一种基于多维遗传算法的重叠社区发现方法*

王 琦[1,2]，温志平[1]
(1. 南京工程学院 计算机工程学院，南京 211167; 2. 南京大学 计算机科学与技术系，南京 210093)

摘 要： 社区结构的发现是社交网络分析研究的重要内容。与传统的重叠社区不同，最近的研究表明某些真实网络中在社区重叠部分要比社区内部节点间的连接更加密集，而现有的算法没有考虑此类社区结构。基于遗传算法，提出了一种新颖的方法来发现此类社区划分。为了刻画节点属于多个社区的重叠现象，首次将多维染色体和均匀块交叉算子引入到社区发现算法中。通过实验证明，提出的算法可以很好地发现社交网络中重叠和非重叠的社区结构。

关键词： 社团发现; 重叠社团; 多维染色体; 从属网络
中图分类号： TP393 **文献标志码：** A **文章编号：** 1001-3695(2016)12-3543-04
doi:10.3969/j.issn.1001-3695.2016.12.006

Multidimensional genetic algorithm for overlapping community detection

Wang Qi[1,2], Wen Zhiping[1]
(1. *College of Computer Engineering, Nanjing Institute of Technology, Nanjing* 211167, *China*; 2. *Dept. of Computer Science & Technology, Nanjing University, Nanjing* 210093, *China*)

Abstract: Community structure identification is an important content of social network analysis. In contrast to traditional definitions of overlapping network community, recent studies have found that overlaps between communities are more densely connected than the non-overlapping parts which are common in real social structures, and existing methods do not consider this kind of community structure. This paper developed an innovative algorithm for detecting dense overlapping communities based on genetic algorithm. In order to characterize the real situation of the nodes belonging to multiple communities, it first introduced a new multidimensional chromosome and block-uniform crossover in community discovery algorithms. It performed several experimental studies to demonstrate that this method successfully captures overlapping as well as non-overlapping communities.
Key words: community detection; overlapping community; multidimensional chromosome; affiliation networks

现在有大量的社交网络如 Facebook、Twitter 和新浪微博等。在社交网络中，人们可以通过社区相互联系和交换信息来展示个人生活。社区结构的发现有助于捕获和跟踪网络的拓扑结构、揭示复杂系统内在的功能特性、预测个体关系和行为的演化趋势，具有很强的应用价值。社区发现算法的研究得到了研究人员的广泛关注，很多经典算法如模块度、介数等被提出来用于有效快速地挖掘不同规模的社区[1~3]。

在很多社交网络中，一个人往往同时属于多个群体或参与多个话题，这就提出了重叠社区发现的问题。现在的社区发现算法都隐含一个共同的假设，认为社区是网络节点集合的若干子集，每个子集内部节点之间的连接相对紧密，而不同子集节点之间的连接相对稀疏[4]。换句话说，这意味着一对节点同时属于的社区越多，它们之间相连的可能性就越低。然而最近 Yang 等人[5]通过对某些真实社区(ground-truth communities)(图 1)的研究发现，一些社交网络中，节点之间在社区间重叠部分的连接往往比社区内部的连接更加紧密[6]。这一研究结果与现有社区的认识形成了鲜明的对比，目前已有的众多社区发现算法均没有考虑此种结构的社区，它们在算法中要么将社区重叠部分分成单独的一个聚类，要么将重叠部分的所有社区合并成一个更大的社区，因而无法正确识别出这类重叠社区。Yang 等人[7]在从属网络模型[8]的基础上提出了一个新的重叠社区发现方法来发现此类社区结构。但是该算法中忽略了社区大小对于网络结构的影响，而社区的规模又是从属网络的一个重要属性；此外在其方法中需要列举不同的社区结构来找到最合适的社区划分。为此，本文引入多维遗传算法来研究社区结构，通过遗传进化过程对网络的最佳社区结构进行搜索，搜索中综合考虑了个体吸引和社区规模的影响。

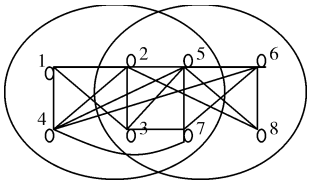


图 1 重叠区域中节点间连接密集的网络

1 相关工作

社区反映了网络中个体行为的局部聚集特征，在社交网络中存在大量的隐性或显性社区，社区结构的研究成为近年来社交网络研究的一个趋势[1~4]。

收稿日期：2015-08-25; 修回日期：2015-11-09 基金项目：国家自然科学基金青年基金资助项目(61401195); 南京工程学院校级青年基金资助项目(QKJB201408)

作者简介：王琦(1980-)，女，讲师，博士研究生，主要研究方向为数据挖掘、社交网络分析研究等(wangq@njit.edu.cn); 温志平(1975-)，女，副教授，硕士，主要研究方向为数据挖掘、社交网络分析研究等.

