一种基于派系过滤的社区进化发现研究*

阎 艳, 黄智兴, 邱玉辉
(西南大学 计算机与信息科学学院, 重庆 400715)

摘要:在对派系过滤方法及其相关原理进行研究基础上,分析了该方法在社区进化发现中存在的参数依赖问题,提出了一种基于派系过滤的社区进化发现方法:通过生成社区树,综合多组参数的社区发现结果,可获取网络中不同耦合度的社区的层次结构,从而发现网络中社区的进化过程。本文将该方法应用在单词关联网络中,实验结果表明,该方法能够发现各社区在进化过程中的规模、成员以及耦合度方面的变化,在一定程度上,克服了传统派系过滤方法对参数的依赖性。

关键词:派系过滤;社区进化;社区树
中图分类号:TP181 文献标识码:A 文章编号:1672-6693(2009)02-0090-04

现实世界中的许多系统都可用网络表示,如 WWW、Internet、引文关系、社会关系等。这些网络都具有社区结构性质,即整个网络由若干个社区构成,社区内部的节点之间连接较为紧密,而社区之间的连接相对稀疏[1]。随着网络的发展,网络中的社区也会发生相应变化。发现网络中的社区并分析其进化过程,对了解动态网络的结构和特性具有重要意义。

派系过滤方法(Clique percolation method,CPM)是一种基于边密度的社区发现方法,它可以发现相互重叠的社区,因而常用于分析大型网络的社区结构。派系过滤方法的结果受参数值的影响。参数的取值必须恰到好处,才可能得到理想的结果。这往往需要人工对多组参数取值的计算结果进行比较和选取。随着网络的不断发展,社区在进化过程中其对应的参数条件也会发生变化。如何有效地发现动态网络中的社区及其进化过程,是派系过滤方法的一个难点。

本文提出一种基于派系过滤的社区进化发现方法,并将该方法应用在单词关联网络中。实验表明,通过建立社区树,综合多组参数的社区发现结果,可获取网络中不同耦合度的社区的层次结构,从而有效地发现网络中社区的进化过程。

1 派系过滤方法

1.1 CPM

派系过滤方法[2-3]是 Palla 等人提出的。Palla 等认为一个社区从某种意义上可以看作是一些相互连通的派系的集合。派系是一个全连通网络。由 k 个节点构成的派系叫做 k-派系(k-clique)。如果两个 k-派系有 k-1 个公共节点,则称它们相邻。若一个 k-派系可以通过若干相邻的 k-派系到达另一个 k-派系则称这两个 k-派系连通。网络中的 k-派系社区可以看成由所有相互连通的 k-派系构成的集合。例如,2-派系可以看作网络中的边,2-派系社区即表示网络中所有连通的子图。类似的,3-派系是网络中的三角形,3-派系社区是由若干个有公共边的三角形构成的子图。由于一个节点可能属于多个不相邻的 k-派系,CPM 能够得到相重叠的社区。例如,图 1 中有两个 3-派系社区分别用黑色节点和灰色节点表示,两个社区有一个节点是重叠的。

图 1 3-派系社区[3]

显然,CPM 的结果受参数 k 影响。随着 k 值的增大,社区的规模会越来越小,但社区内部的耦合度会更高,社区结构也更加紧凑。CPM 处理带权重的

* 收稿日期:2009-01-20
资助项目:国家重点基础研究发展计划(973)(No. 2003CB317008)
作者简介:阎艳,女,硕士研究生,研究方向为语义网环境下的资源管理与发现;通讯作者:邱玉辉,E-mail:yhqiu@swu.cn。

网络时,另一个重要的参数是 w^*。w^* 为边的阈值,权值小于 w^* 的边将会被忽略。随着 w^* 的增大,节点间连接变得松散,社区亦随之缩小甚至分解。因此,必须选择合适的参数取值,才能有效发现网络的社区结构。

1.2 CPMw 与 CPMd

CPMw[4] 与 CPMd[5] 是 CPM 的两种变型。CPMw 用于处理带权重的网络。与 CPM 设置边的权重阈值 w^* 不同,CPM 对派系的强度(Intensity)权重设置阈值,社区由强度高于该阈值的社区构成。社区强度计算为 $I(C)=(\prod_{i,j\in C,j<j}w_{ij})^{2/(k(k-1))}$。可见,CPMw 允许派系中包含权重较小的边。CPMd 是派系过滤方法在有向图的应用。它利用有向 k- 派系(Directed k-clique)发现有向图中的社区。在有向派系中,存在一个节点顺序,使得任意两个节点间存在一条边,该边的方向与两点的顺序一致。CPMw 与 CPMd 和 CPM 一样,社区发现的结果受参数影响。

