社交网络中FN算法结果的后处理研究

倪 涵 白清源
(福州大学数学与计算机科学学院 福州 350108)

摘 要 在复杂网络问题的研究中,众多聚类算法的横向比较与改进研究方面的工作在近几年吸引了大量注意并得到深入研究。其中,基于模块度的算法被广泛应用,而模块度也作为评价聚类的一项指标。在这类算法中,基于模块度的快速Newman算法(Fast-Newman algorithm,FN)显得较为突出,许多相关的深入研究由此展开,但多数工作是基于算子改进、应用领域等方向展开的,而对于算法结果的研究工作则更多偏向于评价、测量和总结。该研究从FN算法的结果入手,对算法的分类结果进行数据的后处理。在研究中发现了FN算法中常见的错误类型,并提出了3种不同的解决方案,使得最终结果更加符合实际,达到更好的聚类效果。在部分案例中准确率可提高至100%。

关键词 社团挖掘,FN算法,后处理,复杂网络,社交网络
中图法分类号 TP393.02 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.054

Research of Post-processing of FN Algorithm Results in Social Networking

NI Han BAI Qing-yuan
(School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract During the last few years, the clustering algorithms' transverse comparison in the field of complex networks has attracted a lot of attention. Among them, the algorithm based on modularity is widely used, and modularity is viewed as an evaluation for a clustering. The Newman fast algorithm based on the modularity (Fast-Newman algorithm, FN for short) is more prominent. Many related researches are based on the FN algorithm, however, most of the works focus on operator improvement, application field extension, etc. For the results of the FN algorithm, most of research works are tend to evaluation, measure and summarize. The study focuses on the post-processing of the classification results of the FN algorithm. The research shows the common feature of the error nodes in the FN algorithm, and proposes three different solutions to make the final results more correspond to the actual situation and achieve a better clustering result. In some cases, the clustering accuracy is up to 100%.

Keywords Community mining, FN algorithm, Post-processing, Complex network, Social network

1 引言

社团结构探测和聚类发现是复杂网络领域中的一个重要课题。一般而言,社团可以包含模块、类、群、组等各种含义[1]。如万维网可以看成是由大量网站社团组成的,同一个社团内部的各个网站讨论的是一些有共同兴趣的话题[2-4]。类似地,在生物网络或者电路网络中,同样可以将各个节点根据其不同的性质划分为不同的社团[5-7]。揭示网络中的社团结构,对于了解网络结构与分析网络特性都很重要。社团结构分析在生物学、物理学、计算机图形学和社会学中都有广泛的应用[8,9]。网络社团结构的研究已经有很长的历史,它与计算机科学中的图形分割(Graph Partition)和社会学中的分级聚类(Hierarchical Clustering)有着密切的关系[10,11]。

因此,由物理学、应用数学、计算机科学、生物学及社会学原理和方法发展出许多社团结构探测算法[12-19]。根据这些算法可得到很多社团结构,具有巨大的应用价值。为了辨别所挖掘的社团结构的优劣,需要一个量化的标准来评估各种社团结构。迄今为止,文献[20]提出的模块度函数Q是最广泛使用的评价函数。同时,基于模块度的理念,Newman提出的FN算法[21]是一种十分有效的用于发掘社团结构的聚类算法。

通过对FN算法的研究发现,在所有错误分类情况中,有一类错误情况高频出现,即:当某点与两个或者两个以上聚类具有相同连接边数时,该点有可能被划分到错误的聚类。这一错误的情况是由于FN算法所基于的模块度算子特性所产生的,该算子的特性以及这一错误分类的原因会在下文中详细阐述。

本文针对FN算法的这一高频错误分类的情况进行了研究,找到了错误原因,并由此提出了3个后处理方案:基于间接关系、基于步长和基于中心关系。以此用于改善FN算法的计算结果数据,能够提高聚类结果的准确率,使其更加符合客观实际中存在的社团结构。在部分案例以及特殊情况下,准确度可达100%。

到稿日期:2014-06-04 返修日期:2014-08-25
倪 涵(1990—),男,硕士生,主要研究方向为复杂网络聚类算法,E-mail:85004195@qq.com;白清源(1964—),男,博士,教授,主要研究方向为数据集成、数据挖掘。

