基于核心图聚类的邮件网络社区发现

彭 玲,徐汀荣,乔志伟
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:根据图聚类节点的密度变化确定核心节点,构成连通子图并确定邮件网络社区划分的个数及初始社区中心点,通过社区中心动态调整的方法将非核心节点划分至所属社区。实验证明了该邮件网络社区划分的有效性和可行性。

关键词:图聚类;邮件社区划分;动态中心

中图分类号:TP301.6 **文献标识码**:B **文章编号**:1674-7720(2010)17-0081-03

A mail community partition based on coring graph clustering

PENG Ling, XU Ting Rong, QIAO Zhi Wei
(College of Computer Science, Suzhou University, Suzhou 215006, China)

Abstract: Based on the greedy algorithm with graph clustering by computing the density variation sequence and identifying core nodes, number of communities, partition the certain nodes to some belonged community with the similarity of characteristics of communication behavior, and readjust the centrality of the communities. The real datasets shows the feasibility and effectiveness of the proposed algorithm.

Key words: graph clustering; mail community partition; dynamic centering

随着互联网的发展,网络成为工作和生活中相互联系的重要工具,与此同时,网络社区[1]也应运而生。网络社区是指基于电脑网络的虚拟社会关系网络的载体,在这种网络中,相同类型的节点之间存在较多的连接,而不同类型节点之间的连接则相对较少。网络社区通常满足六度分割理论及150法则[2]等特征,通常是现实社区的近似。因此对网络社区的发现,对了解现实社会中的社会关系网络有着特别重要的意义。

邮件社区作为一种网络社区,同样与现实的社会关系网络同构,满足小世界网络模型[3],由于电子邮件本身的优势[4],有利于发现社会关系网络的动态特性。目前有关网络社区发现的研究较多,具有代表性的方法有Girvan和Newman提出的基于去边的G-N算法[5]、Aaron和Newman的层次聚类的算法以及基于三角环的Radicchi方法[5]等。但这些算法时间复杂度高,不易于处理大规模的网络。

本文从图论的角度出发,首先采用基于贪心的图形聚类算法,通过计算节点序列密度变化确定社区划分的个数及每个社区的核心代表节点,然后以这些核心节点为中心,采用基于节点相似度的动态中心调整算法将节点划分到其所属的社区中,从而完成对邮件社区的划分。

1 基本概念

邮件网络是社会网络的一种,可以采取社会网络的相关方法对邮件网络进行社区划分。为了便于算法的描述,首先对文中使用的社会网络图及相关概念进行数学描述和说明。

(1)网络图。通信结构包括收件人和发件人的联系信息,为了表示通信次数和方向,本文选择有权有向网络图对邮件网络进行表示。设$G=(V,E,W)$,其中,V是全部顶点v_i的集合,表示电子邮件的收件人或发件人;E是所有边e_{ij}的集合,其中与每个顶点v_i直接相连的顶点被定义为A_i,则$A_i=\{v_j|e_{ij}\in E\}$;W为边的权重集合,任意两个节点v_i、v_j若有$e=(v_i,v_j)$或$e=(v_j,v_i)$则表明v_i、v_j之间存在通信联系,而$W(e)\in W$,可以用邮箱v_i、v_j的通信次数来描述。

(2)节点的度。节点v_i的度$\deg(v_i)$定义为与顶点v_i具有连接关系的所有顶点的数量:$\deg(v_i)=\sum_j e_{ij}$,出度outdeg

(v_i)为该节点向外发送邮件的数量的度量，入度 $\mathrm{indeg}(v_i)$为节点接收邮件的数量度量。

(3)节点的密度。设节点 $i \in H \subseteq V$，则 i 关于 H 的局部密度为：

$$d(i,H)=\frac{1}{|H|}\left(\sum_{j\in H} w_{ij}+\sum_{j\in H} w_{ji}\right) \tag{1}$$

