文章编号：1007-791X (2014) 05-0440-06

基于关系强度的复杂网络社团结构发现算法比较

张 涛[1,*]，魏昕宇[1]，唐 迪[2]

（1. 燕山大学 信息科学与工程学院，河北 秦皇岛 066004；2. 中国联合网络通信有限公司秦皇岛市分公司，河北 秦皇岛 066004）

摘 要：实际复杂网络数据的合理预处理是社团结构分析的基础与前提，对划分结果有着显著的影响。本文通过定义节点间的关系强度，将社会调查得到的社交网络数据处理成社团结构算法中通用的邻接矩阵，提出了处理实际复杂网络数据使之适用于经典社团算法分析的一种方法。同时，本文进一步给出了关系强度强联系定义和关系强度弱联系定义，并比较了算法在这两种定义下展示出的不同性能。

关键词：关系强度；社团结构；复杂网络

中图分类号：TP319.40 **文献标识码**：A DOI：10.3969/j.issn.1007-791X.2014.05.011

0 引言

复杂网络理论是研究各类实际复杂网络的普适方法，而复杂网络社团结构分析又是复杂网络研究的三大重点研究内容之一。在大数据时代，复杂网络的规模以前所未有的速度扩大，节点类型更加复杂。如何发现规模庞大，结构复杂，数据类型多样的复杂网络下的社团结构成为重要的问题。

目前社团结构划分算法已经比较成熟，形成了以 GN 算法[1] 为代表的分裂算法，以 K-L 算法[2] 为代表的贪婪算法，以 Newman 快速算法[3] 为代表的凝聚算法，以基于 Laplace[4] 或 Normal[5] 矩阵的谱平分算法为代表的谱分析算法这四大类经典算法。除以上四类经典算法外，比较实用的算法还有信息中心度算法[6]。

近年来，在这 4 种算法的基础上，涌现出了大量社团结构划分算法：GN 算法的经典改进算法[7-8]，GN 算法的并行化[9]；FN 算法的改进算法[10]；以聚类思想为核心的新算法[11-14] 等。其着眼点都是寻找一种改进 4 类基本算法缺陷的有效手段，以提高算法划分的准确度或运算速度。

复杂网络由节点和节点之间的边构成，从这两个角度出发，学者们提出了基于节点特性的社团结构算法[15-19] 和基于边特性的社团结构算法[20-22]。以节点或边的某一种度量值为依据，逐步探寻出网络的社团结构。但是，目前社团划分算法的主要分析以 karate 俱乐部模型数据和计算机仿真网络数据为代表的标准网络。在实际应用中，将组织结构复杂、数据量庞大、形式繁杂的实际网络处理成可应用于各类社团划分算法计算的标准网络，尚没有简单有效的方法。这正是本文的研究重点。

本文给出了关系强度定义，并在定义条件下处理了一组真实社交网络数据，同时利用社团结构划分的经典算法对获得数据进行分析、比较，详细分析了 4 类经典算法在两种定义条件下的性能差异。

1 社团结构基础理论

1.1 社团结构的定义

社团结构划分首先需要明确社团的数学定义，而这一直是社团分析研究中的一个难题。

社团结构的定性定义为：每个社团内部节点之间的连接相对非常紧密，但是社团之间的联系比较稀疏[23]。这一定义被广泛认可和引用，但是如何

收稿日期：2014-07-02 **基金项目**：国家自然科学基金资助项目 (61273019)；河北省自然科学基金资助项目（F2013203368）；首批“河北省青年拔尖人才”资助项目

作者简介：*张 涛 (1979-)，男，河北唐山人，博士，副教授，主要研究方向为信息融合、可视化模式识别、图像处理，Email：zhtao@ysu. edu. cn。

