基于局部思想的在线社区划分算法

张 瑜,蔡国永
(桂林电子科技大学 广西可信软件重点实验室,广西 桂林 541004)

摘 要:为了快速准确地对在线社会网络进行社区划分,提出了一种基于局部思想的社区划分算法。该算法利用节点和社区聚集系数的性质,结合局部模块度将节点划分成相对独立的社区。算法运行时,只需要了解与目标节点相关的局部网络信息,时间复杂度相对较低,并且也可以用来对整个在线社会网络进行社区划分。利用该算法分别对 Zachary 空手道俱乐部网络和在线社会网络进行划分实验,得到满意的结果。

关键词:复杂网络;在线社会网络;社区划分;聚集系数;局部社区
中图分类号:TP301 **文献标识码**:A **文章编号**:1674-7720(2013)13-0076-04

Online community partition based on localized network

Zhang Yu, Cai Guoyong
(Guangxi Key Lab.of Trusted Software, Guilin University of Electronic Science and Technology, Guilin 541004, China)

Abstract: In order to quickly and accurately divide the online social networks, this paper proposes a partitioning algorithm based on local community theory. This algorithm uses the property of node and community clustering coefficient, and combines with local modularity to divide the nodes into relatively independent community. When it runs, we only need to know the local network information of the target nodes. The time complexity is relatively low, and this method can also be used to divide the whole online social network community. We make experiments on Zachary Karate Club network and online social networks respectively, and satisfied the results are reached.

Key words: complex network; online social network; community detecting; clustering coefficient; local community

在线社会网络是以计算机和网络为中介进行社交、联系和协作形成人与人之间的社会网络,如 Facebook、人人网等。国内外学者研究发现在线社会网络具有明显的社区结构和小世界特性[1],具有节点度的幂律分布特性和高聚类特性[2]。在这样一个异常庞大复杂的系统中,如何按照 Newman 等人给出的经典定义[3],即社区内部连接的紧密程度大于社区间连接的紧密程度,将在线社会网络中联系比较紧密的节点划分成一个社区,使之成为若干个稀疏的子系统,具有重要的意义。对在线社会网络进行社区划分,可以帮助更好地了解网络结构,协调各个社区之间的关系,为信息的查询、搜索提供更为方便快捷的途径。

关于社区划分的算法已经提出很多,如谱平分法[4-5]、GN 算法[6]、Newman 快速算法[7]及利用堆结构的贪婪算法[8]等。这些算法都是从网络的全局信息出发,寻找整个网络的社区结构,而在线社会网络是一个异常庞大、动态变化的复杂系统,获得全局信息是非常困难的;同时很多时候,人们并不需要获得整个网络的社区划分,而只关心某一个节点所在的局部社区。例如,在人人网中,只关心某个人所在的社区,或者是某一个特定的社区。在这种情况下,就没有必要消耗过多的时间计算和寻找全网的社区结构。近年来复杂网络中运用局部的概念来划分社区的方法也有很多[9-11]。参考文献[9]提出“局部模块性”的概念,通过最大化局部模块度为目标来划分局部社区,但单纯的最大化局部模块度难以得到正确的社区划分,并且时间复杂度较高。参考文献[10]提出一种广度搜索方法来寻找某个节点所在的局部社区,称为 BB 算法,该算法的不足之处在于它把社区整个一层邻居节点全部加入或全部排除在社区之外。参考文献[11]提出以节点度为优先的快速局部社区划分算法,但

仅仅是通过度数较大的节点来吸引度数较小的节点并加入到它所在的社区，对于边缘节点和度数较大的节点来说，并不能得到正确的网络划分。

本文提出的在线社会网络局部社区划分算法，与之前提出的局部社区划分算法相比，本质上是一种聚类算法。由于寻找的是局部社区，所以不用计算确定网络的中心节点，而是从任意初始节点开始，通过搜索它的邻居节点，选择符合条件的节点加入社区。算法中采用计算社区聚集系数的方法，即假如邻居节点属于目标社区，比较各邻居节点加入后的社区聚集系数，选择使得聚集系数最大的邻居节点加入社区，并且采用 Clauset 引入的局部模块度 Q 对社区进行终止判断[9]。局部社区划分算法只需要节点的局部信息就可以得到社区划分，对于规模庞大的在线社会网络中寻找局部社区来讲，是十分快捷方便的。但是由于采用的是局部社区划分，所以得到的社区往往是局部最优而不是全局最优的。若要得到整个网络的社区划分，就需要不断迭代此算法。