其中，w_{ij} 为节点 i 向节点 j 发送邮件的次数，而 w_{ji} 为节点 j 向节点 i 发送邮件的次数。

$$D(H)=\min_{i\in n} d(i,H) \tag{2}$$

$D(H)$表示 H 内密度最弱的节点集合。

(4)网络社区中心。在一个包含 m 个节点 $v_1,\cdots,v_m$ 的社区中，设 X_i 记录邮箱 v_i 与其他邮箱的通信次数，则社区中心：

$$\bar{v}=\frac{1}{m}\sum_{i=1}^{m} X_i \tag{3}$$

它是社区中的节点与其他节点相互关联的平均值，是整个社区的代表节点。

(5)社区密度及社区有效直径[3]。设 G_k 是 G 的子图，表示一个社区。社区 G_k 的密度记作 $D(G_k)$，定义为所有节点出入度之和与节点个数之比，即：

$$D(G_k)=\sum_{i=1}^{n}(\mathrm{indeg}(v_i)+\mathrm{outdeg}(v_i))/n \tag{4}$$

社区有效直径记作 $R(G_k)$，定义为 G_k 中至少有 90% 以上的节点对，它们的距离小于或等于 $R(G_k)$。

(6)节点 v 与社区中心 $\bar{v}$ 的相似度。为了便于公式的描述，设 X 记录了节点 v 与其他节点的通信次数，Y 记录了社区中心与其他节点的平均通信次数。则：

$$\mathrm{Sim}(v,\bar{v})=\frac{XY^{\mathrm{T}}}{|X||Y|} \tag{5}$$

2 挖掘社会关系网络

基于电子邮件的社会网络分析可以分为两个步骤：(1)利用邮件日志信息构建有向有权社会网络，网络图中顶点表示电子邮件的收件人或发件人，连线表示两节点之间存在收发联系；(2)使用改进的算法来挖掘隐含在此网络图中的社会关系，即对网络图进行社区挖掘。

2.1 构建社会关系网络

根据电子邮件地址直接构造一个有向有权图，图中顶点表示联系人，连线表示顶点之间联系的收发频率。为了减少噪声节点对整个网络的影响，在构图之前，通过设定阈值，选取收发邮件大于该阈值的节点，然后构造社会关系网络图并通过邻接表和逆邻接表进行存储，从而有利于节省空间并方便节点的出入度计算(在本文中，选取 $\mathrm{indeg}(v_i)\geqslant 3$ 以及 $\mathrm{outdeg}(v_i)\geqslant 3$ 的节点，即阈值为 3)。

2.2 邮件社区划分

邮件社区划分的基本思想：从图的划分以及聚类的角度出发，分别从节点的密度变化和节点对之间的相似度进行考察，并采取社区中心动态调整的技术，主要分为以下步骤：

(1)通过检测网络图中节点的密度变化，确定聚类个数及聚类核心；

(2)节点的划分和各社区中心的调整。

2.2.1 聚类个数及聚类核心求解算法

假设每一个邮件网络图中都存在一个被称为“聚类核心点”的高密集区域，这些聚类核心点被密度稀疏的点所包围。则聚类核心中的节点被称为核心节点，核心节点的集合为核心节点集，而包含这些核心节点的子图叫做核心子图。

根据式(1)、(2)求解出每个节点的局部密度以及集合 H 中密度最弱的节点集合。因此，通过分析最小密度值 $D(H)$的变化，可以近似地求出所有的核心节点集。即如果密度最小的节点存在于稀疏区域，则当删除该节点时 D 值会增大，那么下一个被删除的节点必将存在于密度更高的节点区域内。如果某个节点的删除引起 D 值的急剧下降，则该节点很有可能存在于密度较高的区域，也有可能因为节点的删除导致了周围其他节点的密度下降。算法(1)给出了计算密度序列变化的执行步骤。

输入：图 $G=(V,E,W)$。

输出：节点密度变化 D 及对应的节点集合 M。

算法：

①H 初始值为所有节点，参数 t；

②repeat；

③根据式(1)、(2)计算 $d(i,H)$和 $D(t)$，$M_t=\{i|i\in H\wedge D(t)=d(i,H)\}$；

④如果集合 M_t 包含 2 个或以上相互连接的单个子图，则取相互之间连接数最小的节点集合；

⑤$H=H-M_t$，$t=t+1$；

⑥until 集合 H 为空

计算出节点密度变化序列之后，可以通过式(6)来识别核心节点集。

$$R_t=(D_t-D_{t+1})/D_t>\partial \tag{6}$$

其中，∂为 0 到 1 之间的可调参数，并且参数∂的选择必须要保证社区划分满足以下两条规则[5]：(1)最小组件规则。社区中节点个数必须≥6；(2)社区稳定性规则。社区节点个数约 120 最为稳定。若集合 M_t 满足式(6)，则集合 M_t 为一核心节点集。

