doi:10.3969/j.issn.1000-436x.2014.z2.008

基于改进蛙跳算法的社区划分方法

王桐，赵昕琳
（哈尔滨工程大学 信息与通信工程学院，黑龙江 哈尔滨 150001）

摘 要：现有的网络社区划分方法以社区为主体，机械地将每一个节点划分到某一个社区，在真实网络中，对于活跃度低的用户进行划分会大大降低划分精确度，同时增加时间复杂度，并具有较小的划分意义。因此，将蛙跳算法与社区划分相结合，通过对青蛙性能的排序，提取活跃度高的用户，从而提高划分精确度。实验结果表明该方法具有良好的性能。

关键词：社交网络；社区划分；蛙跳算法；社区结构
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1000-436X(2014)Z2-0048-05

Improved shuffled frog-leaping algorithm based network community detection method

WANG Tong, ZHAO Xin-lin
(College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001,China)

Abstract: Existing community method aims to divide nodes into a community mechanically. In a real network, it will reduce the classification accuracy greatly for the low active users, while increasing the time complexity. It has small significance. Therefore, this paper will combine shuffled leap-frog algorithm with community detection method. It will extract active users by sorting on properties of frog, so as to improve the efficiency of division. Experimental results show that the method has good performance.

Key words: social networks; community detection; shuffled frog-leaping algorithm; community structures

1 引言

随着互联网技术的发展以及传统 PC、智能手机的普及，截至 2014 年 6 月，我国网民数已由 1997 年的 62 万激增至 6.32 亿，社交网站用户更是达到 2.57 亿之多。在如此庞大用户的背景下，如何有效地利用复杂的网络资源成为国内外研究的热点。

日常生活中，社交网络拥有巨大的影响力，人们的生产生活受其影响并改变。国内的新浪微博，国外的 Twitter、Facebook 应运而生。在这样的复杂网络环境下，划分网络的社区结构对于了解整个网络的形态以及各用户之间的关系具有重要的意义。Kernighan-Lin 算法[1]、谱平分法[2]、GN 算法[3]和 Newman 快速算法[4]等为现在主流的社区划分算法。Kernighan-Lin 算法是基于贪婪算法原理的二分法，为一种试探性的优化算法。谱平分法是基于 Laplace 矩阵的划分算法。GN 算法引进了边介数(betweenness)[5]的概念，定义为网络中经过某条边的最短路径的数目，其代表了这条边在网络中的重要程度，是一种分裂算法。在社区划分过程中，每一次迭代后，边介数最大的边将被移出，直到移除网络中所有的边。Newman 快速算法同样是基于贪婪思想的凝聚算法。该算法引进增益函数 Q[6]，目的是将复杂的网络社区划分为 2 个已知大小的社

收稿日期：2014-10-22
基金项目：国家自然科学基金资助项目(61102105)；中国博士后科学基金资助项目(20080440840)；教育部博士点基金资助项目(20102304120014)；黑龙江省自然科学基金资助项目(F201029)
Foundation Items: The National Natural Science Foundation (61102105); China Postdoctoral Science Foundation (20080440840); Doctoral Fund of Ministry of Education (20102304120014); The Natural Science Foundation of Heilongjiang Province (F201029)

区，搜寻使 Q 值最大的划分方法。根据贪婪算法[7]的原理，社区向着使 Q 值增大最多或者减小最少的方向融合。它适用于拥有 100 万以上节点的网络，因此解决了大规模数据的社区划分问题。

此外，朱文强提出了一种粗糙集与蚁群算法在网络社区结构发现中的应用研究[8]算法；胡正华等人提出了基于单亲遗传算法的加权复杂网络社区划分问题研究[9]算法。上述几种社区划分算法除了存在诸如时间复杂度高，只能将网络划分成 2 个社区等问题外，还有一个缺陷无法避免，即这些社区划分算法试图将每一个节点都划分进某一个社区，这就对数据具有较高的要求。然而在真实的社交网络中，存在许多低活跃度用户，这些用户的好友数目少，发表博文数目少，动态更新频率低。这时，机械性地将该类用户划分到某一社区，既无法达到高准确度的划分，也降低了划分社区的意义。因此，本文将蛙跳算法（SFLA, shuffled frog leaping algorithm）与社区划分相结合，通过对每个节点性能的排序，提取活跃度相对较高的用户，从而提高划分效率。

