软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(3):451−464 [doi: 10.3724/SP.J.1001.2012.03996]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

复杂网络簇结构探测——基于随机游走的蚁群算法*

金 弟[1,2], 杨 博[1,2], 刘 杰[1,2], 刘大有[1,2+], 何东晓[1,2]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012)

Ant Colony Optimization Based on Random Walk for Community Detection in Complex Networks

JIN Di[1,2], YANG Bo[1,2], LIU Jie[1,2], LIU Da-You[1,2+], HE Dong-Xiao[1,2]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

+ Corresponding author: E-mail: dyliu@jlu.edu.cn

Jin D, Yang B, Liu J, Liu DY, He DX. Ant colony optimization based on random walk for community detection in complex networks. *Journal of Software*, 2012,23(3):451−464. http://www.jos.org.cn/1000-9825/3996.htm

Abstract: Community structure is one of the most important topological properties in complex networks. The network clustering problem (NCP) refers to the detection of network community structures, and many practical problems can be modeled as NCPs. So far, lots of network clustering algorithms have been proposed. However, further improvements in the clustering accuracy, especially when discovering reasonable community structure without prior knowledge, still constitute an open problem. Building on Markov random walks, the paper addresses this problem with a novel ant colony optimization strategy, named as RWACO, which improves prior results on the NCPs and does not require knowledge of the number of communities present on a given network. The framework of ant colony optimization is taken as the basic framework in the RWACO algorithm. In each iteration, a Markov random walk model is taken as heuristic rule. All of the ants' local solutions are aggregated to a global one through clustering ensemble, which then will be used to update a pheromone matrix. The strategy relies on the progressive strengthening of within-community links and the weakening of between-community links. Gradually, this converges to a solution where the underlying community structure of the complex network will become clearly visible. The performance of algorithm RWACO was tested against a set of benchmark computer-generated networks, and as well on real-world network data sets. Experimental results confirm the validity and improvements of this approach.

Key words: complex network; network clustering; community structure; random walk; ensemble learning; ant colony optimization

摘 要: 网络簇结构是复杂网络最普遍和最重要的拓扑属性之一,网络聚类问题就是要找出给定网络中的所有类

* 基金项目: 国家自然科学基金(60873149, 60973088, 61133011); 教育部博士研究生学术新人奖(450060454018)
收稿时间: 2009-10-23; 修改时间: 2010-07-06; 定稿时间: 2011-01-31

簇.有很多实际应用问题可被建模成网络聚类问题.尽管目前已有许多网络聚类方法被提出,但如何进一步提高聚类精度,特别是在没有先验知识(如网络簇个数)的情况下如何发现合理的网络簇结构,仍是一个未能很好解决的难题.针对该问题,在马尔可夫随机游走思想的启发下,从仿生角度出发提出一种全新的网络聚类算法——基于随机游走的蚁群算法 RWACO.该算法将蚁群算法的框架作为 RWACO 的基本框架,对于每一代,以马尔可夫随机游走模型作为启发式规则;基于集成学习思想,将蚂蚁的局部解融合为全局解,并用其更新信息素矩阵.通过"强化簇内连接,弱化簇间连接"这一进化策略,使网络簇结构逐渐地呈现出来.实验结果表明,对一些典型的计算机生成网络和真实网络,该算法能够较准确地探测出网络的真实类簇数,与一些有代表性的算法相比,具有较高的聚类精度.

关键词: 复杂网络;网络聚类;簇结构;随机游走;集成学习;蚁群算法

中图法分类号: TP18 **文献标识码**: A

现实世界中的诸多系统,如社会网、生物网、Web 网络等,都被称为复杂网络,复杂网络已成为当前最重要的多学科交叉研究领域之一[1−3].与小世界性[1]、无标度性[2]等基本统计特性相并列,网络簇结构是复杂网络最普遍和最重要的拓扑结构属性之一,具有同簇结点相互连接密集、异簇结点相互连接稀疏的特点[3].复杂网络聚类问题就是要揭示出复杂网络中真实存在的网络簇结构.目前已有很多实际应用问题被建模为网络聚类问题,如恐怖组织识别、蛋白质功能预测、新陈代谢途径(pathway)预测、Web 社区挖掘、链接预测等等[4].

复杂网络聚类研究对分析复杂网络的拓扑结构、理解复杂网络的功能、发现复杂网络中的隐藏规律和预测复杂网络行为有着重要的理论意义和广泛的应用前景,它吸引了不同领域的众多研究者.目前,已有许多各具特色的网络聚类算法被提出,按照文献[5]的观点,它们中的大多数可以按照所采用的基本求解策略归纳为两大类:基于优化的方法和启发式方法.

基于优化的方法主要有:FN(fast newman)算法[6]、GA(guimera-amaral)算法[7]、EO(extremal optimization)算法[8]、FUA(fast unfolding algorithm)算法[9]和 ITS(iterated tabu search)算法[10]等.

启发式方法主要有:GN(girvan-newman)算法[3]、CPM(clique percolation method)算法[11]和 LPA(label propagation algorithm)算法[12]及其改进[13,14]等.近来,基于马尔可夫随机游走理论的启发式求解策略也被广泛应用于网络聚类问题.这方面的研究工作主要有:2000 年,Dongen[15]提出了 Markov 聚类算法 MCL,该算法通过在马尔可夫矩阵上施加简单的代数操作,实现了对网络的自然划分;2006 年,Pons 等人[16]首先提出了一种基于随机游走的网络结点间相似性的度量方法,进而给出一个面向大规模网络的凝聚聚类算法;2007 年,Gunes 等人[17]将随机游走思想与 *agent* 方法相结合,用 *agents* 通过随机游走的方式探测网络簇结构信息;2008 年,Rosvall 等人[18]将网络中随机游走的概率流视为真实系统中的信息流,通过压缩概率流描述将网络分解成模块;2008 年,Weinan 等人[19]依据对马尔可夫链(与网络动力学特性相关联)的优化预测,提出了一种网络优化划分策略;2007 年, Yang 等人[20]针对符号网络聚类问题(包括正负权值的网络)提出了基于马尔可夫随机游走模型的启发式符号网络聚类算法 FEC;2008 年,Yang 等人[21]又分析了复杂网络簇结构和马尔可夫随机游走模型动力学特性的内在联系,进而基于大偏差理论提出了分析网络簇结构的谱方法.

尽管目前已提出许多复杂网络聚类方法,但如何进一步提高聚类精度,特别是在未知网络簇个数的情况下如何发现合理的网络簇结构,仍是一个未能很好解决的难题.针对这一问题,本文在文献[20]的基础上提出了一种基于随机游走的蚁群算法(RWACO),用于探测网络簇结构.在该算法中,蚂蚁以随机游走的转移概率作为启发式规则,探测其所在的类簇;每代中所有蚂蚁通过聚类集成[22]产生当前解,并用其更新信息素矩阵;当算法收敛时,可通过分析信息素矩阵得到网络聚类结果.