1 相关概念定义

为了阐述本文算法的方便，下面首先给出算法中几个关键概念的定义和计算。

定义 1 **节点聚集系数** 在一个网络图中，假设节点 i 有 k_i 条边和其他节点相连，如果这 k_i 个节点之间也都相互连接，那么它们之间最多有 $k_i(k_i-1)/2$ 条边。将 k_i 个节点之间实际存在的边数 E_i 与其可能的总边数之比定义为节点 i 的聚集系数，用 C_i 表示，则 $C_i=2E_i/(k_i(k_i-1))$。

聚集系数用来描述网络中节点的聚集情况，是衡量网络集团化程度的重要参数，即网络紧密的程度。聚集特性表示节点 i 的邻居节点间彼此也可能是邻居的情况。

网络中所有节点聚集系数的平均值，即 $C=\frac{\sum_{i=1}^{n}C_i}{n}$，称为网络聚集系数，或叫做社区聚集系数。其中 n 表示网络中的节点数。显然，节点的聚集系数和网络的聚集系数的取值范围都在 0 和 1 之间。网络的聚集系数越大，该网络的耦合性越强。

本文采用 Clauset 引入的局部模块度 Q 的思想对初始社区进行判断[9]。局部模块度的定义如下[12]：

定义 2 **局部模块度** 在一个社区或网络中，其两个顶点都在社区内的边数与整个社区关联的边数之比，称局部模块度，用 Q 表示，则 $Q=\frac{L_{in}}{L_{in}+L_{out}}$。其中 L_{in} 表示两个顶点都在社区内的边的数目，L_{out} 表示只有一个顶点在社区内的边的数目。

一般说来，一个具有社区结构的网络，社区内部的连接密度要远远大于社区间的连接密度。因此，局部模块度 Q 越大，则社区结构越明显。局部模块度只需要已知节点的局部信息而不是网络的全局信息，所以适用于局部社区的划分，和 Newman 等人提出的全局模块度相比[13]，可以在一定程度上降低算法的复杂度。

2 算法设计

本文提出的局部社区划分算法不需要确定中心节点，而是将指定节点加入结果社区，得到一个局部社区，通过计算其邻居节点加入该社区后的社区聚集系数，选择聚集系数最大的节点加入此社区，形成新的局部社区，继续选择邻居节点，并重复计算。在将节点加入局部社区的过程中，算法重复计算社区的局部模块度 Q，直到 Q 值不再增大为止，此时该局部社区形成。为了降低算法的复杂度，做如下关于邻居节点的规定：如果一个节点一半以上的邻居节点都在结果社区中，那么不再计算社区聚集系数，将此节点直接加入社区。算法步骤如下：

（1）初始化。对于网络中每一个节点，将其保存为一个如下的线性动态链表，包括节点的度、节点聚集系数、社区聚集系数、邻居节点集、局部模块度及节点标号（初始值为 0）。如表 1 所示。

表 1 线性动态链表

节点	度	节点聚集系数	社区聚集系数	邻居节点集	局部模块度	节点标号
i	D	C_i	C_{ij}	N_i	Q	Num

（2）从指定节点开始，作为一个局部社区。搜索其邻居节点，依次计算各邻居节点加入此社区后社区的聚集系数 C_{ij}，选取使得 C_{ij} 最大的一个节点 j 加入（约定若最大的 C_{ij} 相等，则任选一个节点加入）；计算此时社区的局部模块度 Q，更新 j 的社区标号为 1；

（3）若有节点超过一半的邻居都在此社区中，不计算，直接加入，更新节点标号和局部模块度；

（4）若节点没有超过一半的邻居都在此社区中，那么计算社区所有的邻居节点的社区聚集系数，并将聚集系数取得最大的节点加入结果社区，并更新 Q 值和节点标号；

（5）当社区的邻居节点集使得 Q 值减小或者不再增加时，算法终止，局部社区形成；

（6）若需要划分整个网络，则从社区标号为 0 的节点中指定一个节点，重复步骤（1）~（5），就可以得到下一个局部社区，并重复此过程，社区划分完毕。

由以上算法得出，将社区全部划分完毕时，本文算法的时间复杂度为 $O(n^2)$，n 为社区的全部节点数。当要寻找的只是某个局部社区时，算法的时间复杂度要低于 $O(n^2)$，GN 分裂算法的时间复杂度为 $O(n^3)$。