2　改进蛙跳算法

蛙跳算法是一种全新的基于社会群体协作的后启发式群体进化算法，具有较强的交互能力和优良的全局搜索能力[10]。算法通过各个青蛙之间的文化交流，寻找不同的石头进行跳跃地来寻找食物较多的地方。文献[11]中，ElbeltagiEmad 等人将 SFLA 算法与蚁群算法 ACO、粒子群算法 PSO、模因算法 MA、遗传算法 GA 相比较，得出了实验结果，其表明在解决某些连续函数问题上，遗传算法的性能明显低于 SFLA 算法；文献[12]中，Eusuff 等人在实验中对比得出，蛙跳算法 SFLA 的收敛速度和精度，在处理组合优化问题时有一定优越性。

在青蛙群体中，全局适应度最好的解为 P_g，子族群中适应度最好的解为 P_b，子族群中适应度最差的解为 P_ω。蛙跳算法首先要在每个子族群中进行局部搜索，寻找子族群中适应度最差的青蛙，并对其更新，当子族群更新迭代到一定阶段以后，再对其进行全局信息交换，直到所设置的条件满足为止[13]。蛙跳算法搜索流程如图 1 所示，更新策略如下。

青蛙更新距离：

$$newD_\omega = oldP_\omega + rand() \times (P_b - P_\omega)$$

$$(-D_{\max} \leqslant D_w \leqslant D_{\max}) \tag{1}$$

其中，D_w 表示青蛙个体的调整矢量，$D_{\max}$ 表示青蛙个体允许改变的最大步长。

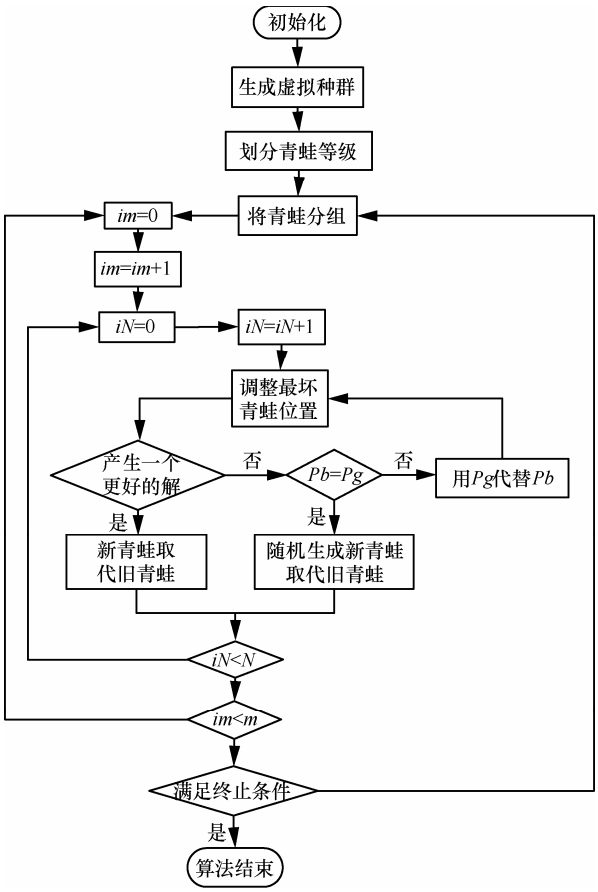

图 1　蛙跳算法搜索流程

蛙跳算法的执行步骤如下。

当已达到标准测试结果，或者达到迭代次数最上限，或者在最近的 k 次全局迭代后全局最优解仍没有明显改善时，算法停止。

在式(1)中，计算距离的更新公式过于简单直接，无法达到精确的位置移动，进而可能越过最优解，无法达到划分最优。经过仔细分析可以发现在考虑将最差位置青蛙趋近子群最优解 P_b 的同时，也趋近与全局最优解 P_g，以便于再次执行全局搜索时，青蛙更具有全局最优性。算法改进后，式(1)变为

$$newD_\omega = oldP_\omega + c_1 r_1 (P_b - P_\omega) + c_2 r_2 (P_g - P_\omega)$$

$$(-D_{\max} \leqslant D_w \leqslant D_{\max}) \tag{2}$$

其中，c_1、c_2 为学习因子，c_1 用以调节青蛙走向子族群最优解，c_2 用以调节青蛙走向全局最优解；r_1、

r_2 为(0, 1)内的 2 个随机数。c_1、c_2 的选取会直接影响青蛙的更新方向与步长，但在实际仿真中，c_1、c_2 的选取比较困难，经过仔细分析后，将 c_1r_1、c_2r_2 分别合并为 m_1、m_2，以此来减少参数选择的困难。变量替换后，式(2)变为

$$newD_\omega = oldP_\omega + m_1(P_b - P_\omega) + m_2(P_g - P_\omega)$$
$$(-D_{\max} \leqslant D_w \leqslant D_{\max}) \tag{3}$$