以数学形式给出社团结构的标准定义一直存在争议。本文采用比较流行的 Radchicchi 等人给出的社团结构定义。

1）强社团定义：如果对任意节点i，子网络V满足

$$k_i^{in}(V)>k_i^{out}(V),\ \forall i\in V, \tag{1}$$

则称V为网络的强社团结构。

2）弱社团定义：如果子网络 V 满足

$$\sum_{i\in V}k_i^{in}>\sum_{i\in V}k_i^{out}, \tag{2}$$

则称V是网络的弱社团结构。

1.2　社团结构算法概述

社团结构是复杂网络的自然属性，所有的实际复杂网络中都存在着不同形式的社团结构。社团结构划分算法的目的就是设法给出一种最符合网络中真实社团结构的社团结构划分结果。

GN 算法定义通过每条边的最短路径数目为该边的边介数，并以此概念为核心，采用分裂算法的思想，逐次移除边介数最大的边，并重新计算移除后网络中每一条边的边介数。重复这一过程，直到网络分裂成一个个孤立的社团。自包含 GN 算法[24]是 GN 算法的一种改进算法，其给出了强社团结构定义和弱社团结构定义，并在 GN 算法的划分结果中挑选符合定义的形式作为最终的划分结果。

K-L 算法和极值优化算法都是基于贪婪算法原理设计的。K-L 算法首先随机将网络划分成两个社团，并定义社团内部的边减去社团之间的边数所得的差值为目标函数Q，从两个社团内各取一个节点构成节点对，假设将此节点对内的两个节点社团归属对调，计算Q函数的增加值。遍历所有可能的节点对，得到能使Q函数增加最多的节点对，将这对节点社团归属对调，并不断重复这一过程。Q函数取得极值的划分形式作为社团划分结果。极值优化函数的思想与此类似，最大的不同是极值优化函数以模块度最大的划分形式作为最终的社团划分结果。

谱平分算法首先构建复杂网络的拉普拉斯矩阵，并求取其特征值和和特征向量。通过特征向量的正负来判断社团归属。

1.3　社团结构划分的评价标准

不同的社团结构划分算法会对同一网络给出不同的划分形式，因此需要一种评价各类算法性能的标准，目前最为通行的标准是 Newman 等人提出的 NG 模块度函数。

模块度函数是基于协调混合（associative mixing）[25] 定义的。首先以某种形式将网络划分为k个社团，定义一个$k\times k$维的对称阵$\boldsymbol{E}=(e_{ij})$，其中e_{ij}表示社团i到社团j之间的边数与原始复杂网络中总边数的比值。如果$i=j$，e_{ij}表示社团i内部的边数与原始复杂网络中的总边数的比值。需要注意的是，对称阵$\boldsymbol{E}$中所有元素的和为 1。

在得到对称阵$\boldsymbol{E}$之后，定义对角线上各元素之和为

$$tr(\boldsymbol{E})=\sum_i e_{ii}, \tag{3}$$

其物理意义为网络中社团内部边占原始复杂网络中边数的比例。

定义每行（或每列）元素之和为

$$a_i=\sum_j e_{ij}, \tag{4}$$

其物理意义为与第i个社团中的节点相连的边在所有边中所占比例。

明确了以上定义，下面给出模块度函数的定义公式

$$Q=\sum_i(e_{ii}-a_i)^2=tr(\boldsymbol{E})-\|e^2\|, \tag{5}$$

其中，$\|\cdot\|$表示矩阵中所有元素之和。Q的取值范围为 0 到 1，实际网络中一般取 0.3 到 0.7 之间。模块度函数的另一个等价公式为

$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,\ C_j), \tag{6}$$

其中，m为网络中的总边数，$\boldsymbol{A}$为网络的邻接矩阵。

一般来说，模块度值较大的划分结果更符合实际情况。

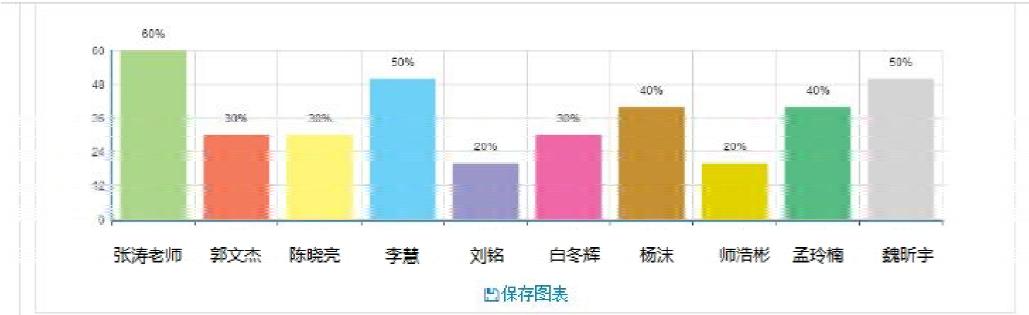
2 复杂网络数据获取与预处理

2.1 课题小组社交网络调查

为了获得实际的社交网络数据，本文详细调查了一个课题小组的社交网络。在调查中，每个受调查人需要报告在过去的一周内是否与其他受调查人联系过，联系的表现形式包括电子邮件、电话、面谈等。所有调查数据均严格保密，以消除其他干扰，确保数据客观真实。