社区结构发现的研究最早是从图分割问题开始的。文献[9]定义了一个目标收益函数，通过贪婪搜索寻找使目标收益函数取值最大的网络分割，并将社区结构用树状图展现出来，揭示了社区的层次化结构。Newman 等人[10]提出了一个新的分裂算法即 GN 算法，通过迭代去除边介数最大的边将网络分裂为若干社区；同时 Newman 等人还在算法中定义了模块度的标准去评价社区结构划分的好坏，这个衡量指标后来被广泛应用，并且围绕着模块度的优化提出了许多基于优化理论的社区发现算法[11~13]。社区结构的发现与图聚类相似，其中比较有效的一类方法是使用标准相似矩阵特征值和特征向量的谱聚类[14~16]，从网络的拉普拉斯矩阵出发研究网络结构的划分。以 Infomap 为代表的算法[17]将网络拓扑结构映射为数据编码问题，通过构造最短编码来划分社区。

针对社交网络中节点同时隶属于多个社区的现象，Baumes 等人[18,19]提出了将社区定义为子图，通过局部优化一个边密度聚类的给定函数来找到重叠社区。Palla 等人[20]提出了利用社区内部边形成的派系去发现社区的 clique percolation method(CPM)算法，接着 Kumpula 等人[21]改进了 CPM 算法；Zhang 等人[22]提出了一个综合谱映射、模糊聚类和质量函数的优化方法；为了提高效率，Raghavan 等人[23]提出基于标签传播的社区发现算法。上述方法都是基于节点的社区发现算法，也有一些基于边的社区发现算法[24,25]。后来，Yu 等人[26]提出了一种基于图聚类的框架，成员之间的关系通过概率计算的方式获得；但是该方法在近似计算社区节点之间相互连接影响时，没有考虑到一对节点处于不同社区的情况，因此模拟出的网络拓扑结构不够精确，同时方法中也需要通过格外的限制信息如模块度来获得社区的精确个数。遗传算法作为解决问题最佳化的搜索算法被引入到社区发现研究中，不同于传统的社区发现算法需要指定社区个数的限制，基于遗传算法的发现方法不需要先验知识就可以自动计算社区的个数，很好地发现非重叠社区。Pizzuti[27]提出了一个基于多目标函数优化的社区发现算法(MOCK)，利用遗传算法优化多个社区评价指标函数。但是 Pizzuti 使用的一维染色体只能描述节点属于单个社区的情况，不能发现重叠社区。Dickinson 等人[28]提出了一个基于边聚类的遗传算法来发现重叠社团，但是仍然是使用一维染色体来描述边的邻接情况，增加了方法的复杂性和可理解性。

上述的社区发现方法都有一个共同的假设前提，即节点间在社区内部的连接要比社区之间重叠部分的连接紧密，现有的方法不能很好地处理前述重叠部分连接密集的社区结构。但是有效发现此类社区结构对于理解真实社交网络结构和功能具有一定意义。本文提出了一种新颖的基于遗传算法的社区发现算法，首次将多维染色体用于描述社区的结构，多维染色体的使用可以简单、准确地刻画真实网络中节点属于多个社区的情况。为了改进已有基于遗传算法方法的不足，本文借助于概率模型的基础上，不仅可以发现非重叠社区，也可以便捷地找出连接密集的重叠社区，且不需要指定任何参数。

2 本文方法

2.1 形式化描述

首先给出全文中需要使用的基本定义和符号。

定义 1 对于一个给定的网络 N，可以被描述成一个无权无向的图 $G(V,E)$，其中，每条节点 $v_i \in V$ 表示一个个体或对象，每条边 $e_{ij} \in E$ 表示一对节点 v_i 和 v_j 之间的连接边。$G(V,E)$ 的结构由一个 $n \times n$ 的邻接矩阵 A 决定(n 是 G 中的节点数)，其中矩阵的每个元素 $a_{ij}=1$ 表示一对节点 v_i 和 v_j 之间存在连接边 $e_{ij} \in E$，否则 $a_{ij}=0$。

定义 2 对于一个给定的图 $G(V,E)$，假设存在一个拟合 $G(V,E)$ 的二分从属网络 $B(V,C,M)$，其中，V 是节点的集合，C 是社区的集合，M 是从节点 $v(v \in V)$ 到社区 $c(c \in C)$ 的边的集合，$(v,c) \in M$ 意味着节点 v 属于社区 c。对于每个社区 c，存在一个独立的概率 $p_c(0 \leqslant p_c \leqslant 1)$，表示社区对于节点间相连概率的影响。