改进后的蛙跳算法具有以下优点：1) 计算公式更加具体，精确；2) 使青蛙更具有全局最优性，在下一次迭代之前，青蛙能更接近最优解。

3　基于改进蛙跳算法的社区划分方法

在蛙跳算法中，其青蛙更新距离的计算如式(1)所示。在式(1)中，距离的计算方法以及参数的使用过于简单。本文针对蛙跳算法进行了局部改进，改进后的公式更严谨，距离计算更精确，在改进后的蛙跳算法基础上，提出一种改进的网络划分算法，在该算法中，首先使用 Python 将初始数据转换成图数据集，再对网络进行社区划分。通过对青蛙性能进行排序，提取活跃度高的用户，从而提高划分精确度。

1) 数据初始化处理

本文的实验数据全部来自爬盟平台由爬虫工具获得。在爬盟平台抓取的新浪微博用户数据的字段信息如表 1 所示，所有抓取到的数据会被写入本地文件中，并上传至爬盟服务器。

在真实的社交网络中，研究人员通常是对图形式的数据进行处理。在实验中所用到的是基于用户与粉丝之间的无向关系图，但在爬盟平台爬取的数据格式并非本文研究所需要的图数据，因此，需要通过一定方法将原始数据中的用户关系提取出来，进而转化为实验所需的图模型如图 2 所示。

图 2　新浪微博用户数据图模型

2) 社区划分算法

基于上述改进后的蛙跳算法和经过 Python 处理的初始化数据，提出一种改进的复杂网络社区划分方法。在每个节点上放置一只青蛙，每只青蛙都具备向任意方向跳跃的能力，但是实际的移动都趋向于靠近适应度最好的青蛙。经过青蛙多次跳跃、变更，也就是算法中的多次迭代，当算法满足收敛条件后，此时的青蛙分布情况，即为社区划分的结果。

表 1　　新浪微博用户信息结构

字段	名称类型	描述
user_id	bigint	用户 ID
screen_name	varchar	用户屏幕名
sex	varchar	性别
base_info	varchar	认证信息
description	varchar	自我介绍
address	varchar	所在地区
username	int	用户名
ttention_num	int	关注数量，默认值为 0
fans_num	int	粉丝数量，默认值为 0
message_num	int	消息数量，默认值为 0
career_info	varchar	工作信息
education_info	varchar	教育信息
profile_image_url	varchar	用户头像 URL
is_verified	tinyint	认证标识
tag	varchar	用户标签
birthday	date	生日
QQ	int	QQ 号码
Msn	varchar	MSN 账号
Email	varchar	Email 账号
create_time	int	用户创建时间
follower_userid	varchar	用户关注人的 ID 列表
is_verified	tinyint	
VIP 标识		
is_daren	tinyint	达人标识
vip_level	tinyint	VIP 等级

通过模拟青蛙的跳跃，将蛙跳问题引入到社区划分的求解过程中来，通过对局部问题的求解，得到整体问题的解决方案，同时可以使本文的社区划分得到更加准确的结果。

本文将复杂网络的社区划分问题经过数学分析，重组成一个函数优化问题，建立如下的数学模

型：$\max_{x\in r}\{Q(X)\}$。其中，Q 是与划分方案 X 相关的一组优化函数，即蛙跳算法中的适应值函数。本文将蛙跳算法中的适应值函数与社区划分的评价标准相结合，即将适应值函数等价于模块度 Q 值。因此，求解复杂网络社区的最优划分方案问题可以转化为最优化与社区划分方案相关的适应值函数的问题。当适应值函数收敛到最优值，此时对应的最优解即为复杂网络社区划分的最优方案，因此，本文把复杂网络社区划分问题直接转化为对最优适应值的求解问题[14]。

假设将网络划分为 k 个社区 $V_1,V_2,\cdots,V_k$，定义一个 $k\times k$ 维的矩阵 $\boldsymbol{e}=(e_{pq})$，其中 e_{pq} 为网络中连接 2 个不同社区 V_p 和 V_q 中节点的边在整个网络所有边中的比例。矩阵对角线上的各元素的和为 $Tr(e)=\sum_p e_{pp}$，每一行中各元素之和为 $a_p=\sum_p e_{pq}$，此时模块度 Q 值的计算式如下

$$Q=\sum_p (e_{pp}-a_p^2)=Tr(\boldsymbol{e})-\left\|\boldsymbol{e}^2\right\| \tag{4}$$