2 研究现状与研究意义

Kernighan-Lin算法是一种试探优化法[22]，是该领域内的第一个算法。它是一种基于贪婪算法原理将网络划分为两个大小已知的社团的二分法。其基本思想就是为网络的划分引进一个增益函数 Q，定义为两个社团内部的边数减去连接两个社团之间的边数，然后寻找使 Q 值最大的划分方法。首先，将网络中的节点随机地划分为已知大小的两个社团。在此基础上，考虑所有可能的节点对，其中每个节点对的节点分别来自两个社团。对每个节点对，计算如果交换这两个节点可能得到的 Q 的增益 $\Delta Q=Q_{after}-Q_{before}$，之后交换最大的 Q 对应的节点对时，记录交换以后的 Q 值。规定每个节点只能交换一次。重复这个交换过程，直到某个社团内所有的节点都被交换一次为止。但其需要事先知道目标聚类的大小，故难以应用于实际。

在Kernighan-Lin算法之后，又发展出了频谱法、分裂法、凝聚法等一系列算法。频谱法所基于的理论基础可见文献[23,24]。一个有 n 个节点的无向图的Laplace矩阵是一个 $n\times n$ 维的对称矩阵 L，其中 L 的对角线上的元素 l_{ii} 是节点 i 的度，而其它非对角线上的元素 l_{ij} 则表示节点 i 和节点 j 的连接关系。如果这两个节点之间有边连接，则 l_{ij} 值为 -1，否则为0。也可以将矩阵 L 表示成 $L=K-A$，其中，K 是一个对角矩阵，其对角线上的元素就对应各个节点的度，而 A 则为该网络的连接矩阵。L 矩阵所有的行与列的和都为0，因此，该矩阵总有一个特征值为0，且其对应的特征向量为 $l=(1,1,1,\cdots)$。可以从理论上证明，不为零的特征值所对应的特征向量的各元素中，同一个社团内的节点对应的元素是近似相等的。在应用中，对于明显带有社团结构的复杂网络，该算法速度很快，但对于不那么明显的则速度慢且效果差。

分裂法的典型代表为GN算法[8]，而凝聚算法的典型代表为FN算法[21]。分裂算法与凝聚算法是一对过程互为相反的算法。分裂算法是由完整图，依照某个算子向下拆分至社团结构；而凝聚算法是从点开始，依照某个算子向上凝聚成社团结构。GN算法的基本思想是通过不断地从网络中移除介数(Betweenness)最大的边将整个网络分解为各个社团。边介数定义为网络中经过每条边的最短路径的数目，它为区分一个社团的内部边和连接社团之间的边提供了一种有效的度量标准。之后应用模块度 Q 来选择截断平面来确定最佳的分裂次数。但是算法复杂度较大，因此对于大型规模的复杂网络难以满足要求。

FN算法与GN算法相比，在效果上不相上下，但是速度却要高出许多，可以分析超过100万节点数的复杂网络，因此现在主流的算法以凝聚算法为主，衍生出了许许多多不同的算法。其中FN算法被视为一个标杆，作为其他算法效果的对比对象来验证算法的优劣程度，但其本身却很少被研究和改进。因此，本文通过观察FN算法的表现，并研究FN算法的算子特点，结合算法的结果，在一定程度上对FN算法进行合理的改进。

3 相关知识

3.1 模块度(Modularity)[20]

为了得到具有实际意义的社团结构，Newman等人定义了社团模块度函数 Q 来衡量网络划分质量。Q 函数的定义为：网络社区内实际存在的边数与完全随机的连接情况下社区内期望的边数之差。那么通过比较实际覆盖度(覆盖度是社团内部连接数占总连接数的比例)与随机连接情况下覆盖度的差异来评估社团结构[9]。给定一个网络 $G=(V,E)$，假设该网络被划分为 k 个社团，则定义一个 $k\times k$ 的对称矩阵 e，其元素 e_{ij} 表示 i 社团的点与 j 社团的点之间的边所占比例。矩阵的元素 e_{ii} 表示所有在相同社团 i 中的边所占的比例。此外，定义了行和 $a_i=\sum_j e_{ij}$，其代表了连接到 i 社团的边的比例。在全随机连接的状态下，可以得到 $e_{ij}=a_i a_j$。因此定义：

$$Q=\sum_i (e_{ii}-a_i^2)$$

如果社团内没有边比随机的好，则得到 $Q=0$。在实验中，Q 值一般位于0.3到0.7之间。

3.2 FN算法[21]

FN算法是Newman基于模块度提出的一种快速挖掘算法，属于凝聚算法的一种贪心算法。根据模块度 Q，遍历连接两个社团，找到增量最大或减量最小的组合，合并成新的社团，并构造层次树。算法过程定义为：

$$\Delta Q=e_{ij}+e_{ji}-2a_i a_j=2(e_{ij}-a_i a_j)$$

在每次合并后，对于新的社团划分，重新更新其对称矩阵 e。算法的最大连接步骤为 $n-1$，则可以构造出层次树，再应用模块度 Q 选择最佳的截断面来划分社团，如图1所示。