在本次调查中，错误填写问卷的受调查人被要求重新填写问卷以确保数据完整、准确。需要指出的是，在实际的复杂网络数据分析中，由于原始数据的规模一般都比较大，错误的数据通常很少，因此可以直接舍弃不必重新调查。为了方便给出复杂网络社团结构的标准划分形式，本次调查采用了一个规模较小的复杂网络。

在调查中，有一种特殊情况：即被调查人甲认为其与被调查人乙在调查期间联系过，而乙认为两人没有联系。这为判定两个结点间是否有连接带来了困惑，为了解决这个问题，本文提出了社交网络联系强度定义。调查问卷的页面如图 1 所示。



选项	数据里	百分比
张涛老师	6	60%
郭文杰	3	30%
陈晓亮	3	30%
李慧	5	50%
刘铭	2	20%
白冬辉	3	30%
杨沫	4	40%
师浩彬	2	20%
孟玲楠	4	40%
魏昕宇	5	50%

回答 **10** (100%)

图 1 调查结果网络页面

Fig. 1 Web of survey results

2.2 关系强度定义

定义 1 关系强度强联系定义。在这个定义中，只有当两个被调查都认为他们联系过的时候才认为联系有效。该定义的数学表达为

$$A_{ij}=\begin{cases}1 & A_{ij}=A_{ji}=1\\ 0 & A_{ij}=0 \text{ 或 } A_{ji}=0\end{cases}, \tag{7}$$

其中，$\boldsymbol{A}$表示社交网络，1 表示联系存在，0 表示联系不存在。

定义 2 关系强度弱联系定义。在这个定义中，只要一个受调查人认为两人曾经联系过，就认为联系有效。仿照强联系定义的形式，给出社交网络的弱联系定义:

$$A_{ij}=\begin{cases}0 & A_{ij}=A_{ji}=0\\ 1 & A_{ij}=1 \text{ 或 } A_{ji}=1\end{cases}。 \tag{8}$$

利用这两个定义，可以得到两个不同的社交网络初始化矩阵，分别称作强联系矩阵和弱联系矩阵。在实验部分，很容易注意到，在不同的关系强

度定义条件下，算法的性能发生了显著地变化。

3　试验结果与分析

3.1　调查问卷整理

综合分析所得的 10 份调查问卷，整理出课题组社交网络的原始调查数据如表 1。

表 1　社交网络原始数据集

Tab. 1　Original data of social network

	1	2	3	4	5	6	7	8	9	10
1		1	1	1	1	1	1	1	1	1
2	1			1			1			
3						1			1	1
4		1					1			
5	1								1	
6	1		1							
7	1	1		1						1
8	1			1					1	1
9			1			1				1
10	1			1			1	1		

表 1 中，每一行或每一列代表一个受调查人，分别用序号 1~10 表示。表格中的元素为 1 表示在问卷假设下两人存在联系，为空则表示联系不存在。以第 1 行为例，受调查人 1 与受调查人 2~10 均存在联系。

3.2　关系强度强联系定义下的初始化

采用关系强度强联系定义处理表 1 中得到的原始数据，得到社交网络的邻接矩阵如表 2 所示。

表 2　关系强度强联系定义下的网络初始化数据

Tab. 2　Initialized network data under strict definition

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	1	1	1	1	0	1
2	1	0	0	1	0	0	1	0	0	0
3	0	0	0	0	0	1	0	0	1	0
4	0	1	0	0	0	0	1	0	0	0
5	1	0	0	0	0	0	0	0	0	0
6	1	0	1	0	0	0	0	0	0	0
7	1	1	0	1	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	1
9	0	0	1	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	1	0	0