1 算法 RWACO

1.1 算法主要思想

令 $N=(V,E)$表示一个复杂网络,V 为结点集,E 为边集.设网络 N 的一个 k-划分定义为$\pi=\{N_1,\ldots,N_k\}$,其中,

$N_1,\ldots,N_k$ 满足 $\bigcup_{1\leqslant i\leqslant k} N_i = N$ 和 $\bigcap_{1\leqslant i\leqslant k} N_i = \varnothing$.如果该划分具有同簇结点相互连接密集、异簇结点相互连接稀疏的特性,则称π为网络 N 的一个簇结构.

假设一个 *agent* 沿网络 N 上的边进行随机游走,那么它每走一步之前都要根据转移概率选择下一步所要到达的位置.假使该 *agent* 当前位于结点 i,其下一步到达邻居结点 j 的转移概率为 p_{ij},若网络 N 的邻接矩阵为 $A=(a_{ij})_{n\times n}$,则有

$$p_{ij}=\frac{a_{ij}}{\sum_k a_{ik}} \tag{1}$$

若使用矩阵表示,设 $D=diag(d_1,\ldots,d_n)$,其中,$d_i=\sum_j a_{ij}$ 表示结点 i 的度,则有转移概率矩阵 $P=(p_{ij})_{n\times n}$ 为

$$P=D^{-1}A \tag{2}$$

从马尔可夫随机理论的角度来说,当复杂网络具有簇结构特性时,假使一个 *agent* 从任意结点出发进行若干步随机游走,如果步数适当的话,那么它当前位于簇内任一结点的概率都应大于位于簇外结点的概率,所以算法 RWACO 中,蚂蚁(蚂蚁与 *agent* 是不同的)以随机游走的转移概率作为启发式规则.就是说,每只蚂蚁都是在转移概率和信息素的双重指导之下去寻找自己的解.尽管在每次迭代中,每只蚂蚁所找到的解只代表了该蚂蚁的局部观点,但通过集成方法将所有蚂蚁的解叠加到一起,就会形成一个全局意义上的解,进而利用其更新信息素矩阵,使得信息素矩阵逐步进化并趋于收敛.最终,收敛后的信息素矩阵就代表了算法所有代、所有蚂蚁信息融合的结果,对其进行简单分析即可实现对复杂网络的聚类.

为清晰揭示上述基本思想,给出一个直观的描述:给定某具有簇结构的网络 N,假设若干蚂蚁在该网络上沿着它的边随机爬行;并且蚂蚁具有生命周期,当所有蚂蚁生命结束后就会产生新的下一代蚂蚁群体.在初始阶段,网络 N 上没有信息素,或者信息素很淡,只是由于受网络簇结构的约束,使得蚂蚁在簇内逗留的概率要大于走出该簇的概率,这时的蚂蚁和随机游走的 *agent* 没什么区别;随着前面所有代蚂蚁信息素的积累及挥发,网络簇内边上的信息素越来越浓,簇间边上的信息素越来越淡,使得蚂蚁变得越来越聪明,它们“在簇内游走而不是离开该簇”的趋势也越来越明显;最终,当信息素矩阵收敛时,就自然地得到了网络 N 的聚类结果.概括地说,算法 RWACO 就是通过“强化簇内连接,弱化簇间连接”使网络簇结构逐渐地呈现出来.

此外,由于算法 RWACO 将随机游走的转移概率作为蚂蚁的启发式规则,所以该算法不仅仅局限于无向、无权网络,它对于有向、加权网络也同样适用.

1.2 算法描述

1.2.1 算法 RWACO

算法 RWACO 可分为探测阶段和分裂阶段两部分.探测阶段即通过蚁群算法的执行,收敛后得到其信息素矩阵;分裂阶段即通过对该信息素矩阵的分析,产生网络的最终聚类结果.我们给出探测阶段的基本算法框架,其中,描述一只蚂蚁生成解的算法 *one_ant* 在第 1.2.2 节中将被分成两部分详细介绍.文中算法大都是通过 Matlab 伪代码的形式给出.为便于理解,算法中对 Matlab 的一些特殊符号给出了注释.

Procedure Exploration_Phase

Input: *A,T,S,*ρ /*,*A* 表示网络的邻接矩阵,*T* 为算法迭代次数,*S* 为每代中蚂蚁个数,ρ为蚂蚁信息素矩阵更新率*/;

Output: *B* /*表示算法收敛后的信息素矩阵,即所有蚂蚁信息融合的最终结果*/.

Begin

```
1    B←ones(n,n)*n;   /*初始化信息素矩阵*/
2    for i=1:T
3        solution←zeros(n,n);
4        for j=1:S
5            solution←solution+one_ant(A,B);   /*one_ant 返回一只蚂蚁产生的解*/
```

6 end /*将当前代所有蚂蚁的局部解集成为全局解*/

7 $B \leftarrow \rho * B + solution$; /*更新信息素矩阵*/

8 end

End □

可以看出,该算法每代将当前所有蚂蚁产生的局部解集成为一个全局解,并用其更新信息素矩阵,使信息素矩阵具有更好地指导作用,以生成更好地下一代蚂蚁群体.随着算法的运行,信息素矩阵逐渐进化,使得蚂蚁变得越来越聪明,它们在簇内爬行而不是离开该簇的趋势也越来越明显.最终,收敛后的信息素矩阵就可以被看作是算法中所有代所有蚂蚁信息融合的结果.

如果从集成学习角度来看,RWACO 也可被视为一个聚类集成算法.由于没有信息素的指导,该算法第一代中所有蚂蚁的解完全是由马尔可夫随机游走方法产生的,我们可以将其视为最初的待集成聚类结果.在蚁群算法框架下,算法每代都将当前待集成聚类结果进行集成,然后用其更新信息素矩阵,使信息素矩阵具有更好的指导作用,而在新的信息素矩阵指导下又可以生成更好的下一代待集成的聚类结果,这是一个互相促进的过程.从这一角度来分析,RWACO 的执行过程就是一个带有进化策略的聚类集成过程,它通过蚁群算法框架将不够精确的初始聚类结果逐步集成为一个高精度的聚类结果.

下一步考虑如何对收敛后的信息素矩阵进行分析,从而得到网络的聚类结果.由于蚁群算法的收敛特性,信息素矩阵的簇结构特征是非常明显的,所以文中采用了简单的分裂方法对其进行聚类.对算法 RWACO 的分裂阶段简单描述如下:

Procedure Divide_Phase

Input: B /*算法收敛后的信息素矩阵*/;

Output: C /*网络的聚类结果*/.

Begin

1 设置分裂界限值ε; /*ε为一个很小的正数,文中取 10^{-2}*/

2 取 B 的第 1 行,将值大于ε的网络结点作为一个簇;

3 将 B 中所有与该簇中结点相对应的行列删除,如果 B 不为空,GOTO step 2;

4 返回由所有簇构成的网络聚类结果 C;

End □

可以如此简单地构建分裂阶段算法,是与探测阶段算法的收敛特性分不开的.因为蚁群算法在探测阶段收敛,所以信息素矩阵的簇结构特性是非常明显的,并且该矩阵中所有同簇结点对应的行列都是基本相同的,只需取任意行,并用一个较小的正数ε对其进行分裂则可得到一个簇.详细分析见第 2.3 节.

1.2.2 一只蚂蚁产生的解

在蚁群算法中,一只蚂蚁产生问题的一个解.所以对网络聚类问题,蚂蚁产生的解就应该是复杂网络的一个聚类结果.