此外，模块性还有一种更简便的表达方式，如下

$$Q=\frac{1}{2m}\sum_{ij}\left(\boldsymbol{A}_{ij}-\frac{d_i d_j}{2m}\right)\varnothing(c_i c_j) \tag{5}$$

其中，m 表示网络中边的个数，$\boldsymbol{A}$ 为网络的邻接矩阵，d_i 和 d_j 分别为节点 V_i 和 V_j 的度，当节点 V_i、V_j 相连时，$\boldsymbol{A}_{ij}$=1，否则为 0，当节点 V_i、V_j 在同一个社区时，$\varnothing(c_i c_j)=1$，否则为 0。经过分析计算，上述公式也等价于

$$Q=\sum_{p=1}^{k}\left[\frac{l_p}{m}-\left(\frac{d_p}{2m}\right)^2\right] \tag{6}$$

其中，l_p 为社区 V_p 内部连接的边的数目，d_p 为社区 V_p 的总度值。

将蛙跳算法与社区划分方法结合的具体流程如下。首先，输入原始微博数据文件，将其转化为实验所用的图模型。然后计算各条边相对于所有源节点的边介数，进行比较后，删除该值最大的边，将青蛙随机分布到各节点上，使最坏青蛙不断向最优青蛙迭代，得到最终的社区划分结果[15]。伪代码如图 3 所示。

4　实验测试

本文衡量社区划分结果为模块度值（modularity），简称 Q。Q 值越大表示社区结构划分得越明显，划分结果越好。Q 值通常介于 0.3 到 0.7 之间，上限为 1。

```
Input:Network G(V,E)
Output:Network communities Cn
capculate the betweenness of all edges
if(the betweenness of the edge is maximum)
    delete the edge
    for all the remains
        update the betweenness of them
    end for
end if
group frogs by grades
ad just the location of the worst one
if (produce a better solution) then
    new frog replaced the old
else
    generate a new frog replaced the old randomly
end if
return Cn
```

图 3　蛙跳算法与社区划分的伪代码

从图 4 中可以看出，在迭代次数增加的同时，本文改进算法得到的模块度增量 Q 值也在增加，并最终趋近于 0.501，传统 GN 算法的模块度增量 Q 趋近于 0.459，说明改进算法可以有效地提高划分性能。

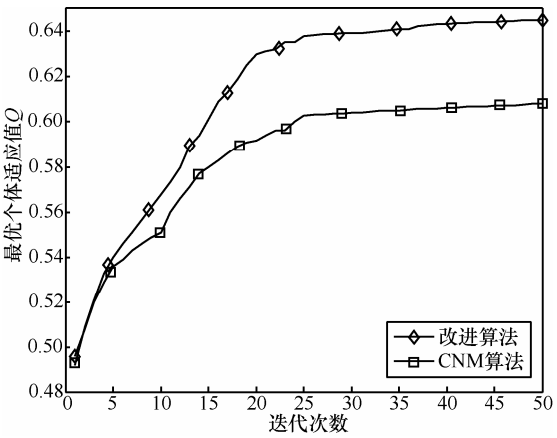

图 4　GN 算法与改进算法性能对比

从图 5 中可以看出，当迭代 7 次以后，改进算法的模块度增量值逐渐高于改进前算法，并最终趋近于 0.591，而 Newman 算法的模块度增量 Q 趋近于 0.576，看出改进算法相对于 Newman 算法的优越性。

从图 6 中可以看出，在前几次迭代过程中，2 种算法模块度增量 Q 值并未见明显区别，但随着迭代次数继续增加，本文改进算法得到的模块度增量 Q 值最终趋近于 0.645，CNM 算法的模块度增量 Q 趋近于 0.608，由此可见，改进后算法得到的社区划分结果更好。