此时，可以使用各类社团结构算法对得到的社交网络数据进行社团结构分析，以 GN 算法为例，经过社团划分，课题小组分裂为两个小的社团，如图 2 所示。

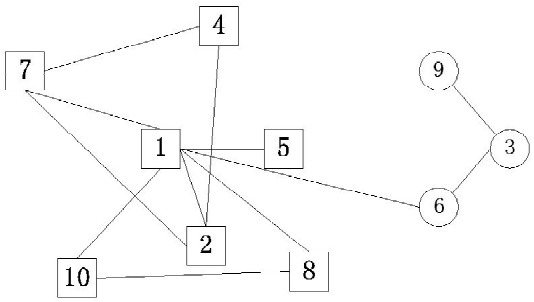


图 2　关系强度强联系定义下 GN 算法划分结果

Fig. 2　Community detecting results under GN and strict definition

从图中可以明显地看出，节点 3、6、9 成为一个独立的社团，节点 3、6、9 组成第二个社团。节点 1 到节点 6 之间的边为两个社团之间唯一的联系。

3.3　关系强度弱联系定义下的初始化

在关系强度弱联系定义下，利用相似的处理收单，将社交网络原始数据被初始化成的邻接矩阵，如表 3 所示。其中每一行或一列的起始元素表示一个节点，表中元素有 0 和 1 两种取值。在拓扑结构图上，0 表现为节点间没有边，1 表示节点间存在一条无权无向的边。

表 3　关系强度弱联系定义下的网络初始化数据

Tab. 3　Initialized network data under loose definition

	1	2	3	4	5	6	7	8	9	10
1	0	1	1	1	1	1	1	1	1	1
2	1	0	0	1	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	1	1
4	1	1	0	0	0	0	1	1	0	1
5	1	0	0	0	0	0	0	0	1	0
6	1	0	1	0	0	0	0	0	1	0
7	1	1	0	1	0	0	0	0	0	1
8	1	0	0	1	0	0	0	0	1	1
9	1	0	1	0	1	1	0	1	0	1
10	1	0	1	1	0	0	1	1	1	0

同样采用 GN 算法进行社团结构划分，在定义条件下，得到划分结果如图 3 所示。

课题组社交网络被明显地划分为两个社团，节点 5 独自构成一个社团，网络中的其他节点构成另外一个社团。值得注意的是，同一种算法，在关系强度强联系定义和弱联系定义的条件下取得的划分结果明显不同。变化的原因是关系强度强联系定义下，许多低效的联系被过滤。

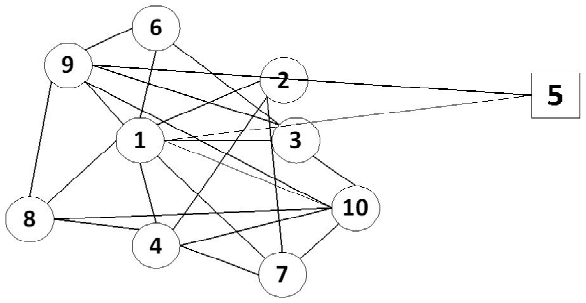


图 3 关系强度弱联系定义下 GN 算法划分结果

Fig. 3 Community detecting results under GN and loose definition

3.4 基于关系强度定义下的算法比较

首先，为了计算算法社团划分的准确率，根据受调查人的研究方向和所处实验室的不同，给出标准的社团划分形式，节点 1、2、4，7、8、10 构成一个社团，节点 3、5、6、9 构成第二个社团，如图 4 所示。因为节点 1 所代表的对象同时为两个实验室的负责人，与两个实验室的成员联系密度相当，也认为把节点 1、3、5、6、9 划分为一个社团也是合理的。

根据给出的两种标准划分形式，可以计算出每种算法划分结果的正确率。其计算方法为正确划分社团的节点数与网络中节点总数的比值。

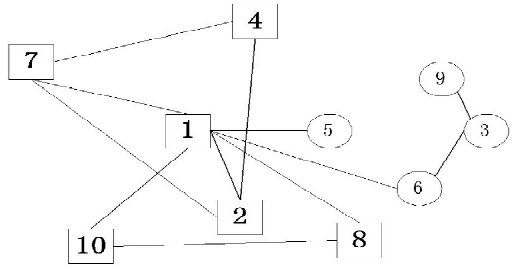


图 4 一种标准社团划分形式

Fig. 4 A standard community structure detecting results

假设已知网络可以划分为两个社团，对于 K-L 算法补充假设社团大小分别为 5，5。对于 6 种算法全部分社交网络强联系和弱联系两种初始化情况讨论。其中自包含 GN 算法需要分别采取社团结构强定义和社团结构弱定义进行分析。需要注意的是社团结构的强弱定义与社交网络的联系强度时两个不同的概念，相关定义算法对比情况如表 4 所示。