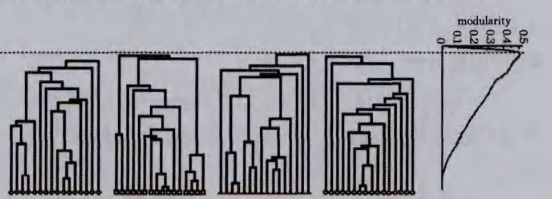

图1 应用模块度(Modularity)选择截断面

4 FN算法问题发现与分析

4.1 算法实验

以典型的空手道俱乐部[25]为例，应用FN算法进行实验，如图2所示。

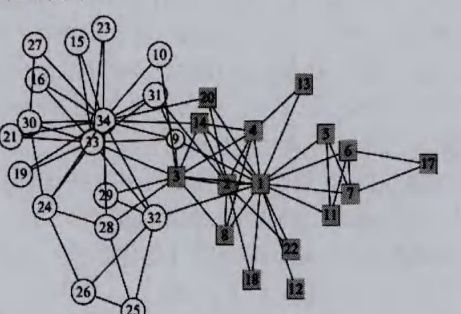

图2 空手道俱乐部成员关系

这个关系图被大量用于社团挖掘领域以作典例。由于一些关系使得俱乐部划分为两个派系，本文并不对此展开详细描述。而这种现实生活中的例子，对于研究社团挖掘算法来说十分便利。教员和管理者由编号1到34，由灰色方格和白色圆圈分别的标识了现实中他们所属的社团。

将网络的邻接矩阵代入FN算法进行计算，可得层次树，如图3所示。

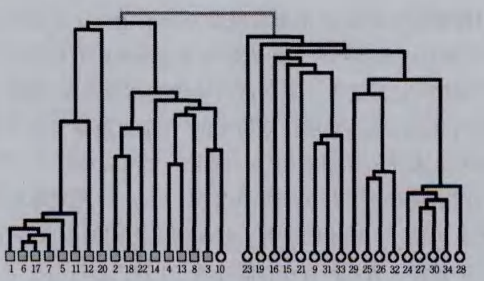

图 3　凝聚算法层次树

利用模块度 Q 选择最佳的截断面来划分社团，可以得到最高值 $Q=0.372$，此时将得到两个社团，分别为：

$c_1=\{1,2,3,4,5,6,7,8,10,11,12,13,14,17,18,20,22\}$

$c_2=\{9,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34\}$

4.2　问题发现

据上文的实验发现，FN 算法在处理空手道俱乐部的社会关系时，总体情况正确，结果令人满意，但却错误地将 10 号点划分到了与现实中他所属社团不相符的社团。

定义 $Algorithm(V_i)$ 为算法将点 V_i 所划入的社团，$Real(V_i)$ 为现实中 V_i 应属的社团。则算法的出错点可以表示为：

$$Algorithm(V_{10})\neq Real(V_{10})$$

这一错误的社团划分情况在许多 FN 的应用场合都会发生，这是由于算法的特性，原因将在下文说明。

4.3　问题分析

在 FN 算法结束后，通过对划分结果进行大量的研究、观察与分析发现：被错误划分社团的点大部分都有着共同的特征性。

首先我们在社团划分结果中找到与现实中不相符的划分点 V_i，即：

$$\{V_i \mid Algorithm(V_i)\neq Real(V_i)\}$$

并观察这种错误的划分出现在什么迭代的哪一次过程中。以空手道俱乐部为例，通过观察层次树图 3 发现，错误的划分发生在第 14 次迭代过程中。在第 13 步骤时，聚类情况为：

$c_1=\{1,5,6,7,11,17\}$，$c_2=\{2,18\}$，$c_3=\{4,13\}$，$c_4=\{25,26,32\}$，$c_5=\{24,27,28,30,34\}$