根据上文描述，对于一个给定的图 $G(N,E)$，社区结构 C 必须很好地满足网络的邻接矩阵 A。对于一对节点 $v_i, v_j \in V$，它们之间产生连接边 $e_{ij} \in E$ 的概率 $p(v_i,v_j)$ 受到它们共属的社区的影响。假设 v_i, v_j 共属的社区的集合是 $C_{v_i,v_j} \subseteq C$，其中 $C_{v_i,v_j} = \{c \mid (v_i,c), (v_j,c) \in M\}$，则 v_i, v_j 相连的概率用式(1)表示。

$$p(v_i,v_j) = p(v_j)\,p(v_i \mid v_j) = 1 - \prod_{c \in C_{v_i,v_j}} (1-p_c) \tag{1}$$

本文的核心思想是利用二分从属网络 $B(V,C,M)$ 来拟合图 $G(V,E)$。社交网络中的社区结构具有传统多模社会网络的结构特征，可以使用多模从属网络的理论来研究网络的社区结构。在从属网络中，成员通常属于不同的子群[7]；而在社交网络中，个体通常属于不同的社区。不同于本文的方法，Yu 等人[26]提出的基于图聚类的社区发现算法(GFC)，借助于随机块模型和概率模型来描述和发现社区结构，但是他们在考虑一对节点间相互连接的概率时，没有考虑到当两个节点分属于不同社区的情况，所以混合模型生成的网络不能准确地刻画网络的社区结构。为了改进这个问题，本文研究一对节点属于各种不同情况下社区的联合概率[7]，这就意味着当节点属于不同社区时，如果两个节点共享的社区数越多，它们之间产生边的概率就越大。基于文献[7]，本文根据节点共享社团的情况定义了一个似然概率函数来模拟网络中节点之间的连接概率，进而拟合整个网络真实的拓扑结构，每个社区独立产生一个影响概率因子。借助于多维遗传算法和凸优化方法，文中提出了一个识别社区结构的拟合算法。给定一个图 $G(V,E)$，利用二分从属网络 $B(V,C,M)$ 和影响概率因子集合 $\{p_c\}$ 来拟合网络 $G(V,E)$ 中节点之间的连接概率，即似然函数 $L(B,\{p_c\}) = P(G \mid B,\{p_c\})$。

$$\arg\max L(B,\{p_c\}) = \prod_{e_{ij} \in E} p(v_i,v_j) \prod_{e_{ij} \notin E} (1-p(v_i,v_j)) \tag{2}$$

2.2 算法描述

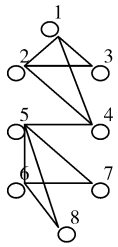
为了计算式(2)，本文使用了多维遗传方法和最优化的方法，具体求解过程如下。

遗传算法(GA)是一种随机自适应的全局搜索算法[29]。当问题的解空间非常大，且不可能通过穷举的方法求解精确解时，遗传算法通过模拟自然界中生物的遗传进化过程，对优化问题的最优解进行搜索，能够以较大的概率得到组合问题的最优解。一个标准的遗传算法通过对种群(即染色体集合)使用选择、交叉和变异的遗传算子来繁衍固定的代数。每一个染色体由一组基因组成，代表了给定问题的一个有效解。典型的遗传算法使用一维染色体，只能描述问题一方面的信息。基于一维遗传算法的社区发现算法只能描述节点属于单个社区的情

况,仅能发现非重叠社区结构。已有一些基于一维遗传算法的方法也试图改进这个缺点,但是都增加了算法的复杂性。本文通过对比发现这个潜在的优化问题(节点和社区)具有典型的二维特征,不易使用一维染色体进行描述,因此,本文首次将多维染色体引入到社区发现中来处理二维问题空间。本文方法基于标准的 Holland 遗传算法,使用了改进的二维遗传算子[30]。

定义 3　给定一个图 $G(V,E)$,基于 $n\times n$ 的邻接矩阵 A,本文假设函数 adiacent(e_{ij}),对于每个边 $e_{ij}\in E$(e_{ij}是节点 v_i,$v_j\in V$的连接边),返回其任一邻边 $e\in E$,即边 e_{ij}和 e 具有共同的节点 v_i 或者 v_j。若边 e_{ij}没有任何邻边,则返回 0。

1) 初始化　在图论描述的基础上,种群的二维染色体可以描述为一个由 $n\times n$ 个基因所组成的矩阵 $R_{n\times n}$,其中 n 是节点的个数($|V|$)。每一个基因 r_{ij}是函数 adiacent(e_{ij})的函数值,其中行 i 表示节点集中第 i 个节点($v_i\in V$),列 j 表示节点集中第 j 个节点($v_j\in V$)。这就意味着如果节点 v_i 和 v_j 存在连接边 e_{ij},则基因 r_{ij}的值就是边 e_{ij}的任一邻接边(即函数 adiacent(e_{ij})的返回值);否则,如果节点 v_i 和 v_j 不存在连接边,则基因 r_{ij}的值为 0。图 2 所示为一个网络及其对应的二维染色体编码。



