

基于 CN-M 的邮件网络核心社团挖掘

胡天天,戴航,黄东旭

(西北工业大学 自动化学院,陕西 西安 710072)

摘要:在当今互联网时代,电子邮件的快速、低耗等特性,使其成为人们生活和工作中的必需工具。为了智能化地提取和分析邮件网络中的海量数据,以从海量邮件数据中挖掘潜在的有价值的信息,将社会网络分析方法应用于邮件网络分析,提出了基于 CN-M (Core Node-Modularity) 的邮件网络核心社团挖掘算法。首先用 JavaMail 对数据进行解析,将解析后的数据保存在数据库中,使用这些数据来构建邮件网络图,根据节点的连接中心度、紧密中心度和中间中心度计算加权中心度,由加权中心度最大的节点开始,根据模块度指标进行核心社团的挖掘。实验结果表明该算法可以很好地挖掘邮件网络中潜在的核心社团。

关键词:社会网络分析;邮件网络;核心社团;加权中心度;模块度

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2014)11-0009-04

doi:10.3969/j.issn.1673-629X.2014.11.003

Mining Core Community from Mail Network Based on CN-M

HU Tian-tian, DAI Hang, HUANG Dong-xu

(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: With the rapid development of the network, e-mail with its fast, low cost and other characteristics, is becoming the necessary tools in people's work and life. In order to extract and analyze the massive data from mail network intelligently, and to grub the potential of valuable information from massive mail data, apply social network analysis method to the mail network analysis, and propose email network core community mining algorithm based on CN-M (Core Node-Modularity). First, use JavaMail to parse mail data, and store the analyzed data in the database. Second, use these data to construct the mail network diagram, according to connection center degree, close center degree and intermediate degree centrality to calculate weighted centrality. Starting from the center of the largest weighted node, based on modularity index, mine the core community. Experimental results show that the algorithm can mine the potential core community from mail network well.

Key words: social network analysis; mail network; core community; weighted centrality; modularity degree

0 引言

在当今互联网时代,电子邮件的快速、低耗等特性,使其成为人们生活和工作中的必需工具,也是人与人之间通信的重要方法。据芯片巨头英特尔调查,在上网的每1分钟时间内,有2.04亿封电子邮件被寄出。如何从海量的邮件中挖掘出有价值的信息,如何迅速而有效地从纷繁复杂的邮件网络中挖掘出隐藏在其中的核心社团是近年来的一个研究热点。国内外研究者和科学家提出一些有效而实用的算法。例如 Girvan 和 Newman 提出基于中介度的社团挖掘算法(GN 算法)^[1], 该方法是社团划分的重要方法之一,但是该算法的时间复杂度过高,不适合海量数据的社团

挖掘。文献[2]提出基于邮件分类的社团挖掘技术,该算法可以很好地挖掘出邮件网络的社团结构,但是并没有找到核心社团。文中将社会网络分析方法^[3]应用于邮件网络分析,提出了基于 CN-M (CoreNode-Module) 的邮件网络核心社团挖掘算法。首先对邮件数据进行解析,使用解析后的数据构建邮件网络图;然后根据各个节点的连接中心度、紧密中心度和中间中心度^[4-5], 计算各个节点的中心度,由节点的中心度计算各个节点的加权中心度,然后由加权中心度最大的节点开始挖掘核心社团,当满足“150 法则”时停止,找到模块度极值最大的节点 Q, 此时的社团即核心社团。实验结果表明,该算法可以很好地挖掘邮件网络中潜

收稿日期:2013-12-23

修回日期:2014-03-28

网络出版时间:2014-07-28

基金项目:2012 教育部博士点基金(20126102110036);中航航空科学基金(2012ZC53042)

作者简介:胡天天(1988-),男,河南人,硕士研究生,研究方向为网络与信息安全;戴航,硕士生导师,研究方向为网络与信息安全。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140728.1221.001.html>