找出了所有满足条件的核心节点集，可以通过多种方法将这些核心节点集划分为核心子图并最终确定聚类个数及聚类核心点。由于邮件网络图为稀疏图，由核心节点所构成的连通子图称为核心子图，核心子图个数为聚类个数，则核心子图中的节点就为聚类核心点。

2.2.2 社区划分算法

算法(2)描述了邮件网络社区划分步骤，ENCD(Email

Network Community Detecting)邮件网络社区划分算法。

输入:图 $G=(V,E,W)$。

输出:社区号及每个社区对应的节点。

算法:

①输入求出的核心子图个数 K 及核心子图节点;

②repeat;

③for $t=(T,T-1,K,2,1)$;

④ for 每个社区中心 c_j; //寻找与集合中相似度最大的社区中心 c_j

⑤ if 集合 M_t 中包含非核心节点记作 x_i,计算 x_i 与 c_j 的相似度,若 $\mathrm{Sim}(x_i,c_j)\geqslant\beta$ 则将 x_i 的社区号记为 j 并将其加入到社区 c_j 中; //考虑了一个节点属于多个社区的情况

⑥for 社区网络中每个社区 community[j]

⑦调整该社区的社区中心; //根据式(3)计算

⑧until 所有的社区中心均未改变。

3 社区划分实验

本文所选实验数据集为 enron 邮件数据集以及苏州大学 2009 年 2 月至 5 月之间的邮件日志内容。其中在 http://www-2.cs.cmu.edu/~enron/可下载到 enron 数据集,它包括预处理后的 151 个节点以及 252 759 条边。苏州大学邮件日志内容有 183 925 个节点以及 391 347 条边,内容含校内邮箱间的邮件收发信息以及相关的校外邮箱的邮件收发信息。

在本实验中,由于考虑到苏州大学邮箱用户的隐私问题,对邮箱地址进行了 MD5 转换,并且对于每个邮箱 mailbox,均由一个邮箱编码 mailboxID 唯一标识。

本实验的实验环境为 2.80 GHz Pentium CPU,1GB 内存,80 GB 硬盘,操作系统为 Microsoft Windows XP,程序开发平台为 Myeclipse。社区划分结果的每一项由两个字段组成:邮箱编码 mailboxID 和社区标号 CommunityID。

图 1 是 enron 数据集构成的社会网络图,图 2 显示了本文参数对 enron 数据集社区划分最终结果的影响。由图可以看出,不同的 ∂ 值对确定社区划分数量影响不大。

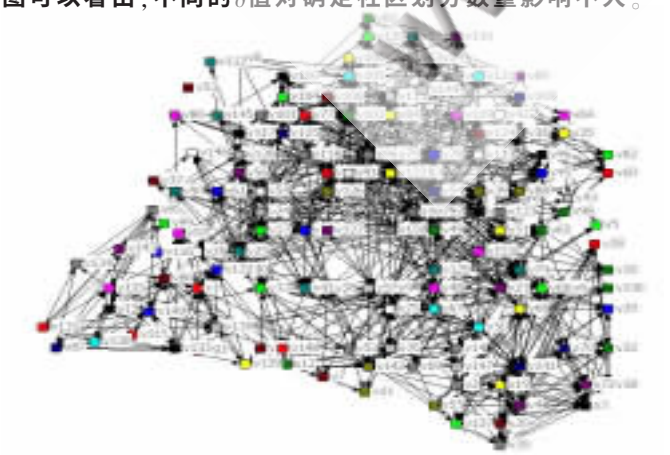

图 1 enron 数据集所构成的社会网络图

表 1 显示了使用本文算法与 G-N 算法在 enron 以及苏州大学邮件数据集上的结果比较,其中 ∂ 取值 0.26。modularity[5]为算法的评价指标之一,通常为(0,1)的小数,且值越大说明社区划分的质量越高。图 3 描述了本文算法与 G-N 算法在 enron 数据集上社区密度的对比。图 4 是在 enron 数据集上社区有效直径的对比图。