表 4 不同算法对比

Tab. 4 Comparison of different algorithms

算法名称	初始化方法	划分错误的节点	划分正确率	模块度	备注
GN	强联系	5	90%	0.2465	—
	弱联系	无	100%	0.2628	剔除核心节点后
强社团定义自包	强联系	—	—	—	无有效划分结果
含 GN	弱联系	无	100%	0.2628	剔除核心节点后
弱社团定义自包	强联系	5	90%	0.3368	—
含 GN	弱联系	—	—	—	无有效划分结果
K-L 算法	强联系	5，8，10	70%	0.2188	—
	弱联系	无	100%	0.1512	—
谱平分算法	强联系	5	90%	0.2465	—
	弱联系	无	100%	0.1352	—
信息中心度算法	强联系	5	90%	0.2465	—
	弱联系	—	—	—	无有效划分结果
极值优化算法	强联系	8，10	80%	0.1248	—
	弱联系	8，10	80%	0.2778	—

首先，需要强调的是，GN 算法的关系强度弱联系定义划分结果和自包含 GN 算法的关系强度强联系定义划分结果是在移除节点 1 的情况下计算得到的。这是因为节点 1 的度数非常大，与网络中每个节点的联系都很紧密，在不移除的情况下爱这两种算法得到的结果没有意义，所以移除干扰划分的节点 1 以获得一种社团划分。由于选择性剔除了一个节点，改变了实验条件，所以这两种情况不参加后面的分析。本文列举出这两种情况只是为了提示在网络中存在度数远远大于其他节点的超级节点时，移除这个节点，有可能会发现潜在的社团结构。例如，在本实验中，如果节点 1 不存在，网络会迅速分裂成两个孤立的社团。

其次，在关系强度强联系定义条件下 GN 算

法、信息中心度算法获得了弱联系定义下更高的准确率。在关系强度弱联系定义下，K-L 算法、谱平分算法取得了更好的划分结果。而极值优化算法对两种定义不敏感。

再次，需要注意的是，K-L 算法和极值优化算法必须给出一个初始的划分形式。由于划分是随机的，所以划分形式的改变可能会引起最终算法的准确率。这是实验中的一个不可控因素。

4　结束语

目前，复杂网络社团结构划分的算法设计和应用已经成为研究热点。算法本身的研究日渐成熟，但是算法设计使用的数据集很有限，主要集中于 karate 俱乐部模型和计算机仿真网络，而且经常忽视设计一种把实际复杂网络处理成标准邻接矩阵的合理方法。本文通过关系强度强联系定义和弱联系定义给出了一种处理实际复杂网络数据的方法，并在定义条件下分析了算法性能的变化。例如，在 K-L 算法下，选择弱联系定义可以调高 30%的正确率。合理地选择关系强度强联系定义和弱联系定义可以提高社团划分的正确率。

参考文献

[1] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2001,99 (12): 7821-7826.

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

[3] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004,69 (6): 066133.

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

[5] Capocci A, Servedio V D P, Caldarelli G, et al.. Detecting community in large networks [J]. Computer Science, 2004,3243: 181-187.

[6] Fortunato S, Latora V, Marchiori M. A method to find community structures based on information centrality [J]. Physical Review E, 2004,70: 056104.

[7] Tyler J, Willkinson D, Huberman B. Email as spectroscopy: Automated discovery of community structure within organizations [M]. International Conference on Communities and Technologies, 2003: 81-96.

[8] Radicchi F, Castellano C, Cecconi F, et al.. Defining and identifying communities in networks [J]. Proc. Natl. Acad. Sci. , 2004,101: 2658-2663.

[9] 杨立文. 基于改进的 GN 算法的社区发现技术 [D]. 长春: 吉林大学, 2012.