在的核心社团。

1 邮件数据的预处理

1.1 电子邮件系统及邮件数据介绍

一个电子邮件系统主要由三个部分组成:用户代理、邮件服务器和邮件发送协议(SMTP)与读取协议(POP3 或 IMAP)。通用因特网邮件扩充 MIME(Multi-purpose Internet Mail Extensions)是目前互联网电子邮件所普遍遵循的邮件技术规范和格式标准。在格式规范中定义了邮件两个主要部分:信封和内容。邮件内容头部有一些关键字段:From:代表发件人的邮箱地址,即通信的发送方;To:代表收件人的邮箱地址,即通信的接收方;Subject:代表邮件的主题,对邮件内容进行概括;Date:代表邮件的发送时间。对于存在邮箱别名的问题暂时不考虑,邮箱别名问题的解决在文献[6]中有详细的解决方案。邮件数据中的每个邮箱的帐号作为一个主体,任何两个主体之间有收发邮件的行为就代表这两个主体之间有联系。一个标准的邮件头信息如图 1 所示。

```
Received: from mail82. jd. com (unknown [58.83.158.28])
  by edm4 (Coremail) with SMTP id icCowEA5dEtcNNS_s2pBQ--
  -.8425S2;
  Mon, 13 Jan 2014 16:14:20 +0800 (CST)
  Date: Mon, 13 Jan 2014 16:14:20 +0800 (CST)
  From: =? utf-8? B? 5Lqs5LicSkQuY29t? = <customer_service
  @jd. com>
  To: test@163. com
  Message-ID: <2061663908.74660521389600860635.JavaMail.
  admin@a01-r02-d1405-i3-112>
  Subject: =? UTF-8? B? 5Lqs5Lic5ZWG5ZOB5Yiw6LSn6YCa55
  +l? =
  MIME-Version: 1.0
  Content-Type: text/html; charset=utf-8
  Content-Transfer-Encoding: quoted-printable
  X-CM-TRANSID: icCowEA5dEtcNNS_s2pBQ--.8425S2
  Authentication-Results: edm4; spf=pass smtp.mail=customer_
  service@jd. c
  om;
  X-Coremail-Antispam: 1Ufl29KBjvdXoW7GF1xXF1xJFWfWw
  47Jw47Jwb_yoWktFcEgr
  W5Z34FgF43ua4UJr1Fqw4UZwn8K3ykCFZxKa1Ivan0kFnFvFsrF
  4D Za4SqrW5tFn8C3Z7
  Z98Z3yYgw48KjkaLaAFLSUrUUUUub8apTn2vfkv8UJUUUU
  8Yxn0WfASr-VFAUDa7-sFnT
  9fnUUvcSsGvfC2Kfnxn1yJw4j6rW3J3BIK3ZELjIFyTuYvjfU
  2fOzDUUUU
```

图 1 标准的邮件头信息

1.2 数据预处理

现实世界中数据大体上都是不完整、不一致的原始数据,无法直接进行数据处理,或处理结果差强人意。为了提高数据挖掘的质量需要进行数据的预处理,数据的预处理就是将数据转化成算法容易理解的格式。对邮件数据的预处理就是对邮件内容头部的解析和提取过程,文中使用 JavaMail 对邮件数据进行预

处理,采用基于 MIME 格式标准对邮件数据内容进行解析,主要是将邮件内容头部的 From、To、Date 字段内容提取出来,将提取出来的内容数据存储在 Mysql 数据库相关表中,然后通过主体之间的通联关系可以构建邮件社会网络图。数据预处理过程如图 2 所示。

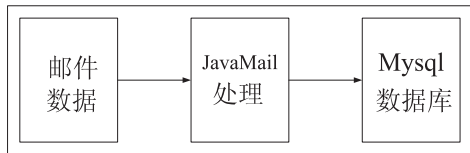


图 2 数据预处理模块

2 相关理论及算法

社会网络分析是研究一组行动者关系的研究方法。一组行动者可以是人、社区、团体、组织或国家等。社会网络分析法通过分析网络中的关系来研究网络的结构及属性特征,包括网络中的个体属性及网络整体属性^[7]。网络的整体属性分析包括小世界效应、小团体研究、凝聚子群等。网络个体属性分析包括点度中心度、接近中心度等;社团的概念相当于模块、群、数据挖掘中的簇等,特点表现为社团内部连接比较紧密,社团之间连接相对稀疏。文中将社会网络分析的方法应用于邮件网络分析中,来挖掘邮件网络中潜在的核心社团。

关系是社会网络存在的重要标志,社会网络结构是通过各种关系建立起来的。搜集网络中的关系数据进行分析,可以得到社会关系的网络结构,通过分析社会网络结构,可以挖掘隐藏的社团结构。在进行核心社团挖掘之前,首先对邮件网络图及相关概念进行形式化的描述和说明。

2.1 邮件网络的加权中心度

(1) 邮件网络图的概念:通信结构包括收件人和发件人的联系信息,为了表示通信次数与方向,选择有向图对社会网络形式化。有向图的描述如下:在邮件网络图 $G(V, E, N)$ 中, V 是网络图中全部顶点的集合,顶点代表收件人或发件人; E 是图中所有边的集合; N 代表网络图中的顶点数。一个简单的邮件网络图如图 3 所示。