2 基于派系过滤的社区进化发现

从派系过滤方法的原理不难看出:相同参数发现的社区之间交集大小是有限的,即小于 k-1;如果一个社区在某参数取值下被发现,那么在参数值更小时,必然能发现某个社区包含该社区;通过增大参数取值,可以发现社区内部耦合度较高的子社区。由此,可以利用社区树对多组参数值的结果进行综合,从而分析社区的进化过程。

2.1 派系过滤生成社区树

社区树是社区集合的树型结构,它清晰地表示了不同规模、不同耦合度的社区之间的层次关系。社区树可用一个四元组(V, E, P, C)表示。其中,V 是一个有穷的节点集,E 是一个有穷边集,P 为参数组集,C 为一个有穷社区集,它包含所有利用 P 中参数发现的社区;V 中每个节点 v 都用一个社区 c 及其对应的最大参数 p 唯一标记($c\in C$, $p\in P$);对于任意的两个节点 v_i, $v_j\in V$,如果存在边 $e_{ij}\in E$, e_{ij}:$v_i\rightarrow v_j$,即 v_i 是 v_j 的父节点,那么 c_j 是 c_i 的真子集且 $p_j>p_i$(p_j,$p_i\in P$)。

社区树生成算法如下。

```
CT(net)
  Input:Network net
  Output:a community tree of net
  create a vertex root;
  set root.community = network, root.parameter = {v = 0, k = 1};
  call CT_construct(root);

CT_construct(vr)
  Input:vertex vr
  Output:a subtree, of which the root is vr
  newpara = increase(vr.parameter);
  find the subcommunity set SC of vr.community, by CMP with newpara;
  if SC is empty, then return;
  for each c in subset {
  if c = vrcommunity {
    vr.parameter = newpara;
    CT_construct (vr)
  }
    else {
      create a vertex vc;
      set vc.community = c, vc.parameter = newpara;
      set vc as a child of vr by adding a new edge:vr→vc;
      CT_construct (vc);
  }
  }
```

2.2 基于派系过滤的社区进化发现

对于一个 t 时刻存在的社区 C_t,要发现其进化过程,关键在于找到 C_t 的前趋 C_{t-1} 和后继 C_{t+1}。社区的查找会用到两个系数,即相对重叠度(Relative overlap)[5] 和覆盖率。社区 A 和社区 B 的相对重叠度定义为 $C(A+B)=\frac{|A\cap B|}{|A\cup B|}$。当 $C(A,B)$ 大于阈值 thr^* 时,称社区 A 和社区 B 具有相关性。社区 A 对社区 B 的覆盖率定义为 $O(A,B)=\frac{|A\cap B|}{|B|}$。通过对相邻时刻的社区树进行搜索,可以获取指定社区相关状态的集合,算法如下。

```
locate(vr, com)
  Input:a community com, and a community tree of which the root is vr
  Output:a set of related community
  S = {};
  if O(vr.community, com) < thr*, then return S;
  if C(vr.community, com) > thr*, then S = S + root;
  for each c which is a child of root {
    S = S + locate(c, com);
  }
  return S;
```

3 实验与分析

数据集选择 2000—2007 年 DBLP 文献目录数

据集。对数据集中的文献标题完成分词、删除冠词和介词、大小写转换、消除后缀、清除非英文单词等预处理工作。基于单词在标题中的同现关系,根据文献发表的年份分别构建单词关联网络[6-11]。所得的网络是一个带权重的无向网络,顶点代表一个单词,边的权值表示两个单词同现的次数。如图 2 所示,单词关联网络中节点的度分布满足幂率分布[12]。

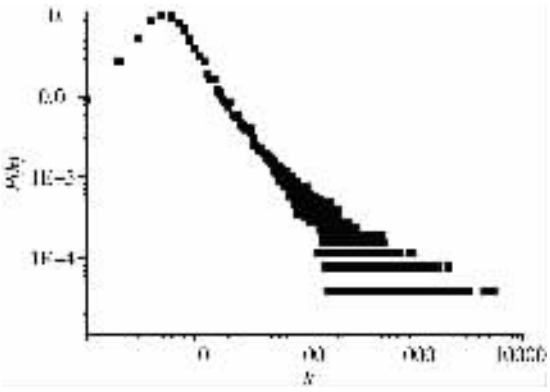

图 2　度的分布

利用 CPM 算法对每个单词关联网络分别建立社区树。其中,w^* 和 k 分别以 25、1 为步长依次进行递增。图 3 给出了 2007 年单词网络的社区树片段。利用各社区树,发现社区的进化过程。为简化起见,仅选取相对重叠率最高的相关社区作为社区的后继。