此时尚未划分的点集为$\{3,8,9,10,12,14,15,16,19,20,21,22,23,29,31,33\}$，点 V_{10} 与独立点 V_3 此时有边 $e(3,10)$，与社团 c_5 中的点 V_{34} 此时有边 $e(34,10)$。

在迭代过程中，将尚未划分的点也视为一个社团。因此点 V_{10} 与两个社团有相同数量的相连的边。

在算法中，每一轮合并都基于下述公式：

$$\Delta Q=e_{ij}+e_{ji}-2a_ia_j=2(e_{ij}-a_ia_j)$$

通过观察公式的构造可以得知，若在某轮的合并中 V_i 与某 2 个或者 2 个以上（其中包含正确和错误的划分）的社团拥有相同数量的连接边时，虽然邻接矩阵中点 V_i 与社团 j 和 k 的 $e_{ij}=e_{ik}$，但是由于社团 j 和 k 所包含的点数不同，也就是说 $a_ia_j\neq a_ia_k$，在出现这种情况时，点 V_i 会被划分给更大 ΔQ 所对应的社团（即点数较少的社团）中，且其不一定是 V_i 应当所属的社团。即此时错误的划分可能出现。

将被错误划分的点独立出来，即它不再属于任何一个社团，而其他的点保持它们所属的社团。此时可以发现，被错误划分的点 V_i，与社团 $Algorithm(V_i)$ 内点所相连的边数，以及社团 $Real(V_i)$ 内点所相连的边数总数相同或更少。如图 4、图 5 所示，图中黑色代表聚类，白色代表被错误划分的点，之间的连线代表边。图中聚类只有 2 个，而实际中可以是两个或者两个以上，边的情况也类似。

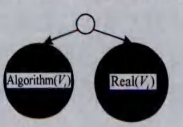

图 4

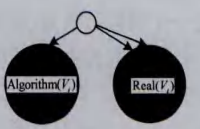

图 5

此外，通过研究其他例子，定义了这一类高频出现的错误点的共同特性为：该点与两个或者两个以上的社团具有相同数量的连接边，或与算法所属社团连接边少于与其他社团连接边。若这一特性出现，则说明该点很有可能被划分到错误的社团。

以上文实验结果为例，可以发现，将点 V_{10} 独立出来后，此时社团划分变为：

$c_1'=\{1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22\}$

$c_2=\{9,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34\}$

此时，点 V_{10} 与社团 c_1 有边 $e(3,10)$，与社团 c_2 有边 $e(34,10)$。则独立出的点 V_{10} 与社团 c_1、c_2 有相同数量的相连的边，即错误划分的点与错误的社团和正确的社团的连接边数总数是相同的。这在其他例子中也能得到体现。因此我们考虑后处理的方法，将结果进行加工，以得到更好的聚类效果。

5　解决方案的提出与实验

5.1　解决方案

在 FN 算法结果中，扫描，并将具有上述特性的点从结果中剥离，再依据一些算法重新给定归属社团，使得聚类结果更加符合实际。

假设我们已经通过扫描获得了 FN 算法的结果，得到两个社团 c_j 和 c_k，并且已找到目标点 V_i 和独立出点 V_i 之后的两个社团 c_j' 和 c_k'。

方案 1　基于间接关系

由于所要处理的点的类型与两个或两个以上的社团都具有相同连接数（边数），因此采用间接联系的方法。步骤如下：

Step1　$P=2$，令步长为 2。

Step2　$T1=\sum_{V_j\in c_j'}(d(V_i,V_j)=P)$ 且 $T2=\sum_{V_k\in c_k'}(d(V_i,V_k)=P)$，其中 $V_j\in c_j'$，$V_k\in c_k'$。当 $d(V_1,V_2)=P$ 成立时取值为 1，否则为 0。即统计步长为 P 的点的个数。

Step3　若 $T1>T2$，则 $V_i\in c_j$，或者 $T1<T2$，则 $V_i\in c_k$。如果 $T1=T2$，则 $P+1$ 并且执行 Step2。

该方案基于如下思想：V_i 与 c_j'、c_k' 的直接联系相同时，若间接联系更强（V_i 可以通过其在某一社团内的直接联系找到更多的间接联系），则更有可能属于间接联系更强的一类，特别是在处理社会关系网络的时候。在最坏情况下算法复杂度为 $O(n)$，由于复杂网络的邻接矩阵一般是稀疏的，因此平均复杂度为 $O(\log n)$。