(a)样图

$$\begin{bmatrix} 0 & e_{13} & e_{23} & e_{45} & 0 & 0 & 0 & 0\\ e_{13} & 0 & e_{24} & e_{45} & 0 & 0 & 0 & 0\\ e_{23} & e_{24} & 0 & 0 & 0 & 0 & 0 & 0\\ e_{45} & e_{45} & 0 & 0 & e_{42} & 0 & 0 & 0\\ 0 & 0 & 0 & e_{57} & 0 & e_{57} & e_{67} & e_{56}\\ 0 & 0 & 0 & 0 & e_{57} & 0 & e_{68} & e_{58}\\ 0 & 0 & 0 & 0 & e_{67} & e_{68} & 0 & 0\\ 0 & 0 & 0 & 0 & e_{56} & e_{58} & 0 & 0 \end{bmatrix}$$

(b)图(a)对应的一种可选染色体编码

图 2　一个网络及其对应的二维染色体编码

算法中的种群通过随机或者贪婪算法产生,每个染色体矩阵 $R_{n\times n}$的每个基因值 r_{ij}是函数 adiacent(e_{ij})的返回值。例如在图 2 中,第一行第二列的对应的值是 e_{13},代表的是边 e_{12}的任一邻边(即 e_{13})。这种初始化的方法是为了使得算法将网络结构朝着节点互连的集合进行划分,通过限制解空间来提高算法的收敛性。

2) 均匀块交叉　交叉是通过交换父染色体中的部分基因产生新的后代。本文使用一个通用的被称为均匀块交叉的二维交叉算子,如图 3 所示。给定两个父染色体,随机产生一个均匀块交叉算子,交叉算子将根据预设的交叉概率交换两个父染色体在矩形块中的基因部分,矩形块的大小和位置是随机产生的。使用均匀块交叉算子的好处是保证交叉产生的后代染色体的每个基因都是已有边的邻接边,使得算法仍然是朝着节点互连的集合进行网络结构划分。

3) 变异　变异算子是以预设的概率改变种群中染色体的某些基因值。随意地改变基因值会使得解空间朝着无效的方向搜索,因此,本文对染色体中每个待变异的基因 $r_{ij}\in R_{n\times n}$,规定基因变异后只能取函数 adiacent(e_{ij})的另一个值,即另一条邻接边。如果函数 adiacent(e_{ij})有且仅有一个值,则变异算子不改变基因 r_{ij}的值。

4) 适应度函数　选择是从种群中根据适应度函数选择优胜的个体,淘汰劣质的个体,目的是把优化的个体(或解)直接遗传到下一代或通过配对交叉产生新的个体再遗传到下一代。本文在搜索进化过程中不需要其他外部信息,仅使用适应度函数来评估个体的优劣,并作为后续遗传操作的依据。适应度函数是判断种群中个体优劣程度的指标,根据所求问题的目标函数来进行评估。根据上文描述,对于一个给定的图 $G(V,E)$,本文的目标是通过不断搜索进化寻找一个二分从属网络 $B(V,C,M)$和概率集合$\{p_c\}$来很好地描述网络 $G(V,E)$的拓扑结构,因此,算法的适应度函数即为似然函数 $L(B,\{p_c\})=P(G|B,\{p_c\})$,通过最大化此似然函数来找到最符合真实的网络拓扑结构,发现最佳的社区结构。

$$\begin{bmatrix} 0 & e_{13} & e_{23} & e_{45} & 0 & 0 & 0 & 0\\ e_{13} & 0 & e_{24} & e_{45} & 0 & 0 & 0 & 0\\ e_{23} & e_{24} & 0 & 0 & 0 & 0 & 0 & 0\\ e_{45} & e_{45} & 0 & 0 & e_{42} & 0 & 0 & 0\\ 0 & 0 & 0 & e_{57} & 0 & e_{57} & e_{67} & e_{56}\\ 0 & 0 & 0 & 0 & e_{57} & 0 & e_{68} & e_{58}\\ 0 & 0 & 0 & 0 & e_{67} & e_{68} & 0 & 0\\ 0 & 0 & 0 & 0 & e_{56} & e_{58} & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & e_{13} & e_{12} & e_{12} & 0 & 0 & 0 & 0\\ e_{13} & 0 & e_{13} & e_{45} & 0 & 0 & 0 & 0\\ e_{12} & e_{13} & 0 & 0 & 0 & 0 & 0 & 0\\ e_{12} & e_{45} & 0 & 0 & e_{41} & 0 & 0 & 0\\ 0 & 0 & 0 & e_{58} & 0 & e_{58} & e_{58} & e_{67}\\ 0 & 0 & 0 & 0 & e_{58} & 0 & e_{68} & e_{58}\\ 0 & 0 & 0 & 0 & e_{58} & e_{68} & 0 & 0\\ 0 & 0 & 0 & 0 & e_{67} & e_{58} & 0 & 0 \end{bmatrix}$$