假设一只蚂蚁在网络 N 上爬行.与随机游走的 *agent* 不同,它每走一步之前不仅要考虑转移概率,还要受到信息素的指导.若网络 N 的邻接矩阵为 $A=(a_{ij})_{n\times n}$,信息素矩阵为 $B=(b_{ij})_{n\times n}$,假设该蚂蚁当前位于结点 i,它下一步访问结点 j 的概率 m_{ij} 则可由公式(3)表示,那么该蚂蚁的访问概率矩阵即可表示为 $M=(m_{ij})_{n\times n}$.值得指出的是,这里的访问概率矩阵 M 与第 1.1 节的转移概率矩阵 P 是不同的,矩阵 M 考虑到了信息素带来的影响,所以它所表现出的簇结构特征会变得越来越清晰.

$$m_{ij}=\frac{a_{ij}b_{ij}}{\sum_r a_{ir}b_{ir}} \tag{3}$$

设任意蚂蚁出发点位置为 s,其爬行的步数为 l,V_s^t 表示该蚂蚁的 $t(t\leqslant l)$步访问概率分布向量,其中,$V_s^t(j)$ 表示蚂蚁从结点 s 出发,经过 t 步爬行后,当前位于结点 j 的概率.很显然,$V_s^0=(0,\ldots,0,1,0,\ldots,0)$,其中,$V_s^0(s)=1$.在矩阵 M 的指导下,V_s^t 即可表示为

$$V_s^t = V_s^{t-1}M \tag{4}$$

由于该蚂蚁是以转移概率作为启发式规则,并在信息素的指导下,在具有簇结构的网络 N 上爬行,如果步数 l 适当,其爬行 l 步之后,当前位于簇内任意结点的概率都应该大于位于簇外任意结点的概率.所以,V_s^l 应近似满足公式(5),其中,C_s 表示起始结点 s 所在的簇.对蚂蚁爬行步数 l 的详细分析见后文.

$$\underset{i\in C_s}{\forall}\ \underset{j\notin C_s}{\forall}\{V_s^l(i) > V_s^l(j)\} \tag{5}$$

考虑到经过任意 t 步后,蚂蚁曾到达过它的起始位置,所以我们在每步操作后将 $V_s^t(s)$ 设置成 1,从而增加了一些对“已到达过”而不是“当前位于”的考虑.实验结果表明,该操作是很有效的.另外,考虑到对 V_s^l 的计算除受到网络簇结构(转移概率)和蚂蚁信息素的正面影响之外,还受到结点度分布不平衡的负面影响.也就是说,如果某结点的度很大,即使它不和起始结点 s 在同一簇内,它的 l 步访问概率值也可能很大.我们采用如下方法来消除这一不良影响:设 $d_i' = \sum_j a_{ij}b_{ij}$,它表示考虑到信息素之后结点 i 的度,那么就可以按照公式(6)对 V_s^l 进行修正.这样计算出的 V_s^l 就可以避免复杂网络中结点度的幂率分布所带来的影响,从而能够更好地满足公式(5).

$$V_s^l(i) = \frac{V_s^l(i)}{d_i'} \tag{6}$$

在上述马尔可夫随机游走理论的基础上,给出计算任意蚂蚁的 l 步访问概率分布向量 V_s^l 的算法,如下:

Procedure Produce_V

Input: s /*蚂蚁出发点位置*/,
A /*网络的邻接矩阵*/,
B /*当前代的信息素矩阵*/,
l /*蚂蚁爬行的步数*/;
Output: V /*该蚂蚁的 l 步访问概率分布向量*/.

```
Begin
1  V←zeros(1,n);
2  M←A.*B;
3  Dv←sum(M,1);
4  Ds←diag(Dv);
5  M←inv(Ds)*M;   /*inv 表示矩阵取逆运算*/
6  for i=1:l
7     V(s)←1;
8     V←V*M;
9  end
10 V←V./Dv;
End                                                    □
```

文中计算蚂蚁 l 步访问概率分布向量 V_s^l 的方法借鉴了文献[20]中算法 FEC 的主要思想,但它们有着本质的不同.首先,FEC 算法中使用了目的结点来计算概率向量 V_s^l,但该方法作为蚂蚁的启发式规则是不合适的,所以文中使用了起始结点来计算该向量.值得指出的是,这两种计算方法以及它们所计算出的概率向量 V_s^l 代表的意义都是完全不同的.另外,我们在计算概率向量 V_s^l 时综合考虑了网络拓扑结构和当前信息素两种因素,并且随着算法迭代次数的增加,计算出的 V_s^l 会越来越好地满足公式(5);而算法 FEC 只考虑到了网络拓扑结构信息,所以它得到的 V_s^l 只能近似满足公式(5).

在计算出 V_s^l 后,接下来的问题是寻找蚂蚁的解,也就是要找到复杂网络的一个聚类结果.但一只蚂蚁只能说明它会以高概率访问其所在类簇内的结点,并不能说明访问概率低的结点就一定在另一簇内.所以,一只蚂蚁

只能以其局部观点找到它所在的簇,而并不需要考虑其余结点的分类情况.

为找出该蚂蚁所在的簇,我们采用如下方法:首先,将网络中所有结点按照 V_s^l 降序排序,得到结点序列 *ix*,并找出该蚂蚁的起始结点 *s* 在 *ix* 中所占的位置 *cut*1.因为该蚂蚁所在的簇必然会包含其起始结点,所以 *ix* 的簇内簇外分裂点位置不会小于 *cut*1;然后,计算出 *ix* 中相邻两结点的访问概率值之差,找出使该差值最大的位置 *cut*2,即认为访问概率值下降最快的位置应更有理由作为 *ix* 的簇内簇外分裂点;最后,取上述两步的最大值作为结点序列 *ix* 的分裂点,由 *ix* 中在该分裂点之前的所有结点构成此蚂蚁所在的簇.但上述分裂方法只考虑了概率向量 V_s^l 的信息,而未充分考虑网络自身的拓扑结构的信息.

2004 年,Newman 等人[23]针对于网络聚类问题提出了网络模块性评价函数(又称 *Q* 函数).该度量标准目前已被大多数研究者广泛接受.但如果用 *Q* 函数作为度量来分裂结点序列 *ix* 是不可行的,因为它会倾向于将所有网络结点均分为两大类,而无法精确地找到该蚂蚁所在的簇.文中采用如下方法:首先,通过上段中的方法找到分裂点所在的局部区间[*cut*1,*cut*2].如果 *cut*1>*cut*2,则将 *cut*1 作为结点序列 *ix* 的最终分裂点;否则,在该局部区间内找出使得 *Q* 函数值最大的位置作为 *ix* 的最终分裂点,由 *ix* 中在该分裂点之前的所有结点构成此蚂蚁所在的簇.这种方法不仅充分考虑了访问概率向量 V_s^l 的信息,还考虑了网络自身拓扑结构的信息,所以其找到的簇是很准确的.图 1 针对随机网络 *RN*(4,32,16,8)给出了一个直观的描述,图 1(a)和图 1(b)分别代表了算法运行过程中的两种典型情况.其中,三角形表示结点序列 *ix* 的最终分裂点位置.