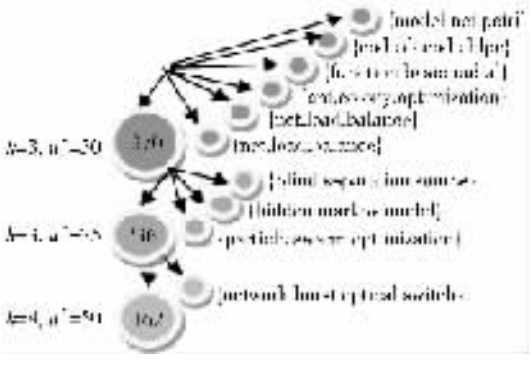

图 3　2007 年单词关联网络的社区树片段

从实验结果看,该方法能够有效发现社区的进化过程。表 1 描述了社区 $\{\text{ad hoc network}\}_{2000}$ 的进化过程。不难看出该社区在规模和耦合度方面的变化。实验还发现,大社区和小社区有着截然不同的进化特性。一个大社区总可以找到它的后继(或前趋),两者相对重叠度在 0.6 ~ 0.9 之间。这意味着大社区总是在发生变化,并且社区内部相当多的节点总是保持不变。这部分静止节点主要由高频词构成,如常用词 approach, framework, system, application, architecture, method,及术语 agent, fuzzy, wavelet, robot, database, retrieval, multimedia distribute 等。和大社区不同,相当一部分小社区是静止不变的,如 {world wide web}、{vector support machine}、{associate rule mine} 等。此外,有少数小社区找不到它的前趋,如 $\{\text{particle swarm optimization}\}_{2005}$, $\{\text{ant colony optimization}\}_{2007}$ 等。

表 1　社区 {ad hoc network} 的进化过程

时间	w^*	k	社区
2000	50	3	Ad Hoc Network
2001	50	4	Ad Hoc Mobile Network Route
2002	50	5	Ad Hoc Mobile Network Wireless
2003	75	6	Ad Hoc Mobile Network Protocol Route Wireless
2004—2007	75	7	Ad Hoc Mobile Network Protocol Route Wireless

4 结语

本文提出一种利用派系过滤建立社区树进行社区进化发现的方法,并把该方法用在单词关联网络中。从实验结果看,该方法能够发现各社区在进化过程中的规模、成员以及耦合度方面的变化。在一定程度上,该方法克服了派系过滤方法对参数的依赖性,但在利用 CPM 建立社区树的过程中,需要考虑参数的递增策略。如何选择合理的参数递增策略,从而在简化计算的同时有效发现网络的社区结构及其进化过程,还有待进一步研究。

参考文献:

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

[2] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435: 814-818.

[3] Derenyi I, Palla G. Clique percolation in random networks [J]. Physical Review Letters, 2005, 94: 160-202.

[4] Farkas I, Palla G, Vicsek T. Weighted network modules[J]. New Journal of Physic, 2007, 9(6): 180-198.

[5] Palla G, Farkas I J. Directed network modules[J]. New Journal of Physic, 2007(9): 186.

[6] Palla G, Vicsek T C. Quantifying social group evolution[J]. Nature, 2007, 446: 664-667.

[7] Zhuge H. Communities and emerging semantics in semantic link network: discovery and learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 99: 1-1.

[8] Porter M. An algorithm for suffix stripping[J]. Program: E-

lectronic Library and Information,2006,14(3):130-137.
[9] Barabasial A,Jeong H. Evolution of the social network of scientific collaborations[J]. Physica A:Statistical Mechanics and its Applications,2002,311(3):590-614.
[10] Church K,Hanks P. Abstract word association norms,mutual information,and lexicography[J]. Computational Linguistics,1990,16(1):22-29.
[11] 杨清平. 网格资源的组织与发现研究[J]. 重庆师范大学学报(自然科学版),2008,25(4):70-73.
[12] Clauset A,Shalizi C R,Newman M. Power-law distributions in empirical data[EB/OL]. [2007-01-07](2009-01-15). http://arxiv.org/abs/0706.1062v1.

Research on Community Evolution Discovery Based on Clique Percolation

YAN Yan, HUANG Zhi-xing, QIU Yu-hui

(Faculty of Computer & Information Science, Southwest University, Chongqing 400715, China)

Abstract: Clique percolation method has been used in many cases for discovering overlapping communities. However, its result is largely affected by parameters; as community size or cohesion changes the parameters needed shift accordingly. To overcome the dependency on parameters, in this paper we propose an approach for community evolution discovery based on community tree constructed by clique percolation. A community tree provides a hierarchical structure of communities discovered under a range of parameters in a given network; related community states can be found by searching a series of community trees thus the life span of a community can be discovered. We apply it to the word association network from DBLP data set and analyze how each community evolves. The outcome shows that this approach can effectively discover the community evolution process and identify the changes in size, membership and intensity. It's also observed that communities of different size have different evolving characteristic features.

Key words: clique percolation; community evolution; community tree

(责任编辑 游中胜)