(a)父染色体的均匀块交叉

$$\begin{bmatrix} 0 & e_{13} & e_{23} & e_{45} & 0 & 0 & 0 & 0\\ e_{13} & 0 & e_{13} & e_{45} & 0 & 0 & 0 & 0\\ e_{23} & e_{24} & 0 & 0 & 0 & 0 & 0 & 0\\ e_{45} & e_{45} & 0 & 0 & e_{41} & 0 & 0 & 0\\ 0 & 0 & 0 & e_{58} & 0 & e_{58} & e_{58} & e_{56}\\ 0 & 0 & 0 & 0 & e_{58} & 0 & e_{68} & e_{58}\\ 0 & 0 & 0 & 0 & e_{67} & e_{68} & 0 & 0\\ 0 & 0 & 0 & 0 & e_{56} & e_{58} & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & e_{13} & e_{12} & e_{12} & 0 & 0 & 0 & 0\\ e_{13} & 0 & e_{24} & e_{45} & 0 & 0 & 0 & 0\\ e_{12} & e_{13} & 0 & 0 & 0 & 0 & 0 & 0\\ e_{12} & e_{45} & 0 & 0 & e_{42} & 0 & 0 & 0\\ 0 & 0 & 0 & e_{57} & 0 & e_{57} & e_{67} & e_{67}\\ 0 & 0 & 0 & 0 & e_{57} & 0 & e_{68} & e_{58}\\ 0 & 0 & 0 & 0 & e_{58} & e_{68} & 0 & 0\\ 0 & 0 & 0 & 0 & e_{67} & e_{58} & 0 & 0 \end{bmatrix}$$

(b)产生的后代染色体

图 3　父染色体的均匀块交叉及产生的后代染色体

下面基于多维遗传算法通过不断迭代来求解上文的优化问题。

对于一定规模的种群,遗传算法使用选择、交叉和变异等遗传算子对个体进化固定的代数。本文引入二维染色体作为种群的个体,它是一个由 $n\times n$ 个基因组成的矩阵。在每一次进化迭代中,遗传算法通过上文描述的遗传算子随机产生一个新的种群。每一个染色体表示一个待选的从属网络 B_i。对于染色体 $R_{n\times n}$,它的每个非零基因 r_{ij}表示边 e_{ij}的邻边 $e(e,e_{ij}\in E)$。对于一个给定的图 $G(V,E)$,如果边 e 和 e_{ij}相邻($e,e_{ij}\in E$),则边 e 和 e_{ij}应该处于同一社区中。在本文中,对于每个二维染色体矩阵 $R_{n\times n}$,从它的第一个元素(表示边 e_{ij})及其值(表示邻边 e)开始,顺着边的连接关系找到属于同一社区的所有其他边。社区的划分使用递归跟踪函数完成,直到找到一个包含所有边的最小的社区结构,每个社区中的边都是相邻边。这个最小的社区结构就是文中需要寻找的从属网络 B_i。

使用上述方法,对于每个染色体 $R_{n\times n}$可以求出对应的从属网络 B_i。对于每个得到的从属网络 B_i,算法的目标是找到一组合适的概率 p_c,使得式(2)的似然函数取得最大值。这个优化问题不属于凸优化范畴,将式(2)转换为式(3)[7]。此时,似然函数的极值问题转换为一个凸优化问题,可以使用梯度下降的方法求解出全局的最优解,即$\{\arg\max L_i\}$的值以及对应的$\{p_c\}$。算法使用每个染色体的适应度函数值选择个体,并借助于遗传算子进行交叉和变异,产生出代表新的解集的种群。末代种群中的最优个体可以作为近似的最优解,即网络的社区结构。为了找到最符合真实拓扑结构的最小社区划分,算法中对概率$\{p_c\}$的求解中引入了惩罚机制。

$$\arg\max_{\{\theta_k\}}\sum_{e_{ij}\in E}\log\left(1-e^{-\sum_{c_k\in C_{v_i,v_j}}\theta_k}\right)-\sum_{e_{ij}\notin E}\sum_{c_k\in C_{v_i,v_j}}\theta_k \qquad (3)$$