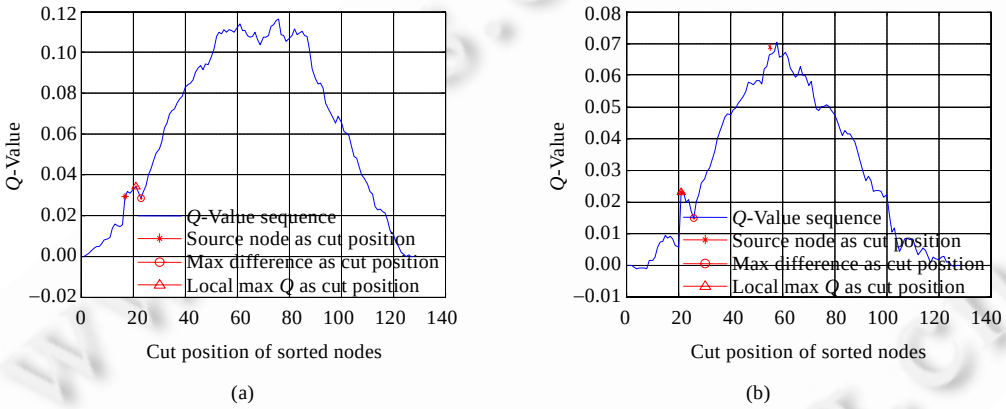


Fig.1 Two typical scenes of computing cut position of *ix* against random network *RN*(4,32,16,8)

图 1 针对随机网络 *RN*(4,32,16,8),计算结点序列 *ix* 分裂点的两种典型情况

计算出 V_s^l 后,给出寻找该蚂蚁的解的算法如下:

Procedure Divide_V

Input: *V* /*蚂蚁的 *l* 步访问概率分布向量*/,

S /*蚂蚁出发点位置*/;

Output: *solution* /*该蚂蚁的解*/.

Begin

1 [*sorted_V*,*ix*]←*sort*(*V*,'*descend*');

2 *cut_pos*1←*find*(*ix*=*s*);

3 *diff_V*←−*diff*(*sorted_V*);

4 [*temp*,*cut_pos*2]←max(*diff_V*);

5 *cut_pos*←max_*Q*(*cut_pos*1:*cut_pos*2);

6 *cluster*←*ix*(1:*cut_pos*);

```
7    solution←zeros(n,n);
8    solution(cluster,cluster)←1;
9    I←eye(cut_pos,cut_pos);   /*生成单位矩阵*/
10   solution(cluster,cluster)←solution(cluster,cluster)−I;
End                                                          □
```

1.3 算法参数设置

算法 RWACO 有迭代次数 T、蚁群规模 S、信息素矩阵更新率ρ、矩阵分割界限值ε和蚂蚁爬行的步数 l 等 5 个参数.其中,前 3 个参数很容易确定,文中将其分别设置为:T 取 20,S 取网络结点数 n,ρ取 0.6.由于在探测阶段蚁群算法(信息素矩阵)收敛,所以对参数ε只需取一个较小的正数即可实现对信息素矩阵的分割,文中取$\varepsilon=10^{-2}$,对该参数的详细分析参见第 2.3 节.然而,我们对参数 l 的设置却是非常困难的,因为它需要保证计算出的 l 步访问概率分布向量 V_s^l 尽量满足公式(5).也就是说,在按照 V_s^l 进行降序排序的结点序列 ix 中,与起始结点 s 在同一簇内的结点应尽可能地排在最前面.我们研究发现,当结点序列 ix 收敛时(不再发生变化),它就能够很好地满足上述要求.因此,本文通过判断结点序列 ix 的收敛情况来确定参数 l 的取值.即如果 ix 不再发生变化,那么蚂蚁爬行结束.实验结果表明,结点序列 ix 很快就会收敛,其一般不会超过 50 步.

1.4 时间复杂度分析

为了便于理解,文中算法均采用矩阵方法进行描述.但复杂网络一般为稀疏图,而算法 RWACO 中的多数操作并不是必须用到矩阵操作,所以为了降低时间复杂度,该算法实际上是采用链表形式来实现的.不妨设网络中的结点数为 n,边数为 m.下面对算法 RWACO 的时间复杂度进行分析.

首先,我们讨论一只蚂蚁生成解的复杂性.其中,Produce_V 中时间复杂度最高的步骤为第 8 步.由于矩阵 M 为稀疏矩阵,所以如果采用链表表示,其复杂性为 $O(lm)$.Divide_V 中时间复杂度最高的步骤为第 1 步(涉及排序操作),其复杂性为 $O(n\log(n))$,所以一只蚂蚁生成解的时间复杂度应为 $O(n\log(n)+lm)$.

在上述分析的基础上可知,如果采用链表来实现,Exploration_Phase 中时间复杂度最高的步骤应为第 5 步,复杂性为 $O(TS(n\log(n)+lm))$.由于文中参数分别设置为 $T=20,S=n,l\leqslant 50$,又由于一般复杂网络都为稀疏图(即 $m=k\times n$,其中,k 为常数),所以 Exploration_Phase 的复杂性也可表示为 $O(n^2\log(n))$.在实现过程中,算法 Divide_Phase 并不需要真的对 B 矩阵的行列进行删除,而只需将所有待删除的行列值全部设置成 0 就可以了.此外,该算法至多需要对长度为 n 的行向量分割 C 次,其中,C 为算法得到的类簇数.所以,Divide_V 的复杂性为 $O(n^2+Cn)$.最终,算法 RWACO 的复杂性即可表示为 $O(n^2\log(n))$.

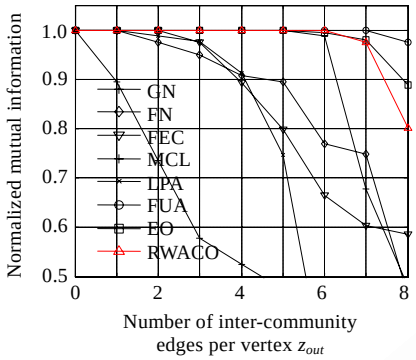
2 实 验

为了定量地分析算法 RWACO 的性能,我们利用人工生成网络和真实世界网络对文中算法进行测试,并给出参数分析.算法实验环境为:处理器 Intel(R)Core(TM)2 4400 2.0GHz,内存 2G,硬盘 160G,操作系统为 Windows XP,编程语言为 Matlab 7.3.