[10] 肖有诰, 屠成宇. 基于启发式函数的分布式 FN 算法 [J]. 计算机系统应用, 2012,21 (10): 122-125.

[11] 沙爱晖, 黄树成, 李甜. 一种基于网络社团结构和模块化函数的聚类方法 [J]. 计算机应用与软件, 2014,31 (4): 274-277.

[12] 姜荣, 赵风霞, 谢福鼎, 等. 一种基于 Normal 矩阵的时间序列聚类方法 [J]. 计算机应用研究, 2010,27 (8): 2926-2928.

[13] 袁辉辉, 曹玉林, 王小明. 基于边聚类的多层社会网络社团发现算法 [J]. 计算机应用研究, 2014,31 (2): 351-353,377.

[14] 杜守印, 李斌, 董传杰, 等. 基于相似性复杂网络社团结构发现算法研究 [J]. 软件, 2014,35 (2): 70-74.

[15] 朱凤辉, 樊瑛. 基于电阻网络的节点重要性判别 [J]. 北京师范大学学报（自然科学版）, 2013,49 (6): 636-639.

[16] 赵之滢, 于海, 朱志良, 等. 基于网络结构的节点传播影响力分析 [J]. 计算机学报, 2014,37 (4): 753-766.

[17] 钟芬芬. 复杂网络社区发现算法研究 [D]. 西安: 西安电子科技大学, 2012.

[18] 郭强. 一种基于节点分裂的重叠社区发现算法 [D]. 哈尔滨: 哈尔滨工程大学, 2013.

[19] 罗明伟, 姚宏亮, 李俊照, 等. 一种基于节点相异度的社团层次划分算法 [J]. 计算机工程, 2014,40 (1): 275-279.

[20] 贾春旭. 复杂网络社团发现算法的研究及其应用 [D]. 兰州: 兰州理工大学, 2012.

[21] 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.

[22] Fortunato S, Latora V, Marchiori M. A method to find community structures based on information centrality [J]. Physical Review E, 2004,70 (5): 056104.

[23] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用 [M]. 北京: 清华大学出版社, 2006.

[24] Radicchi F, Castellano C, Cecconi F, et al.. Defining and identifying communities in networks [J]. Proceedings of the National Academy of Sciences, 2004,101 (9): 2658-2663.

[25] Newman M E J. Mixing patterns in networks [J]. Physical Review E, 2003,67 (2): 026126.

（下转第 454 页）

[13] 户菲菲, 张介眉. 论葱白的通阳作用 [J]. 现代中西医结合杂志, 2009,18 (20): 2444.

[14] 吴谦. 医宗金鉴·订正仲景全书伤寒论注 [M]. 北京: 中国中医药出版社, 1998: 106.

New discovery on anti-diarrhea effect of fistular onion stalk in treatise on febrile diseases based on formal concept analysis

ZHANG Jue[1], LIU Min[2], XU Sun-jing[1]

(1. Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510000, China; 2. The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, Guangdong 510000, China)

Abstract: Based on the formal concept analysis theory and the study on the classical books in traditional Chinese medicine such as Treatise on Febrile Diseases, the research finds out the inner correlation between Baitong Tang and Ganjian Fuzi Tang and explores the anti-diarrhea mechanism of Fistular Onion Stalk, which demonstrates the broad prospect of formal concept analysis in the age of big data. It is found that the application of formal concept analysis to visual data mining in traditional Chinese medical system will provide a reliable technological guarantee for the discovery of the prescription, formula and syndrome, and the indication of medicine in TCM.

Key words: big data; formal concept analysis; Treaties on Febrile Diseases

(上接第 445 页)

Comparisons among complex network community algorithms based on relationship strength

ZHANG Tao[1], WEI Xin-yu[1], TANG Di[2]

(1. School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China; 2. China United Network Communications Corporation Qinhuangdao Branch, Qinhuangdao, Hebei 066004, China)

Abstract: The data pre-processing of actual complex network is an important basis of researching community structure detecting algorithm, which has a significant influence on the accuracy of algorithm. A new method of process is proposed to transmit the original network data into standard adjacent matrix through the relationship strength definition. Several classical community detecting algorithms under the definition are analyzed.

Key words: relationship strength; community structure; complex network