3 算法示例

下面给出一个简单网络对该算法进行简要的分析说明，该社区包括 19 个节点，37 条边，可以划分成三个社区。如图 1 所示。

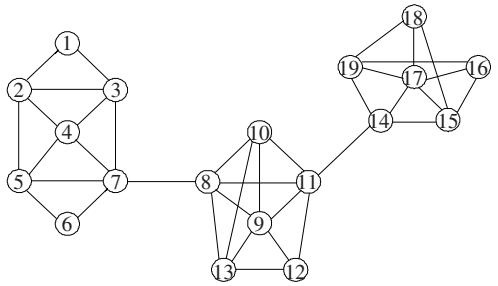

图 1 具有 19 个节点的简单网络图

假如现在想要寻找 19 节点所在的社区。首先，选择 19 节点作为初始节点，那么节点 19 形成一个初始社区 $C(S)$，S 表示社区节点集合。此时，S={19}。19 节点有 4 个邻居节点 N_{19}={14，17，16，18}，分别计算各个邻居节点加入社区 $C(S)$后的社区聚集系数 C_{ij}，见表 2，此时 18 节点和 16 节点的社区聚集系数相等。按照之前的约定，任选一个节点加入社区，如选择 18 节点加入，则 S={19，18}，局部模块度 Q=1/6。在社区 $C(S)$中，现在的邻居节点集为{17，16，14，15}，重新计算这 4 个节点加入社区 $C(S)$后的社区聚集系数 C_{ij}，见表 3，此时发现 16 节点的社区聚集系数最大，将 16 节点加入，则 S={19，18，16}，此时 Q=1/4。观察此时的社区情况，17 节点有一半的邻居节点 N_{17}={19，18，16}在目的社区中，此时不用计算直接将 17 节点加入，Q=1/2。同理，加入 15 节点和 14 节点，局部模块度 Q 分别为 8/11 和 11/12，此时 S={19，18，16，17，15，14}，见表 4 所示，表中黑色粗体部分为节点在局部社区中的邻居节点集。现在的社区还有一个邻居节点{11}，假如现在{11}节点属于这个局部社区，计算社区的局部模块度 Q，得出 Q=9/12，和之前的局部模块度相比，处于降低的趋势（之前为 11/12），此时算法终止，局部社区形成。同样的道理，可以划分出另外 2 个社区。

表 2 S={19}时各邻居节点的计算

i	D	Ci	C_{ij}	N_i	Q	Num
18	3	2/3	7/12	{15，17，19}	1/6	1
17	5	3/5	11/20	{14，15，16，18，19}	---	0
16	3	2/3	7/12	{15，17，19}	---	0
14	4	1/3	5/12	{11，15，17，19}	---	0

表 3 S={19，18}时各邻居节点的计算

i	D	Ci	C_{ij}	N_i	Q	Num
17	5	3/5	53/90	{14，15，16，18，19}	---	0
16	3	2/3	11/18	{15，17，19}	1/4	1
14	4	1/3	1/2	{11，15，17，19}	---	0
15	4	1/3	5/9	{14，17，18，16}	---	0

表 4 节点{17，15，14，11}加入社区 $C(S)$后的局部模块度

i	D	Ci	C_{ij}	N_i	Q	Num
17	5	3/5	---	{14，15，16，18，19}	1/2	1
15	4	1/3	---	{14，17，18，16}	8/11	1
14	4	1/3	---	{11，15，17，19}	11/12	1
11	5	2/5	---	{8，9，10，12，14}	9/12	0（不加）