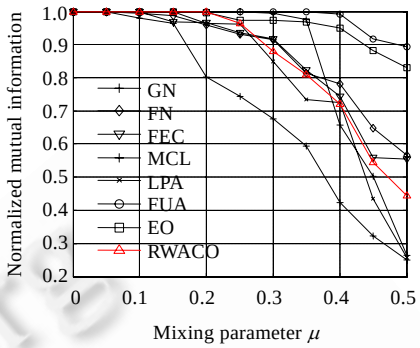
2.1 计算机生成的网络

文中采用已知簇结构的随机网络(Newman 模型)测试算法 RWACO 的聚类精度,该实验方法被相关工作广泛采用,已成为测试复杂网络聚类算法准确性的一种基准方法[3].已知簇结构的随机网络定义为 $RN(C,s,d,z_{out})$,其中,C 表示网络簇的个数,s 表示每个簇包含结点的个数,d 表示网络中每个结点的度,z_{out} 表示每个结点与簇外结点的连接数.可以看出,随着 z_{out} 的增大,网络簇结构越来越模糊,同时也给网络聚类算法带来了越来越大的挑战.特别地,当 $z_{out}>8$ 时,认为该随机网络不具有簇结构.由于不同算法得到的类簇数是不同的,未必等于网络的真实类簇数.有的算法倾向于将真实的网络簇进一步细分,有的算法会将若干真实网络簇分为一类.文献[24]认为,在该情况下,基于信息理论的精度度量标准 Normalized Mutual Information(NMI)较其他精度度量标准更加公平

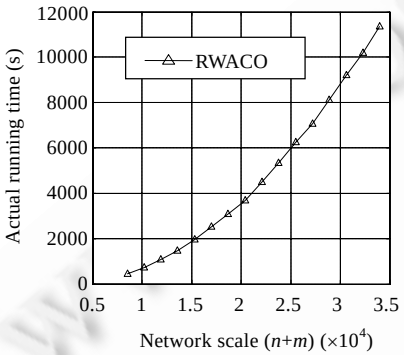
合理.所以,文中采用NMI对不同算法的聚类精度进行评估.按照该精度度量标准,我们将文中算法RWACO与当前一些具有代表性的优秀算法 GN[3],FN[6],FEC[20],MCL[15],LPA[12],EO[8],FUA[9]进行了比较.图 2(a)给出了实验结果,这里所采用的随机网络是被普遍采用的基准随机网络 $RN(4,32,16,z_{out})$.图中 y 轴表示聚类精度,x 轴表示 z_{out},曲线上的每个数据点是采用不同算法聚类 50 个随机网络得到的平均准确率.可以看出,文中算法 RWACO 的聚类精度稍逊于算法 EO 和 FUA,而优于其他 5 种算法.



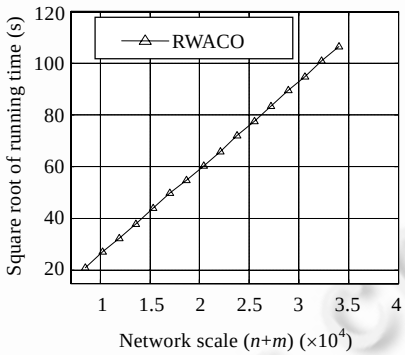
(a) 采用 Newman 模型 $RN(4,32,16,z_{out})$,算法 RWACO 与算法 GN,FN,FEC,MCL,LPA,FUA,EO 聚类精度的比较



(b) 采用新的异质网络模型 $HRN(200,15,2,1,\mu)$,算法 RWACO 与算法 GN,FN,FEC,MCL,LPA,FUA,EO 聚类精度的比较



(c) 采用 Newman 模型 $RN(C,100,16,5)$,算法 RWACO 的运行时间随网络规模的变化趋势



(d) 采用 Newman 模型 $RN(C,100,16,5)$,算法 RWACO 运行时间的平方根随网络规模的变化趋势

Fig.2 Testing performance of RWACO against random networks

图 2 采用随机网络测试算法 RWACO 的性能

为了进一步测试算法 RWACO 的性能,我们采用了一种更加新颖的网络模型[25],该模型可以产生异质网络结构,同时也能够比 Newman 模型[3]更好地验证分辨极限(resolution limition)[26]问题.该网络模型可被定义为 $HRN(n,k,\gamma,\beta,\mu)$,其中,$n$ 表示网络中的结点数,k 表示结点的平均度,γ表示结点度的幂率分布系数,β表示社区规模的幂率分布系数,μ表示任一结点和簇外结点的连接数占该结点度的百分比(也被称为混合参数).可以看出,随着μ的增大,网络簇结构越来越模糊,同时也给网络聚类算法带来了越来越大的挑战.特别地,当μ>0.5 时,认为该随机网络不具有簇结构.图 2(b)给出了实验结果,这里所采用的随机网络是异构随机网络 $HRN(200,15,2,1,\mu)$.图中 y 轴表示聚类精度,x 轴表示μ,曲线上的每个数据点是采用不同算法聚类 50 个随机网络得到的平均准确率.可以看出,新的异质网络模型较经典的 Newman 模型更加难以聚类,文中算法 RWACO 在该异质网络上的聚类精度与那些启发式算法相当,但是要逊于基于优化的算法 EO 和 FUA.

计算速度是另一个评价聚类算法性能的重要指标.第 1.4 节已给出了算法 RWACO 的时间复杂度为 $O(n^2\log(n))$,本节从实验角度出发来评价该算法的运行效率.图 2(c)给出了文中算法实际运行的时间随网络规模

变化的趋势,图 2(d)给出了该算法实际运行时间的平方根随网络规模变化的趋势.实验中采用了随机网络 $RN(C,100,16,5)$进行测试,该网络的类簇结构确定,但其网络簇的个数可由 C 值调节,共包括 $100C$ 个网络结点,$800C$ 条网络连接.图 2(d)中,y 轴表示算法实际运行时间的平方根,x 轴表示网络规模(结点数+连接数).可以看出,由于 $\log(n)$对算法运行效率的实际影响较小,所以该算法的运行时间与网络规模的平方近似成正比.这也进一步验证了算法 RWACO 的时间复杂度为 $O(n^2\log(n))$的正确性.

2.2 真实世界网络

由于真实网络与计算机生成网络具有不同的拓扑特性,我们通过 3 个已知簇结构的真实网络来进一步测试文中算法的性能.由于算法 RWACO 是一个确定性的聚类算法,其每次运行结果都是相同的,所以我们可以对每个网络只进行一次实验.表 1 给出了算法 RWACO 与算法 GN[3],FN[6],FEC[20],MCL[15],LPA[12],EO[8],FUA[9]进行比较的实验结果.表 1 中除使用 NMI 作为精度度量标准外,还使用了 Newman 等人经常采用的 Fraction of Vertices Identified Correctly(FVIC)[6]作为辅助的精度度量标准.这是由于除 RWACO 以外的其他算法通常会对网络的真实簇结构进行不合理的细分,而 FVIC 方法是将算法在每个真实网络簇中找到的最大模块视为正确划分,因此它把对网络真实簇结构的不合理细化也看作是正确分类.所以说,该度量标准要偏向于那些对真实网络簇结构进行不合理细分的网络聚类算法.如果算法 RWACO 的 FVCI 精度都要高于这些算法,这更能说明文中算法的有效性.