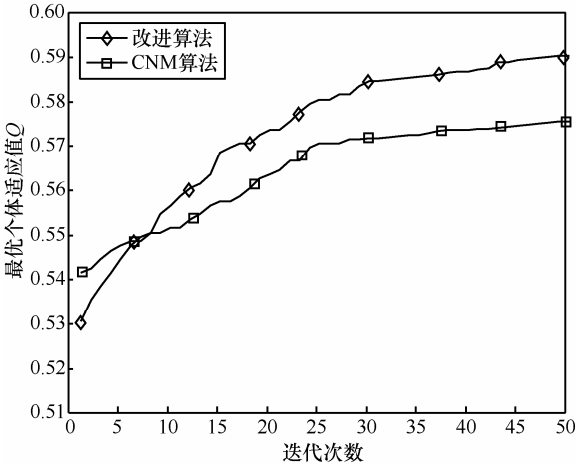

图 5　Newman 算法与改进算法性能对比

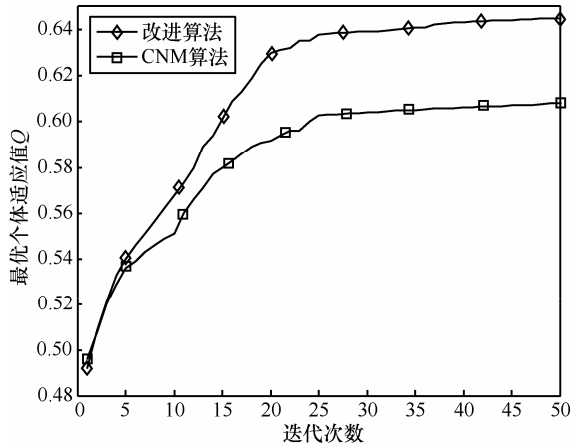

图 6　CNM 算法与改进算法性能对比

5　结束语

本文对蛙跳算法进行改进，改进之后的蛙跳算法公式更严谨，距离计算更精确。基于这种改进算法，本文将蛙跳算法结合到传统社区划分算法中，提出一种新的网络划分算法，以提高推荐效率。本文提出的算法与传统 GN 算法、Newman 算法、CNM 算法相比较，实验结果表明该方法在对社区进行划分时具有较高的准确度，从而更好地在实践中为用户提供更加优质的服务。

参考文献：

[1] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49:291-307.

[2] BARNARD S T, SIMON H D. Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems[J]. Concurrency: Practice and Experience, 1994, 6:101-117.

[3] MORRISON N, LORD R T, INGGS M R. The Gauss-newton algorithm applied to track-while-scan radar[A]. 2007 IET International Conference on Radar Systems[C]. 2007.1-5.

[4] EGHBAL M, SAHA T K, HASAN K N. Transmission expansion planning by meta-heuristic techniques: a comparison of shuffled frog leaping algorithm, PSO and GA[A]. 2011 IEEE Power and Energy Society General Meeting[C]. 2011.1-8.

[5] ZHANG N. Community Structure in Complex Networks Partitioning Algorithm Research[D]. Dalian University of Technology, 2009.

[6] GALKOWSKI P J, ISLAM M A. An alternative derivation of the modified gain function of Song and Speyer[J]. IEEE Transactions on Automatic Control, 1991, 36: 1323-1326.

[7] EDMONDS J. Matroids and the greedy algorithm[J]. Mathematical Programming, 1971, 1:127-136.

[8] ZHU W Q. Research and Application of Rough Sets and Ant Colony Algorithm in Finding Community Structure in the Network[D]. Soochow University, 2011.

[9] HU Z H. Based on the Partheno Genetic Algorithm of the Weighted Complex Network Community Division[D]. University of an Inner Mongolia, 2012.

[10] AMIRI B, FATHIAN M, MAROOSI A. Application of shuffled frog-leaping algorithm on clustering[J]. The International Journal of Advanced Manufacturing Technology, 2009, 45:199-209.

[11] BHATTACHARJEE K K, SARMAH S P. Computational Collective Intelligence Technologies and Applications[M]. Springer Berlin Heidelberg, 2012:513-522.

[12] AMIRI B, FATHIAN M, MAROOSI A. Application of shuffled frog-leaping algorithm on clustering[J]. The International Journal of Advanced Manufacturing Technology, 2009, 45:199-209.

[13] ZHEN Z, WANG Z, GU Z, *et al*. Advances in Computation and Intelligence[M]. Springer Berlin Heidelberg, 2007.

[14] LUO Y R. Multi Population Genetic Algorithm and Its Application in the Community Partition of Complex Networks[D]. Jiangxi University of Science, 2012.

[15] TP K, BE B, J N. Staff Expertation in a community services division[J]. The Nebraska State Medical Journal, 1965:116-120.

作者简介:

王桐（1977-），男，黑龙江哈尔滨人，哈尔滨工程大学副教授，主要研究方向为无线网络、云计算与信息安全。

赵昕琳（1990-），女，黑龙江哈尔滨人，哈尔滨工程大学硕士生，主要研究方向为社交网络，推荐系统。