方案 2　基于步长

该方案基于如下思想：在一个社团中，两个成员之间的互相联系的代价要小于跨社团之间交流的代价。因此，采用Dijkstra最短路径算法计算点V_i与c_j'、c_k'中所有点的最短路径（最短路径并不跨社团，即在计算过程中将c_j'、c_k'之间相连的边删去，独立考虑c_j'、c_k'），并按照c_j'、c_k'分别求和，算法步骤如下：

Step1　计算$D1=\sum_{V_j\in c_j'} Dijkstra(V_i,V_j)$和$D2=\sum_{V_k\in c_k'} Dijkstra(V_i,V_k)$，其中$Dijkstra(V_1,V_2)$表示点$V_1$和$V_2$之间的最短路径。

Step2　若$D1<D2$，则$V_i\in c_j$；或者$D1>D2$，则$V_i\in c_k$。

特别地，当$D1=D2$时，按照各自的社团最短路径的降序排列，依次从$D1$、$D2$中减去最长路径，再执行Step2。在最坏情况下，算法复杂度为$O(n^2)$。而对于稀疏网络来说，其平均算法复杂度为$O(n\log n)$。

方案3　基于中心关系

该方案基于社团中心的思路。假设社团中存在一个或若干个核心点，这个核心点类似于交通枢纽，一般处于社团的中心地位（繁忙地位）。

应用广度优先最短路径算法，计算出一个社团内任意两个成员之间的最短路径，并记录最短路径所经过的点。那么处于中心地位的点出现的频率应当是最高的，记为$CoreV(c)$。再应用Dijkstra最短路径算法来计算点V_i到$CoreV(c)$之间的距离，来确定V_i的归属社团。算法步骤如下：

Step1　分别计算社团c_j'和c_k'内部成员之间两两的最短路径，并记录所经过的点（社团之间的路径不参与计算）。

Step2　统计出各自社团的中心点$CoreV(c_j)$和$CoreV(c_k)$（一个或若干个）。

Step3　计算$D1=Dijkstra(V_i,CoreV(c_j'))$和$D2=Dijkstra(V_i,CoreV(c_k'))$。

Step4　若$D1<D2$，则$V_i\in c_j$；或者$D1>D2$，则$V_i\in c_k$。

特别地，当$CoreV(c)$有多个时，有两种决策来计算D值：1）最近中心点；2）平均距离。需要采用哪种决策，则根据实际情况和需求来确定。

在上述方案执行完毕后，再次扫描更新后的聚类结果，找出新聚类结果中的目标点（已处理过的点不再扫描）。

因此，提出的算法如下：

1）获得复杂网络的邻接矩阵并代入FN算法，获得层次树图。

2）利用模块度Q选择层次树图的最佳截断平面来获得聚类结果。

3）扫描聚类结果中具有所提到特性的点。

4）将扫描到的点从社团中独立出来（每次只独立1个）。

5）根据需求，选择3种方案中的一种进行计算，重新划分其归属聚类。

6）重复执行4），直到所有的点处理完毕。

7）重复执行3），直到扫描不到未处理过的点。

伪代码如下：

```
adjacent_matrix：复杂网络的邻接矩阵
Fast-Newman()：FN算法函数
GetModularity()：获得最佳阶段面（即获得聚类结果）
Scan()：扫描函数
Queue：存储目标点队列
PostP()：后处理函数
1. Tree=FastNewman(adjacent_matrix)
2. R=GetModularity(Tree)
3. Scan(R) to Queue
4. while(Queue!=empty)
5. s=get first element from Queue
6. NewR=PostP(adjacent_matrix,R,s,n) //n=1,2,3 分别对应3种方案
7. delete s from Queue
8. Scan(NewR) to Queue
9. output NewR
```

5.2　实验结果

以上文所提到的空手道俱乐部网络为例进行实验。根据所观察的错误点的特性，找到了点V_{10}以及将其独立后的两个社团：

$c_1'=\{1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22\}$

$c_2=\{9,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34\}$

方案1

当$P=2$时，$T1=\sum_{V_j\in c_1'}(d(V_{10},V_j)=P)=5$，$T2=\sum_{V_k\in c_2}(d(V_{10},V_k)=P)=13$。