3 实验和分析

通过一系列的试算过程，文中遗传算法中的交叉概率设为 0.8，变异概率设为 0.2，固定种群的个数设为 20 000，进化迭代的次数设为 200。为了验证文中方法的有效性，将本文算法(TGA)与最相近算法 MOCK 进行比较[27]，评价指标使用归一化互信息 NMI(式(4))[31]，在实验的过程中不需要指定任何参数(如社区个数)等。文中使用已知社区结构的人工数据集进行各种实验来验证方法的准确性，所使用的数据集是 Girvan 等人[32]采用的评价基准。这个数据集包含 128 个节点，均分为四个社区，节点的平均节点度 avgDegree 为 16，每个节点拥有一些不属于各自社区的连接。本文设计了 10 种不同连接情况的上述数据集。为了验证本文方法对于上述社区结构划分的有效性，本文从下面三个角度进行实验。为了产生不同的网络结构，本文引入了参数 z，用来表示节点拥有的社区外的连接数。第一类实验本文设定参数 z 的值为 0，第二类实验设定参数 z 的均值为 5，第三类实验设定参数 z 的均值为 12。

$$\mathrm{NMI}(C_1,C_2)=\frac{-2\sum_{i=1}^{C_{C_1}}\sum_{j=1}^{C_{C_2}}\log(C_{ij}N/C_{i.}C_{.j})}{\sum_{i=1}^{C_{C_1}}C_{i.}\log(C_{i.}/N)+\sum_{j=1}^{C_{C_2}}C_{.j}\log(C_{.j}/N)} \quad (4)$$

由于文中网络的社区结构都是已知的，可以直接通过式(4)计算得出算法推出的社区结构与事实的社区结构之间的差异，即 NMI 值。NMI 值越高，说明算法划分的社区结构与真实的网络结构就越相近，当 NMI 值为 1 时，说明推出的社区结构与事实的社区结构完全相同。在进行多轮重复实验后，本文方法和 MOCK 算法的平均实验情况如图 4 ~6 所示。

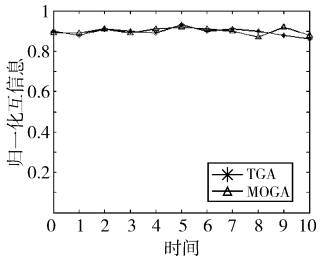


图 4 $z=0$ 时的 NMI 曲线

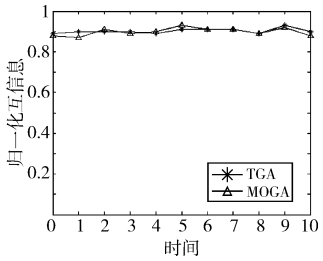


图 5 $z=5$ 时的 NMI 曲线

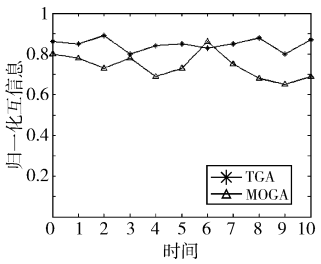


图 6 $z=12$ 时的 NMI 曲线

对于每一类实验，参数 z 的均值保持不变，均设计了 10 种不同节点连接情况的数据集。从实验结果可以发现，当参数 z 为 0 时(图 4)，说明不同社区间的节点之间没有连接，社区之间是互相不重叠的，从图 4 的 NMI 曲线变化可以看出本文的方法和 MOCK 算法都可以较准确地获得当前网络的社区结构。当参数 z 为 5 时(图 5)，说明节点有属于社区外的连接，但数目较少，即社区节点在重叠部分的连接要比社区内部稀疏。此时，从图 5 的 NMI 曲线变化亦可以看出本文的方法和 MOCK 算法都可以较准确地获得当前网络的社区结构。前述两类实验说明，本文方法和 MOCK 算法可以很好地发现非重叠社区和稀疏连接社区的结构。当参数 z 为 12 时(图 6)，这时节点在社区之间的连接要比社区内部多，即节点在社区重叠部分的连接要比社区内部连接紧密，此时社区的划分难度增大。从图 6 的 NMI 曲线变化可以看出，本文方法对此类社区结构的划分要比 MOCK 算法相对精确，这就验证了本文方法对于上述具有连接密集型重叠部分的社区结构划分更加准确。

4 结束语

社交网络中社区结构的发现越来越成为各领域专家研究的重点，分析社区的结构和构成有利于研究网络拓扑结构的特点、找出个体聚集的模式和影响因素，更好地进行网络信息检索和推荐。本文基于遗传算法提出了一个发现连接密集型重叠区域的社区发现算法，首次将多维染色体引入到社区发现算法中。多维遗传算子尤其是均匀块交叉算子的使用简化了研究的问题，减小了算法的复杂性。实验的结果也验证了本文方法的有效性和有用性。社交网站中的节点和边可能是多种类型的，后续工作将在本文的基础上进一步进行异构网络的研究。

参考文献:

[1] Fortunato S. Community detection in graphs [J]. Physics Reports, 2010, 486(3-5): 75-174.