4 算法应用及分析

4.1 Zachary 空手道俱乐部网络

在 20 世纪 70 年代初，ZACHARY W 通过观察美国大学空手道俱乐部成员间的人际关系，并根据俱乐部成员间平时的交往状况建立了一个网络[14]。网络包含 34 个节点、78 条边，如图 2 所示。节点代表俱乐部的成员，边代表成员间的关系。由于突发原因，俱乐部主管与校长之间因是否提高收费问题产生争执，并最终导致俱乐部分裂成两部分。其中节点 34 和节点 1 分别代表了俱乐部校长和主管，白色和灰色的节点分别代表分裂后的两个社区节点。Zachary 空手道俱乐部网络是用来判断社区划分效果的常用实验网络，用来检验测试算法能否准确划分网络结构。

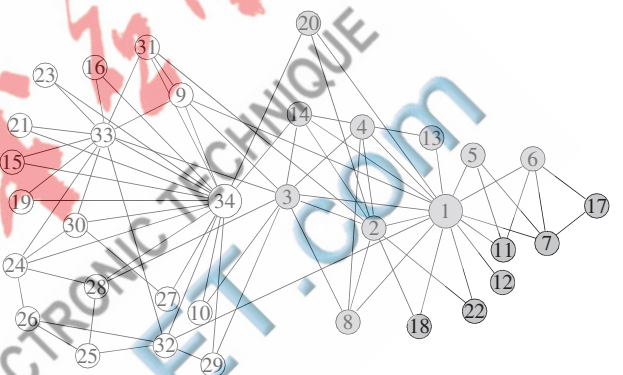

图 2 Zachary 空手道俱乐部网络

根据本文算法，任意选择一个节点，如选择 13 节点作为初始社区 $C(S)$。此时 13 节点有两个邻居 N_{13}={1，4}，4 节点具有更大的社区聚集系数 $C_{13,4}$ 并且加入社区后使得社区局部模块度 Q 增大，所以选择 4 节点加入社区。此时，S={13，4}。同样的道理，依次加入节点{8，14，18，2，22，3，1}。此时观察 20 节点共有两个邻居 N_{20}={1，2}在社区中，按照之前的约定，若有节点超过一半的邻居都在此社区中，不计算，直接加入。然后进一步加入{5，11，6}。与加入 20 节点同样的道理，将节点{7，17，12}直接加入。最后搜索此时的社区邻居节点集{34，31，9，33，28，10，32，29}，发现邻居节点中没有符合条件的节点加入社区。于是，局部社区形成。如图 2 所示灰色部分节点。循环利用此算法，将网络划分成两个社区，和初始结果一致。

图 3 表示在划分第一个社区（灰色部分节点）的过程中，试探加入社区邻居节点后局部模块度 Q 的变化情况。其中，曲线表示的是经过计算社区聚集系数和局部模块度最终加入目的社区的节点{13，4，8，14，2，18，22，3，1，12，20，5，11，7，6，17}，其余部分表示经过计算试探未能加入社区的节点。从图中可以看出，最终加入社区的节点都使得社区的局部模块度 Q 增加，直到局部模块度不增加为止，最终形成局部社区。

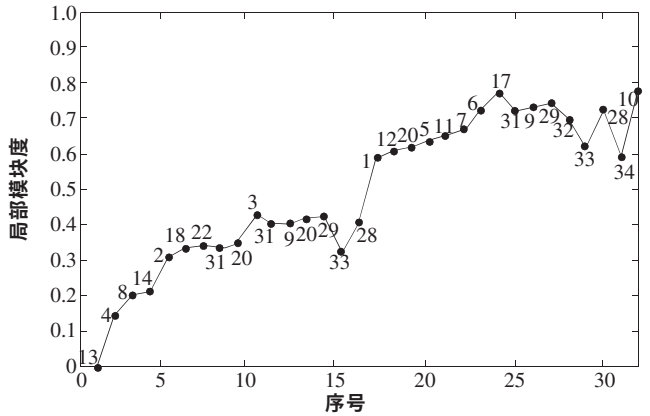

图 3 加入各个节点后社区局部模块度的变化

4.2 在线社会网络示例分析

为验证算法对真实在线社会网络分析时的有效性和可靠性，设计了一组试验，试验数据集取自人人网，该数据集有 166 个节点，468 条边，网络直径为 9，平均度为 5.6，密度为 0.034，平均聚类系数为 0.322。节点代表人人网中的个人节点，边表示朋友关系。由于人人网是非常具有影响力的大学生交友网站，数据集分析的对象限制于好友圈子，即这 166 个节点是某个人的 166 个朋友。从图 4 中可以直观地看出好友网络已经被划分成 5 个相对独立的子社区，这与平时对人人网的直观理解相符合。而人人网的好友关系基本都是真实线下关系的反映，很自然可以划分成小学同学、初中同学、高中同学、大学同学等。由于之前对数据进行了筛选处理，即只选择了小学同学、初中同学、高中同学、大学同学、研究生 5 个社区的朋友节点，划分结果如图 4 所示。