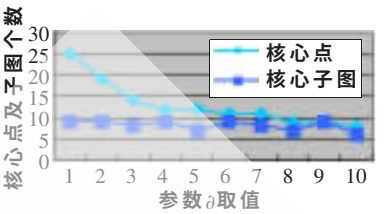

图 2 参数 ∂ 对结果的影响性

表 1 G-N 与 ENCD 算法在相同数据集上的结果对比

算法	数据集	Modularity	社区数
G-N	enron	0.372	8
	苏州大学邮件	0.296	47
ENCD	enron	0.369	9
	苏州大学邮件	0.301	50

图 3 不同算法的社区密度对比图

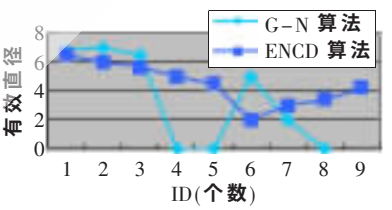

图 4 不同算法在 enron 数据集上的有效直径对比图

由表 1 可以看出,本文提出算法在社区划分个数方面与 G-N 相当。但是,G-N 算法用于 enron 数据集所发现的社区结果中,有两个社区中的节点只有 1 个,还有一个社区的节点个数为 2 个。这显然与社区划分步骤(1)矛盾,而本文提出的算法所求出的每个社区中节点个数则相对比较平均。图 3 中 G-N 算法得出的社区密度最小为 5,最大约为 20;而本文算法得出的最小值约为 10,并且最大峰值为 25。可见,本文算法在 enron 上划分的社区内部联系更加紧密。图 4 说明了 G-N 算法与 ENCD 算法得出的社区有效直径均相当,因此本文算法用于网络社区划分是可行的。

4 分析与评估

邮件社区的划分本质上是稀疏图聚类的问题,而此

类划分又是 NP 完全问题[6]。Girvan 和 Newman 提出的基于去边的 G-N 算法，在图划分上得取得了很好的效果，但是时间复杂度较高，为 $O(E^2V)$，其中，E 为网络图中边数、V 为节点数。因此，不利于大规模的网络社区发现。而本文算法(1)由于使用了邻接表及逆邻接表结构，所以，时间复杂度为 $O(|E|+|V|\log|V|)+O(|V|)+O(|E_c|)$，其中 E_c 为核心图中的连接边数。算法(2)时间复杂度为 $O(E)$。所以最坏的间复杂度为 $O(|E|+|V|\log|V|)$。苏州大学邮件数据集上测试的结果表明，本文方法执行效率要优于 G-N 算法。

本文提出了一种新型的基于核心图聚类算法的邮件网络社区划分，首先通过计算节点的密度变化找出满足条件的核心节点，然后将这些核心节点集划分为核心图，最后通过节点相似度将未划分的节点划分到最相似的子图中。在 enron 以及苏州大学邮件数据集上的结果表明，本文算法在社区划分的质量上与 G-N 算法相当，但是执行效率要高于 G-N。此外，本文算法还支持一个节点属于多个社区的情况，而这种情况在现实生活中是极为常见的。

参考文献

[1] ZHANG Y C, YU Xj, HOU L Y. Web communities: Analysis and construction[M]. Berlin: Springer, 2005.

[2] STROGATZ S H. Exploring complex networks [J]. Nature, 2001(410):268-276.

[3] 陈绍宇，宋佳兴，刘卫东，等.关系网络:一种基于小世界模型的社会关系网络[J].计算机应用研究，2006，23(5):194-197.

[4] ALSTYNE M V, ZHANG J. EmailNet: A system for automatically mining social networks from organizational email communication[C]. In NAACSOS2003, 2003.

[5] MARK E J, NEWMAN M. Finding and evaluating community structure in networks [M]. Physical Review E,69. 026113, 2004.

[6] GIRVAN M, NEWMAN M. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci. USA 2002 (99):8271-8276.

(收稿日期：2010-03-05)

作者简介：

彭玲，女，1984 年生，硕士研究生，主要研究方向：智能化信息处理。

徐汀荣，男，1959 年生，硕士生导师，教授，主要研究方向：智能化信息处理、网络。