[2] Aggarwal C C. Social network data analytics [M]. [S.l.]: Springer, 2011.

[3] Kurka D B, Godoy A, Von Zuben F J. Online social network analysis a survey of research applications in computer science [J]. Computer Science, 2015(4).

[4] 方滨兴，等. 在线社交网络分析 [M]. 北京: 电子工业出版社, 2014.

[5] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth [J]. Knowledge and Information Systems, 2012, 42(1): 746-754.

[6] Yang J, Leskovec J. Structure and overlaps of communities in networks [J]. Computer Science, 2012, 356(17): 3530-3538.

[7] Yang J, Leskovec J. Community-affiliation graph network model for overlapping community detection [C]//Proc of the 12th IEEE International Conference on Data Mining. 2012: 1170-1175.

[8] Lattanzi S, Sivakumar D. Affiliation networks [C]//Proc of ACM Symposium on Theory of Computing. 2009: 427-434.

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

[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2): 026113.

[11] Clauset A. Finding local community structure in networks [J]. Physical Review E, 2005, 72(2): 026132.

[12] Schuetz P, Caflisch A. Multistep greedy algorithm identifies community structure in real world and computer generated networks [J]. Physical Review E, 2008, 78(2): 026112.

[13] Newman M E J. Modularity and community structure in networks [J]. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577-8582.

[14] Zha Hongyuan, He Xiaofeng, Ding C, *et al*. Spectral relaxation for K-means clustering [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 1057-1064.

(下转第 3553 页)

本文利用 Markov 逻辑网强大的逻辑表达能力和推理能力，从维度内容、事件表象和数据源等方面制定规则，分为两步骤推理优选维度内容，充分利用集成数据解决事件表象不完整、不一致、不精确问题。实验结果表明，该方法有效地提高了事件表象统一的准确率和召回率。

参考文献：

[1] 张传言，洪晓光，彭朝晖，等. 基于 SVM 和扩展条件随机场的 Web 实体活动抽取 [J]. 软件学报，2012，23(10)：2612-2627.

[2] Wu Juebo，Wu Zongling. Comprehensive approach to semantic similarity for rapid data integration [J]. International Journal of Control，Automation and Systems，2014，12(3)：680-687.

[3] Xia Y，Leung H. A Fast learning algorithm for blind data fusion using a novel L2-norm estimation [J]. IEEE Sensors Journal，2014，14(3)：666-672.

[4] 潘泉，于昕，程咏梅. 信息合并理论的基本方法与进展 [J]. 自动化学报，2003，29(4)：599-609.

[5] Ayat N，Akbarinia R，Afsarmanesh H，*et al*. Entity resolution for probabilistic data [J]. Information Sciences，2014，277(2)：492-511.

[6] Park C Y，Laskey K B，Costa P C G，*et al*. Multi-entity Bayesian networks learning for hybrid variables in situation awareness [C] // Proc of the 16th International Conference on Information Fusion. 2013：1894-1901.

[7] Fillmore C J. Frames and the semantics of understanding [M]. [S. l.]：Quademi di Semantica，1985：222-253.

[8] 徐永东，徐志明，王晓龙. 基于信息融合的多文档自动文摘技术 [J]. 计算机学报，2007，30(11)：2048-2054.

[9] 毕凯，王晓丹，邢雅琼. 基于 D-S 证据理论的模糊聚类集成 [J]. 系统工程与电子技术，2014，36(7)：1446-1452.

[10] 徐从富，郝春亮，苏保君，等. 马尔可夫逻辑网络研究 [J]. 软件学报，2011，22(8)：1699-1713.

[11] 杨国宁，冯秀芳. D-S 证据理论中冲突证据融合新方法 [J]. 计算机应用与软件，2014，31(2)：82-85.

[12] 董永权，李庆忠，丁艳辉. 一种基于证据理论和任务分配的 Deep Web 查询接口匹配方法 [J]. 模式识别与人工智能，2011，24(2)：262-271.

[13] 刘大有，于鹏，高滢，等. 统计关系学习研究进展 [J]. 计算机研究与发展，2008，45(12)：2110-2119.

[14] Singla P，Domingos P. Entity resolution with Markov logic [C] //Proc of the 6th Industrial Conference on Data Mining. 2006：572-582.

[15] 张永新，李庆忠，彭朝晖. 基于 Markov 逻辑网的两阶段数据冲突解决方法 [J]. 计算机学报，2012，35(1)：101-111.

[16] Yang Jiangming，Cai Rui，Wang Yida，*et al*. Incorporating site-level knowledge to extract structured data from Web forums [C] //Proc of the 18th International Conference on World Wide Web. 2009：181-190.