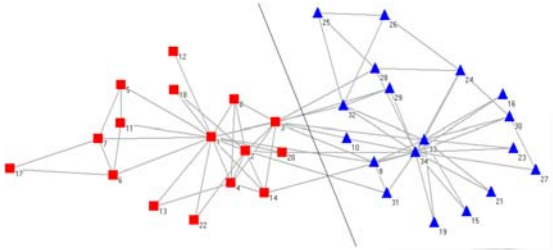
第 1 个为空手道俱乐部网络(Zachary's karate club)[27],该网络是 Zachary 通过对一个美国大学空手道俱乐部历时两年的观测而构建的.它以俱乐部中的 34 个成员作为结点,如果两个成员之间存在友谊关系,那么它们对应的顶点之间就会有一条边相连.后来由于意见分歧,该俱乐部最终分裂为分别以管理者和教练为核心的两个新俱乐部.由表 1 和图 3(a)可以看出,算法 RWACO 与 MCL 完全正确地将该网络分为两类.而其他 6 种算法会对该网络进行不合理的细分.即使不考虑细分所带来的影响(使用 FVCI 作为精度度量标准),其他算法仍然无法对该网络进行完全正确的分类.第 2 个为海豚网络(dolphin social network)[28],该网络是 Lusseau 对居住在新西兰神奇湾的 62 个宽吻海豚历时 7 年的观察建立的.每个海豚代表一个顶点,如果两个海豚间联系频繁,那么它们对应的顶点之间就会有一条边相连.这些海豚天然地分为雄性海豚和雌性海豚两大类.由表 1 和图 3(b)可以看出,算法 RWACO 得到的 NMI 精度是最高的,它近乎完美地将这些海豚分为两类,而仅分错了一只名为 sn89 的海豚,这或许是由于该海豚和两个簇都仅有一条边连接的缘故.而其他 7 种算法均将该网络分为了更多的簇(>2),尤其是算法 GN 和 MCL,都将该网络分为了 13 个类.即便是不考虑细分所带来的影响(使用 FVCI 作为精度度量标准),算法 RWACO 的聚类精度也仅略低于算法 MCL 和 FUA.第 3 个为足球联盟网络(American college football)[3],该网络是 Newman 根据 2000 年秋季常规赛季的比赛计划构建的.网络中的结点代表球队,边权代表两个球队在该赛季的比赛次数,它共包含 115 个结点和 616 条边.这些球队被分为 12 个联合会(conference),一般来说,同一个联合会的球队之间的比赛次数要多于不同联合会的球队间的比赛次数,所以每个联合会代表一个真实的网络簇.由表 1 和图 3(c)可以看出,算法 RWACO 将该网络分为了 12 类,除 Sunbelt 和 IA Independents 两个联合会之外,其他 10 个联合会中的球队几乎完全被正确分类.Sunbelt 中,7 只球队被分为两类,文献[3]认为这是合理的划分.IA Independents 中的 5 支球队被分布在 3 个类簇中,这主要是由于这些球队是独立球队,它们与其他联合会中球队的比赛要比与联合会内球队的比赛还要多[3].无论是采用 NMI 还是 FVCI 作为度量标准,算法 RWACO 的聚类精度都仅稍逊于算法 MCL,而要优于其他 6 种算法.图 3 中,相同灰度级相同形状的结点属于同一个真实类簇,而每一个结点堆是算法 RWACO 得到的一个簇

从整体来看,针对这 3 个已知社区结构的真实网络,算法 RWACO 与 MCL 得到的聚类精度相当,而优于其他 6 种算法.同时,算法 RWACO 还可以较准确地获得网络的真实类簇数,无论是对于类簇数少的情况(2 分类)还是对于类簇数多的情况(12 分类)都是如此.这都说明了本算法的有效性.

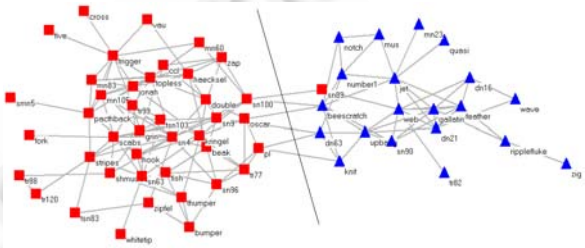
Table 1 Comparing RWACO with GN, FN, FEC, MCL, LPA, EO, and FUA on three real-world networks with known community structure

表 1 对 3 个簇结构已知的真实网络,算法 RWACO 与算法 GN,FN,FEC,MCL,LPA,EO,FUA 的比较

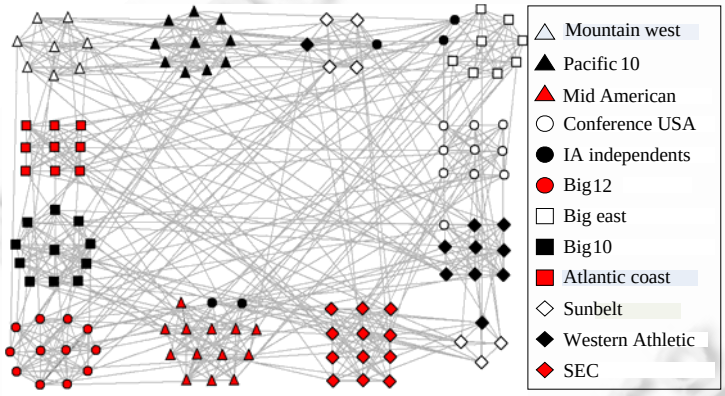
Algorithms	Karate network			Dolphin network			Football network		
	NMI (%)	FVCC (%)	Cluster number	NMI (%)	FVCC (%)	Cluster number	NMI (%)	FVCC (%)	Cluster number
GN	57.98	97.06	5	44.17	98.39	13	87.89	83.48	10
FN	69.25	97.06	3	50.89	96.77	5	75.71	63.48	7
FEC	69.49	97.06	3	52.93	96.77	4	80.27	77.39	9
RWACO	100	100	2	88.88	98.39	2	92.69	93.04	12
MCL	100	100	2	42.39	100	13	93.45	95.65	16
LPA	67.75	96.47	3.78	52.29	97.06	6.5	89.20	87.51	11.22
FUA	58.66	97.06	4	63.63	100	4	89.03	86.96	10
EO	58.66	97.06	4	57.92	98.39	4	88.49	86.26	10



(a) RWACO 对空手道网络的聚类结果



(b) RWACO 对海豚网络的聚类结果



(c) RWACO 对足球联盟网络的聚类结果

Fig.3 Clustering results of RWACO against three real-world networks with known community structure

图 3 算法 RWACO 对 3 个簇结构已知的真实网络之聚类结果

由于目前网络模块性函数 Q 作为评价社区结构优劣的量化标准已被多数研究者所广泛接受,我们再以 Q 函数作为评价标准,采用 5 个真实网络作为测试数据集,将文中算法与其他 7 种算法进行比较.这 5 个真实网络中除包含上面的 3 个网络之外,还包含两个簇结构未知的网络.其中:Krebs 美国政治书网络(Krebs polbooks network)[29]是由 Krebs 建立的,网络中的结点代表亚马逊网上书店(Amazon.com)卖出的有关美国政治的图书,网络中的边将购买者经常一起购买的书连接起来,该网络共包含 105 个结点和 441 条边;爵士乐队协作网(Jazz musicians network)[30]是 Gleiser 和 Danon 根据一个爵士乐唱片的数据库建立的乐队协作网,网络包含 198 个乐队,每个乐队用一个结点表示.如果两个乐队有一个或多个共同乐师的话,则在代表两个乐队的结点之间添加一条边,网络中总共有 5 484 条边.表 2 给出了实验结果.可以看出,以 Q 函数作为目标函数的算法 FUA,EO 所获得的 Q 函数值要明显优于启发式网络聚类算法 RWACO,MCL,FEC,LPA 所获得的 Q 函数值.