因此，$V_{10}\in c_2$，这与实际分类情况相同。通过修改之后社团划分准确度达到100%。

方案2

计算可得$D1=\sum_{V_j\in c_1'} Dijkstra(V_{10},V_j)=42$，$D2=\sum_{V_k\in c_2} Dijkstra(V_{10},V_k)=35$，$D1>D2$。因此，$V_{10}\in c_2$，这与实际分类情况相同。通过修改之后社团划分准确度达到100%。

方案3

通过计算，得到$CoreV(c_1)=V_1$以及$CoreV(c_2)=V_{34}$。$D1=Dijkstra(V_{10},V_1)=2$以及$D2=Dijkstra(V_{10},V_{34})=1$，可得$D1>D2$。因此，$V_{10}\in c_2$，这与实际分类情况相同。通过修改之后社团划分准确度达到100%。

根据以上3种解决方案的计算结果将V_{10}按照算法重新划分后，聚类结果变为：

$c_1=\{1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22\}$

$c_2=\{9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34\}$

第一轮处理完毕后，通过扫描行的聚类结果发现，V_3也符合我们所观测的错误分类点的特点，尽管在这里并没有划分错误。对此，依然按照算法进行计算，之后得到的结果并未改变，这说明算法并没有划分错误，也说明在未知时候划分正确时，我们提出的算法可以将具有本文中所讨论的特点的点划分到正确的聚类中。再次扫描，除了V_3与V_{10}外，并无其他符合该特征的点，故处理结束。此时，虽然模块度Q的值从0.371795轻微下降到0.371466，但准确率为100%。

将解决方案同样作用于锯木厂员工关系网[26]和海豚关系网[27]，来验证本文所提出的后处理方案的可行性与普遍性。

首先对锯木厂员工关系网进行实验，其关系网如图6所示。

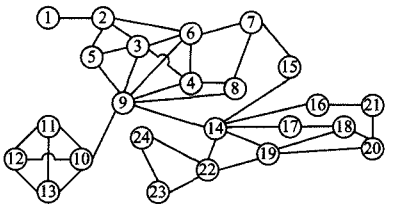

图 6　锯木厂员工关系网

将该网的邻接矩阵代入 FN 算法，结果如表 1 所列。

表 1　锯木厂员工关系网的实验结果

聚类	原始结果	FN 算法结果	错误点	缺失点
1	1,2,3,4,5,6,7,8,9	1,2,3,4,5,6,7,8,9,15	15	无
2	10,11,12,13	10,11,12,13	无	无
3	14,15,16,7,18,19,20,21,22,23,24	14,16,7,18,19,20,21,22,23,24	无	15

FN 算法结果的准确率为 95.8%，模块度为 Q=0.545。

对算法结果进行扫描与处理。通过扫描，发现点 15 属于本文提及的错误点类型。将该点按照上文的 3 种方案进行，处理结果表明，3 种方案都能正确地将点 15 划分到正确的聚类中，由于计算过程类似上文所述，不再重复。点 15 重新划分后，聚类的准确率达到了令人满意的结果，即 100%，模块度为 Q=0.5475，略有上升。

接下来对海豚关系网进行实验，其关系网如图 7 所示。

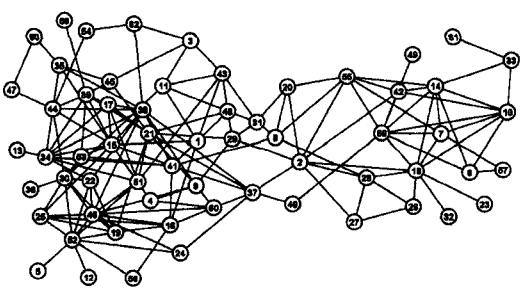

图 7　海豚关系网

将该网的邻接矩阵代入 FN 算法，结果如表 2 所列。

表 2　海豹关系网的实验结果

聚类	原始结果	FN 算法结果	错误点	缺失点
1	2,6,7,10,14,18,19,20,23,26,27,28,32,33,42,49,55,57,58,61	2,6,7,8,10,14,18,20,23,26,27,28,29,31,32,33,42,49,55,58,57,61	8,29,31	19
2	1,3,4,5,8,9,11,12,13,15,16,17,21,22,24,25,29,30,31,34,35,36,37,38,39,40,41,43,44,45,46,47,48,50,51,52,53,54,56,59,60,62	1,3,4,5,9,11,12,13,15,16,17,19,21,22,24,25,30,34,35,36,37,38,39,40,41,43,44,45,46,47,48,50,51,52,53,54,56,59,60,62	19	8,29,31