[17] 张玉芳，黄涛. Markov 逻辑网在重复数据删除中的应用 [J]. 重庆大学学报，2010，33(8)：36-41.

[18] Yin Xiaoxin，Han Jiawei，Yu P S. Truth discovery with multiple conflicting information providers on the Web [C] //Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2007：1048-1052.

[19] 朱毅华，侯汉清，沙印亭. 计算机识别汉语同义词的两种算法比较和测评 [J]. 中国图书馆学报，2002，28(4)：82-85.

[20] 夏天. 汉语词语语义相似度计算研究 [J]. 计算机工程，2007，33(6)：191-194.

[21] 付剑锋. 面向事件的知识处理研究 [D]. 上海：上海大学，2010.

[22] 程传鹏，吴志刚. 一种基于知网的句子相似度计算方法 [J]. 计算机工程与科学，2012，34(2)：172-175.

[23] 江敏，肖诗斌，王弘蔚，等. 一种改进的基于《知网》的词语语义相似度计算 [J]. 中文信息学报，2008，22(5)：84-89.

[24] Poon H，Domingos P. Sound and efficient inference with probabilistic and deterministic dependencies [C] //Proc of the 21st National Conference on Artificial Intelligence. 2006：458-463.

[25] Singla P，Domingos P. Discriminative training of Markov logic networks [C] // Proc of the 20th National Conference on Artificial Intelligence. 2005：868-873.

[26] Lowd D，Domingos P. Efficient weight learning for Markov logic networks [C] //Proc of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. 2007：200-211.

[27] Kok S，Singla P，Richardson M，*et al*. The Alchemy system for statistical relational AI [R]. Seattle：University of Washington，2010.

(上接第 3546 页)

[15] Donetti L，Munoz M. Detecting network communities：a new systematic and efficient algorithm [J]. Journal of Statistical Mechanics，2004(10)：10012.

[16] Capocci A，Servedio V D P，Caldarelli G，*et al*. Detecting communities in large networks [J]. Physica A：Statistical Mechanics and Its Applications，2005，352(2-4)：669-676.

[17] Rosvall M，Bergstrom C. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences，2008，105(4)：1118-1123.

[18] Baumes J，Goldberg M K，Krishnamoorthy M S，*et al*. Finding communities by clustering a graph into overlapping subgraphs [C] // Proc of the IADIS International Conference on Applied Computing. 2005：97-104.

[19] Baumes J，Goldberg M，Magdon-Ismail M. Efficient identification of overlapping communities [C] //Proc of IEEE International Conference on Intelligence and Security Informatics. 2005：27-36.

[20] Palla G，Derényi I，Farkas I，*et al*. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature，2005，435(7043)：814-818.

[21] Kumpula J M，Kivel M，Kaski K，*et al*. Sequential algorithm for fast clique percolation [J]. Physical Review E，2008，78(2)：026109.

[22] Zhang Shihua，Wang Ruisheng，Zhang Xiangsun. Identification of overlapping community structure in complex networks using fuzzy C-means clustering [J]. Physical A：Statistical Mechanics and Its Applications，2007，374(1)：483-490.

[23] Raghavan U N，Albert R，Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E，2007，76(3)：036106.

[24] Evans T S，Lambiotte R. Line graphs，link partitions and overlapping communities [J]. Physical Review E，2009，80(1)：016105-016112.

[25] Ahn Y Y，Bagrow J P，Lehmann S. Communities and hierarchical organization of links in complex networks [J]. Nature，2009，466(2)：761-764.

[26] Yu Shipeng，Yu Kai，Tresp V. Soft clustering on graphs [C] // Advances in Neural Information Processing Systems. 2005：1553-1560.

[27] Pizzuti C. A multiobjective genetic algorithm to find communities in complex networks [J]. IEEE Trans on Evolutionary Computation，2012，16(3)：418-430.

[28] Dickinson B，Valyou B，Hu Wei. A genetic algorithm for identifying overlapping communities in social networks using an optimized search space [J]. Social Networking，2013，2(4)：193-201.

[29] Holland J H. Adaptation in natural and artificial systems：an introductory analysis with applications to biology，control，and artificial intelligence [M]. [S. l.]：Michigan Press，1975.

[30] Anderson C A，Jones K F，Ryan J. A two-dimensional genetic algorithm for the Ising problem [J]. Complex Systems，1991，5(3)：327-333.

[31] Danon L，Díaz-Guilera A，Duch J，*et al*. Comparing community structure identification [J]. Journal of Statistical Mechanics：Theory and Experiment，2005(9)：P09008.

[32] Girvan M，Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences，2002，99(12)：7821-7826.