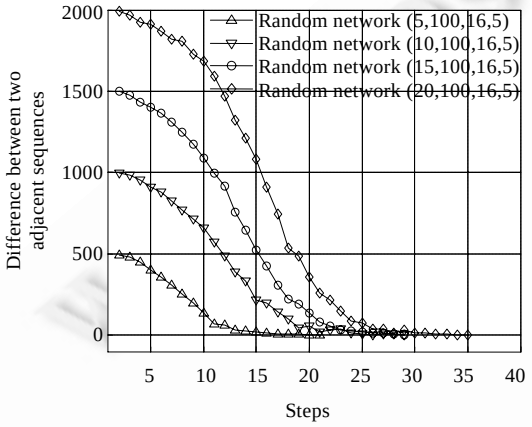
Table 2 Comparing RWACO with GN, FN, FEC, MCL, LPA, EO, and FUA on five real-world networks in term of function Q

表 2 以 Q 函数作为度量标准,针对 5 个真实网络,算法 RWACO 与算法 GN,FN,FEC,MCL,LPA,EO,FUA 的比较

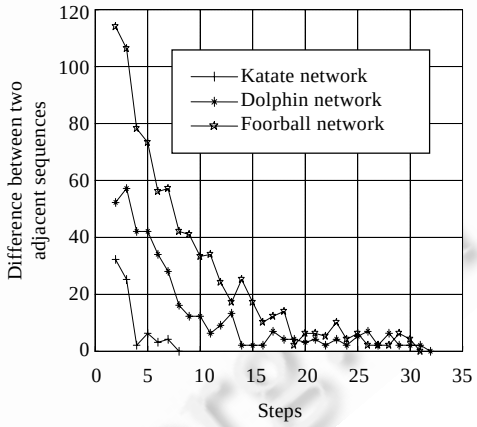
Q-Value	Karate network	Dolphin network	Polbooks network	Football network	Jazz network
GN	0.401 3	0.470 6	0.516 8	0.599 6	0.405 1
FN	0.252 8	0.371 5	0.502 0	0.454 9	0.403 0
FEC	0.374 4	0.497 6	0.490 4	0.569 7	0.444 0
RWACO	0.371 5	0.377 4	0.456 9	0.601 0	0.413 3
MCL	0.371 5	0.439 0	0.508 9	0.582 8	0
LPA	0.370 5	0.480 6	0.504 2	0.588 4	0.368 4
FUA	0.418 8	0.526 8	0.498 6	0.604 6	0.443 1
EO	0.418 8	0.526 9	0.526 2	0.601 5	0.436 7

2.3 参数分析

参数 l 是 RWACO 算法的重点,第 1.3 节已给出了参数 l 的取值方法,本节再从实验的角度对其进行详细分析.图 4 中,x 轴表示蚂蚁爬行步数 l 的取值,y 轴表示两相邻时刻所对应的结点序列(按照 V_s^l 降序排序)的差异.若 l 时刻的结点序列为 p,l–1 时刻的结点序列为 q,那么 $y(l)=nnz(p-q)$,其中,$nnz(v)$为返回向量 v 中非零元素的个数.由图 4 可以看出,对于不同规模的网络,不管是随机网络还是真实网络,其结点序列随步数 l 的收敛速度都是很快的,即使对于具有 2 000 个结点 16 000 条边的网络也可以在 35 步内收敛.



(a) 通过不同规模的随机网络分析参数 l 的敏感性

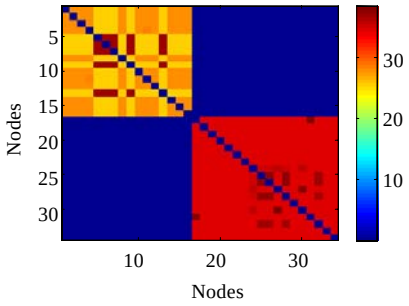


(b) 通过 3 个真实网络分析参数 l 的敏感性

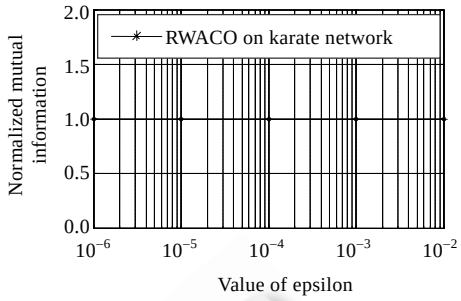
Fig.4 Sensitivity analysis of ant's step number l

图 4 对蚂蚁爬行步数 l 的敏感性分析

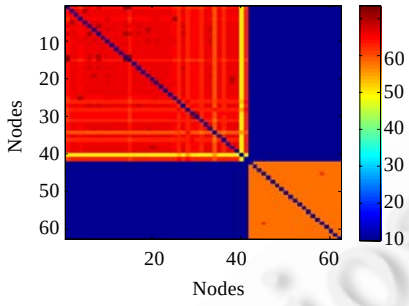
参数ε是 RWACO 算法的另一个重要参数,其设置可能会对 RWACO 的聚类精度产生较大影响.在此,我们第 2.2 节中 3 个已知簇结构的真实网络为例,对参数ε的鲁棒性进行分析.由于参数ε应为一个很小的正数,文中将其分别设置为 10^{-6},10^{-5},10^{-4},10^{-3},10^{-2},并在上面 3 个真实网络上进行实验,结果如图 5 所示.其中,左侧的 3 个图(图 5(a)、图 5(c)、图 5(e))分别为算法 Exploration_Phase 针对这 3 个网络得到的信息素矩阵(按网络的真实社区结构排序,使得同簇结点相邻),右侧的 3 个图(图 5(b)、图 5(d)、图 5(f))分别为算法 Divide_Phase 对 3 个信息素矩阵采用ε进行分割得到的聚类精度.可以看出,算法 Exploration_Phase 得到的信息素矩阵均具有很好的收敛特性,算法 Divide_Phase 对分割参数ε的取值也不敏感.



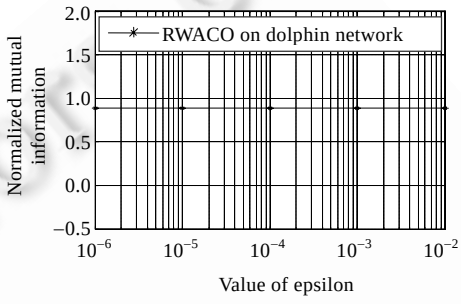
(a) 对于空手道网络,算法 Exploration_Phase 得到的信息素矩阵



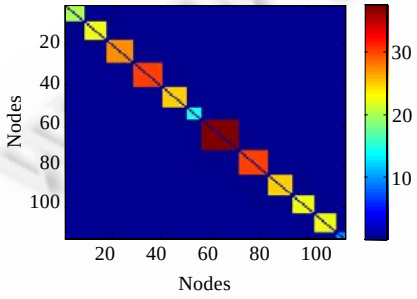
(b) 对于空手道网络,矩阵分割参数ε在区间[10^{-6},10^{-2}]中变化时的聚类精度