FN 算法结果的准确率为 93.5%，模块度为 Q=0.3854。

接下来，对算法结果进行扫描与处理。

第一轮扫描，会发现点 29。按照 3 种方案进行处理，能够正确地将点 29 划分给 2 号聚类。计算过程同上，不再复述。第二轮扫描，可将点 31 正确划分。第三轮则可将点 8 正确划分。第四轮扫描后无该特征点。算法结束。

至此，成功地将 3 个被错误划分的点即 8,29,31 通过本文所提出的 3 类方案成功地划分到正确的聚类中，处理之后，算法结果的准确率从 93.5%提高到了 98.4%，只有 1 个点未正确归类，由于 19 号点的错误类型并不在讨论范围内，故没有将其正确归类；而模块度为 Q=0.3534，略有下降。

结束语　本文针对 FN 算法结果中所出现的错误划分情况进行了研究。发现，由于 FN 算法的特性使得点被倾向于划分给较小的聚类，其结果也会偏向于形成几个大小相似的社团结构。此外，通过实验发现准确率提高时，模块度 Q 的值在大部分实验中会轻微下降，只有部分会上升，这说明模块度 Q 所的假设并不全对，更大的模块度 Q 并不意味着聚类一定更好，而是在一定层面上来说，获得较大的模块度 Q 的聚类要优于获得较小的模块度 Q 的聚类。

基于本文所提出的 3 种后处理方案，都能够将本文中所描述的 FN 算法错误划分的正确划分到社团中，提高了结果准确率。在多个实验中，对于顶点数较少的例子，后处理方法能使得精确度达到 100%，结果令人满意。

本文针对 FN 进行了研究，提出了优化结果的方案。同时也发现这种错误点类型也出现在 GN 算法和 CPM 算法中，并且经过所提出的后处理方案的处理，这一类型的错误点能够被正确地划分到与实际归属相同的社团，证明该方案可以延伸到 GN 算法与 CPM 算法中来进行优化处理。这表明了所提出的后处理优化方案具有一定的移植性。

参 考 文 献

[1] Xie Zhou, Wang Xiao-fan. An Overview of Algorithms for Analyzing Community Structure in Complex Networks[J]. Complex Systems and Complexity Science, 2005, 2(3): 1-12

[2] Gibson D, Kleinberg J, Raghavan P. Inferring web communities from link topology [C]// Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. 1998: 225-234

[3] Flake G W, Lawrence S R, Giles C L, et al. Self organization and identification of web communities [J]. IEEE Computer, 2002, 35(3): 66-71

[4] Adamic A L, Adar E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3): 211-130

[5] Shen-Orr S, Milo R, Mangan S, et al. Network motifs in the transcriptional regulation network of Escherichia coli [J]. Nature Genetics, 2002, 31(1): 64-68

[6] Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: simple building blocks of complex networks [J]. Science, 2002, 298(5594): 824-827

[7] Holme P, Huss M, Jeong H. Sub network hierarchies of biochemical path ways[J]. Bioinformatics, 2003, 19(4): 532-538

[8] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci, 2001, 99(12): 7821-7826

[9] Gleiser P, Danon L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6(4): 565-573

[10] Garey M R, John son D S. Computers and Intractability: A Guide to the Theory of NP-Completeness [M]. San Francisco: W. H. Freeman Publishers, 1979

[11] Scott J. Social Network Analysis: A Handbook(2nd ed)[M]. London: Sage Publications, 2002

[12] Wang Xiao-fan, Liu Ya-bing. Overview of Algorithms for Detecting Community Structure in Complex Networks[J]. Journal of

Electronic Science and Technology Photonic Sensors, 2009, 38 (5):537-543

[13] Newman M E J. Random graphs with clustering [J]. Phys Rev Lett, 2009, 103:058701

[14] Newman M E J. Detecting community structure in networks [J]. Eur Phys J B, 2004, 38(2):321

[15] Sales-Pardo M R, Guimera A, Moreira A, et al. Extracting the hierarchical organization of complex systems [J]. Proc Natl Acad Sci USA, 2007, 104:15224

