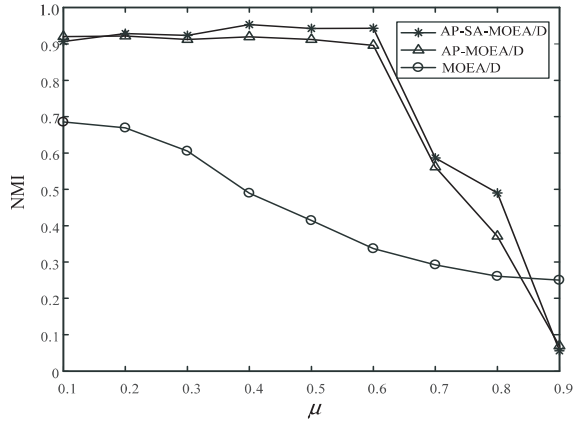

图 3 Karate-Club 网络三种算法 NMI 对比曲线

表 3 Dolphins 网络三种算法 NMI 均值对比

μ	AP-SA-MOEA	AP-MOEA	MOEA
0.10	0.72	0.69	0.69
0.20	0.68	0.67	0.67
0.30	0.68	0.66	0.61
0.40	0.69	0.65	0.49
0.50	0.65	0.63	0.41
0.60	0.66	0.62	0.34
0.70	0.59	0.43	0.29
0.80	0.44	0.36	0.26
0.90	0.29	0.24	0.25

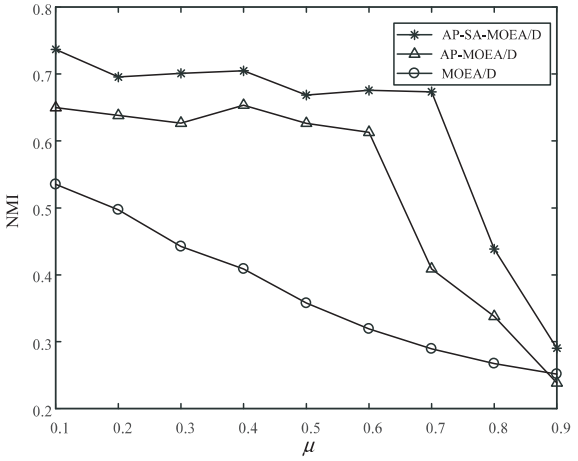

图 4 Dolphins 网络三种算法 NMI 对比曲线

通过对比可以发现,总体上 AP-SA-MOEA/D 算法显然比 AP-MOEA/D 算法、MOEA/D 算法的社区发现效果明显,因而文中提出的算法对于社区发现是有效的。

5 结束语

提出一种改进的复杂社区检测多目标进化算法,首先利用近邻传播(AP)聚类算法半监督产生初始解以及聚类数目,克服传统的通过随机方式产生的初始解聚类效果不稳定的缺点;进而利用模拟退火(SA)算法对 MOEA/D 算法进行改进,提高全局搜索能力;最后通过仿真及算法对比,证明该算法效果更佳。

参考文献:

[1] LYZINSKI V, TANG M, ATHREYA A, et al. Community detection and classification in hierarchical stochastic blockmodels[J]. IEEE Transactions on Network Science & Engineering, 2017, 4(1): 13-26.

[2] RAHIMI S, ABDOLLAHPOURI A, MORADI P. A multi-objective particle swarm optimization algorithm for community detection in complex networks[J]. Swarm and Evolutionary Computation, 2018, 39: 297-309.

[3] MCCALLUM A, WANG X, CORRADA-EMMANUEL A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30: 249-272.

[4] CHEN Dongming, YAN Yanbin, WANG Dongqi, et al. Community detection algorithm based on structural similarity for bipartite networks[C]//2016 7th IEEE international conference on software engineering and service science (ICSESS). Beijing: IEEE, 2016: 98-102.

[5] NI L, MANMAN P, JIANG W, et al. A community detection algorithm based on multi-similarity method[J]. Cluster Computing, 2018(12): 1-10.

[6] 张 虎,吴永科,杨陟卓,等. 基于多层节点相似度的社区

(下转第 108 页)