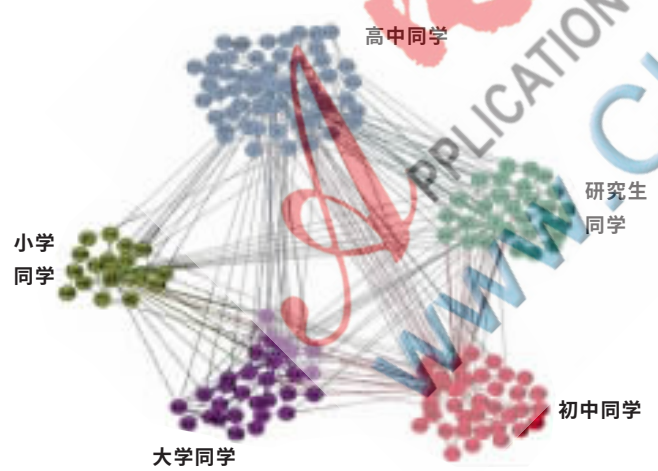

图 4 朋友网的社区划分结果

在社区划分的结果中，共划分成 5 个社区，与之前的社区基本一致，只有少数节点被错分（正确率为 88%），但整体的划分结果相对来说比较满意。用 GN 算法对数据集进行划分，共得到 7 个社区，正确率仅为 83%。利用本文算法来划分人人网朋友关系数据，不仅降低了算法的复杂度，也得到了较为正确的划分结果。

本文基于聚集系数和局部的思想，提出了一种划分在线社会网络社区结构的方法。将本文方法和目前基于全网的社区划分方法相比，该方法不需要事先知道划分的社区数目，也不需要寻找中心节点，而是从任何一个节点出发进行社区划分，从而降低了算法复杂度。与之前提出的一些局部社区划分算法相比，不仅能够适用于大规模在线社会网络，得到正确的划分结果，并且时间复杂度较低。对 Zachary 空手道俱乐部网络和人人网朋友关系网络的划分结果与实际结果基本相符，说明算法是可行的。

参考文献

[1] ADAMIC L A, BUYUKKOKTEN O, ADAR E. A social networkcaught in the Web[J]. First Monday, 2003,6(8):1-22.

[2] MISLOVE A, MARCON M, GUMMADI P K, et al. Measurement andanalysis of online social networks [C]. Internet Measurement Conference, 2007.

[3] NEWMAN M E J. Detecting community structure in networks[J]. Eur Phys J B, 2004, 38:321-330.

[4] FIEDLER M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory [J]. Czechoslovak Mathematical Journal, 1973, 23(298):619-633.

[5] POTHEN A, SIMON H D, LIOU K P. Partitioning sparse matrices with eigenvectors of graphs [J]. Siam Journal On Matrix Analysis And APPlieations, 1990, 11(3):430-452.

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

[7] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev E, 2004, 69(6):066133.

[8] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Phys Rev E, 2004, 70(6):066111.

[9] CLAUSET A. Finding local community structure in networks[J]. Phys Rev E, 2005, 72(2):26132-26137.

[10] BAGROW J P, BOLLT E M. A local method for detecting communities[J]. Phys Rev E, 2005, 72(4): 046108.

[11] 解诒，汪小帆.复杂网络的一种快速局部社团划分算法[J]. 计算机仿真，2007，24(11)：82-85.

[12] Wang Xutao, Chen Guanrong, Lu Hongtao. A very fast algorithm for detecting community structures in complex networks[J]. Physica, 2007, A384:667-674.

[13] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69 (2):026113.

[14] ZACHARY W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473.

（收稿日期：2013-03-26）

作者简介：

张瑜，女，1983 年生，硕士研究生，主要研究方向：社会计算，数据挖掘。

蔡国永，男，1971 年生，教授，博士，主要研究方向：软件形式化方法，社会计算，可信计算。