[16] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of community hierarchies in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10:10008

[17] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A, 2009, 388: 1706-1712

[18] Zhang S, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering [J]. Physica A, 2007, 374:483-490

[19] Zhang S, Wang R S, Zhang X S. Uncovering fuzzy community structure in complex networks [J]. Phys Rev E, 2007, 76(4): 046103

[20] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Phys Rev E, 2004, 69(2):026113

[21] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev EstatNonlin Soft Matter Phys, 2004, 69(6):06133

[22] Kernighan B W, Lin S. A efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2): 291-307

[23] Fiedler M. Algebraic connectivity of graphs [J]. Czech Math J, 1973, 23(98):298-305

[24] Pothen A, Simon H, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM J Matrix Anal Appl, 1990, 11 (3):430-452

[25] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33:452-473

[26] Michael J H. Labor dispute reconciliation in a forest products manufacturing facility[J]. Forest Products Journal, 1997, 47:41-45

[27] Lusseau D. The emergent properties of a dolphin social network [J]. The Royal Society, 2003, 270(Suppl):186-188

(上接第 232 页)

[11] Ju Sheng-feng, Wang Zhong-qing, Li Shou-shan, et al. A Comparative Study on Different Active Learning Strategies for Sentiment Classification[C]// Advances of Computational Linguistics in China(2009-2011). 2011:506-511

[12] Song H, Yao T. Active Learning Based Corpus Annotation [C] // IPS-SIGHAN Joint Conference on Chinese Language Processing. Beijing, China, 2010:28-29

[13] Long Jun, Yin Jian-ping, Zhu En, et al. An Active Learning Algorithm by Selecting the Most Possibly Wrong-Predicted Instances[J]. Journal of Computer Research and Development, 2008, 4(3):472-478

[14] Church K W, Hanks P. Word association norms, mutual information and lexicography[J]. Computational linguistics, 1990, 16 (1):22-29

[15] Liu Bin, Huang Tie-jun, Cheng Jun, et al. A New Statistical-based Method in Automatic Text Classification[J]. Journal of Chinese Information Processing, 2002, 16(6):18-24

[16] Zhu xiao-juan. The Researeh on Chinese Word Segmentation System Based on SVM[D]. Central South University, 2007

[17] Yan Hui, Zhang Xue-gong, Li Yan-da. Kenal-based maximal-margin clustering algorithm[J]. Journal of Tsinghua University (Natural Sciences), 2002, 42(1):36-38

[18] Dai Y, Loh T E, Khoo C S G. A new statistical formula for Chinese text segmentation incorporating contextual information[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999:82-89

[19] Cover T, Hart P. Nearest Neighbor Pattern Classification[J]. IEEE Trans on Information Theory, 1967, 13(1):21-27

[20] Zhao Ying, Liu Hong-xing, Wang Zhong-yu, et al. An Improved Nearest Neighbor Searching Method for Classification Problems [J]. Journal of Nanjing University: Natural Sciences, 2009, 45 (4):455-462

[21] Wang Zhen-yu, Wang Xi-zhao. Active Learning Algorithm Based on Neighborhood Entropy[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1):97-102

[22] Carter T. An introduction to information theory and entropy [EB/OL]. http://astarte. csustan-edu/tom/ SFICSSS

[23] Bai Long-fei, Wang Wen-jian, Guo Hu-sheng. A noval Support Vector maching Active Learning strategy[J]. Journal of Nanjing University: Natural Sciences, 2012, 48(2):182-189

[24] http://crfpp. sourceforge. net/

[25] Qiu Sha, Wang Fu-yan, Shen Hao-ru, et al. Chinese Named Entity Recognition Based on Part of Speech Feature with Edges [J]. Computer Engineering, 2012, 38(13):128-130

[26] Li Shou-shan, Huang Hu-ren. Word Boundary Decision with CRF for Chinese Word Segmentation [C]//23th PACLIC. 2009: 726-732

[27] Xiao Qin, Liang Zong, Wu Yu-qian, et al. CRF-based Experiments for Cross-Domain Chinese Word Segmentation at CIPS-SIGHAN-2010 [C]//Proceedings of CLP 2010. 2010

[28] Zhong Ke-li, Zhou Xue, Li Hang-yu, et al. Cascaded Chinese Weibo Segmentation Based on CRFs[C] // Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China, 2012:69-73