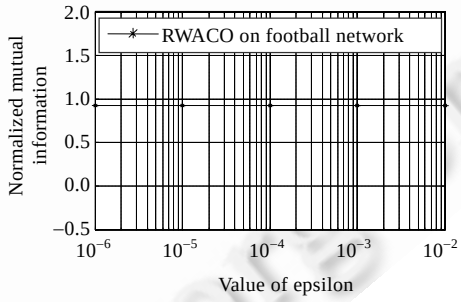
(c) 对于海豚网络,算法 Exploration_Phase 得到的信息素矩阵



(d) 对于海豚网络,矩阵分割参数ε在区间[10^{-6},10^{-2}]中变化时的聚类精度



(e) 对于足球联盟网络,算法 Exploration_Phase 得到的信息素矩阵



(f) 对于足球联盟网络,矩阵分割参数ε在区间[10^{-6},10^{-2}]中变化时的聚类精度

Fig.5 Sensitivity analysis of cut parameter ε

图 5 对矩阵分割参数ε的敏感性分析

3 结 论

本文提出一种蚁群算法来探测网络簇结构.在该算法的每一代中,每只蚂蚁以随机游走模型作为启发式规则,并在信息素的指导下寻找它所在的类簇;然后,基于集成学习思想将所有蚂蚁找到的局部解(它所在的类簇)融合为一个全局解,并用其更新信息素矩阵,使信息素矩阵具有更好的指导作用;在启发式规则及新的信息素矩阵的指导下,又可产生更好的下一代蚂蚁群体.这是一个双向促进的进化过程.随着蚁群算法迭代次数的增加,网络簇内边上的信息素越来越浓,簇间边上的信息素越来越淡,算法收敛后即得到网络聚类结果.总之,该算法就是通过“强化簇内连接,弱化簇间连接”使网络簇结构自然地呈现出来.

在计算机生成的网络和已知簇结构的真实网络上进行测试,文中算法 RWACO 与其他算法相比都表现出了较高的聚类精度.然而,该算法的运行效率不是很高,使其仅适用于对聚类精度要求较高的中小型网络分析任务.此外,对于包含细粒度社区的网络,该算法可能倾向于发现较大的社团,而无法得到更细微的社区划分结构.因此,我们以后的工作主要集中在:1) 分析并改进 RWACO 的效率的瓶颈,试图进一步提高算法的运行效率;2) 通过一些细粒度的基准网络对 RWACO 进行测试,并对其进行改进,以期能够更好的发现较小规模社团,从而进一步提高该算法的聚类质量.

References:

[1] Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature, 1998,393(84):440−442. [doi: 10.1038/30918]

[2] Adamic LA, Huberman BA. Power-Law distribution of the World Wide Web. Science, 2000,287(5461):2115a. [doi: 10.1126/science.287.5461.2115a]

[3] Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of National Academy of Science, 2002, 9(12):7821−7826. [doi: 10.1073/pnas.122653799]

[4] Fortunato S. Community detection in graphs. Physics Reports, 2010,486(3-5):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[5] Yang B, Liu DY, Liu JM, Jin D, Ma HB. Complex network clustering algorithms. Journal of Software, 2009,20(1):54−66 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[6] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[7] Guimerà R, Amaral LAN. Functional cartography of complex metabolic networks. Nature, 2005,433(7028):895−900. [doi: 10.1038/nature03288]

[8] Duch J, Arenas A. Community detection in complex networks using extremal optimization. Physical Review E, 2005,72(2):027104. [doi: 10.1103/PhysRevE.72.027104]

[9] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008,2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]

[10] Lü ZP, Huang WQ. Iterated tabu search for identifying community structure in complex networks. Physical Review E, 2009,80(2):026130. [doi: 10.1103/PhysRevE.80.026130]

[11] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[12] Raghavan UN,, Albert R, Kumara S. Near linear-time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3):036106. [doi: 10.1103/PhysRevE.76.036106]

[13] Leung IXY, Hui P, Liò P, Crowcroft J. Towards real time community detection in large networks. Physical Review E, 2009,79(6):066107. [doi: 10.1103/PhysRevE.79.066107]

[14] Barber MJ, Clark JW. Detecting network communities by propagating labels under constraints. Physical Review E, 2009,80(2):026129. [doi: 10.1103/PhysRevE.80.026129]

[15] van Dongen S. Graph clustering by flow simulation [Ph.D. Thesis]. Utrecht: University of Utrecht, 2000.

[16] Pons P, Latapy M. Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications, 2006,10(2):191−218.

[17] Gunes I, Bingol H. Community detection in complex networks using agents. 2006, arXiv:cs/0610129v1.

[18] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of National Academy of Science, 2008,105(4):1118−1123. [doi: 10.1073/pnas.0706851105]

[19] E WN, Li TJ, Vanden-Eijnden E. Optimal partition and effective dynamics of complex networks. Proceedings of National Academy of Science, 2008,105(23):7907−7912. [doi: 10.1073/pnas.0707563105]

[20] Yang B, Cheung WK, Liu JM. Community mining from signed social networks. IEEE Trans. on Knowledge and Data Engineering, 2007,19(10):1333−1348. [doi: 10.1109/TKDE.2007.1061]

[21] Yang B, Liu JM, Feng JF, Liu DY. On modularity of social network communities: The spectral characterization. In: Proc. of the 2008 IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2008). Sydney: IEEE/WIC/ACM Press, 2008. 127−133. [doi: 10.1109/WIIAT.2008.70]

[22] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining partitionings. In: Proc. of the AAAI. Edmonton: AAAI Press, 2002. 93−98.

[23] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[24] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005,P09008. [doi: 10.1088/1742-5468/2005/09/P09008]

[25] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

[26] Fortunato S, Barthélemy M. Resolution limit in community detection. Proceedings of National Academy of Science, 2007,104(1): 36−41. [doi: 10.1073/pnas.0605965104]

[27] Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977, 33(4):452−473.

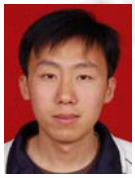
[28] Lusseau D. The emergent properties of a dolphin social network. Proceedings of the Royal Society B: Biological Sciences, 2003, 270(Suppl2):186−188. [doi: 10.1098/rsbl.2003.0057]

[29] Newman MEJ. Modularity and community structure in networks. Proceedings of National Academy of Science, 2006,103(23): 8577−8582. [doi: 10.1073/pnas.0601602103]

[30] Gleiser PM, Danon L. Community structure in jazz. Advances in Complex Systems, 2003,6(4):565−573. [doi: 10.1142/ S0219525903001067]

附中文参考文献:

[5] 杨博,刘大有,Liu Jiming,金弟,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66. http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]



金弟(1981－),男,河北乐亭人,博士生,主要研究领域为数据挖掘,复杂网络分析.

刘大有(1942－),男,教授,博士生导师,主要研究领域为知识工程与专家系统,Agent 系统,时空推理,数据挖掘.

杨博(1974－),男,博士,教授,博士生导师,主要研究领域为 Agent 系统,数据挖掘,复杂网络分析.

何东晓(1984－),女,博士生,主要研究领域为数据挖掘,复杂网络分析.

刘杰(1973－),男,博士,副教授,主要研究领域为数据挖掘,机器学习.