(2) 连接中心度和连接度:连接中心度代表这一个节点的连接能力和它在网络中的活跃程度,可以使用与该节点(Node)有通联关系的节点数目来表示,连接中心度越大,代表与该节点有通联关系的节点越多,节点 k 的连接中心度表示如下:

$$\sum_i g_{ik}, i \neq k \quad (1)$$

其中, g_{ik} 表示节点 i 与节点 k 相连接的节点。

则节点 k 的连接度定义为:

$$G_c(k) = \frac{\sum_i g_{ik}}{N}, i \neq k \quad (2)$$

其中, $G_c(k) \in [0, 1]$ 。

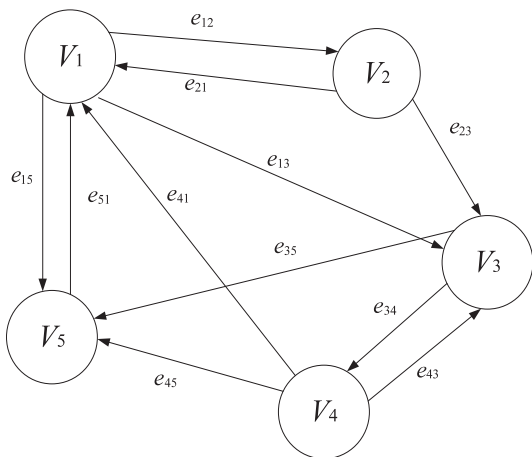


图 3 简单的邮件网络图

(3) 中间中心度和中间度: 中间中心度描述节点在网络中所能起到的“传递”能力, 中间中心度可以简单理解为经过该节点的最短路径数。中间中心度越大代表其他节点越有可能通过它建立通联关系。节点 k 的中间度表示如下:

$$G_b(k) = \sum_i \sum_j \frac{q_{ikj}}{q_{ij}}, i \neq k \neq j \quad (3)$$

其中, q_{ikj} 表示节点 i 到节点 j 的最短路径经过节点 k ; q_{ij} 表示节点 i 到节点 j 的最短路径; $G_b(k) \in [0, 1]$ 。

(4) 紧密中心度及紧密度: 紧密中心度表示一个节点到网络图中其他所有节点的速度, 可以使用该节点到其他节点的最短路径之和来表示, 一个节点的紧密中心度越小, 说明该节点到其它节点速度越快。紧密度表示如下:

$$G_i(k) = \sum_i \sum_j \frac{h_{ik}}{h_{ij}}, i \neq k \neq j \quad (4)$$

其中, h_{ik} 表示节点 i 到节点 k 的最短路径之和; h_{ij} 表示节点 i 到节点 j 的最短路径之和; $G_i(k) \in [0, 1]$ 。

(5) 节点 k 的中心度及加权中心度。

节点 k 的中心度可以表示为:

$$G_c(k) = G_c(k) + G_b(k) - G_i(k) \quad (5)$$

节点 k 的加权中心度表示为:

$$G_{cw}(k) = G_c(k) + \frac{\sum_i G_{ki}(k)}{n} \quad (6)$$

其中, $G_{ki}(k)$ 表示与节点 k 相连接的节点 i 的中心度; n 表示与节点 k 相连接的节点数目。与节点 k 相连接的节点中心度越大, 则节点的加权中心度也就越大。

2.2 邮件网络模块度

模块度^[8]: E_{ij} 表示邮件网络中连接两个不同社团的节点的边在所有边中所占的比例, 这两个节点分别为社团 i 和社团 j 。 E_{ii} 就代表社团 i 内部的所有边, $\sum_i E_{ii}$ 表示所有的社团内部的边。 $a_i = \sum_j E_{ij}$ 定义为所有与社团 i 内任意节点相连接边与所有边中的比值, 模块度表示为:

$$Q = \sum_i (E_{ii} - a_i^2) \quad (7)$$

模块度越大表明社团划分的越好, Q 最佳值的范围为 0.3 ~ 0.7, 采用基于贪婪算法进行凝聚社团划分时, 社团划分的结果是一个树状图^[9], 在树状图中找到模块度 Q 的最大值, 此时就表示划分的社团结构是最优的。

2.3 小社团稳定性

罗宾·丹巴通过研究不同形态的原始社会, 发现村落中的成员都在 150 名左右, 人们将他的理论称为“150 法则”^[10]。 社交网站 Facebook 中人们的平均好友数量为 120 人^[11], 也符合“150 法则”的理论。

基于 150 法则, 文中认为在社会网络中, 一个稳定的社团人数一般不会超过 150 人。 人数过多的时候, 大社团往往可以分裂为几个更小的社团^[12-13], 而这些小社团的模块度比大社团更大。

2.4 核心社团挖掘算法

将数据预处理过后的每个邮箱账号看作是一个主体。 通过主体之间的通联关系来构建邮件网络图。 至此, 可以进行核心社团的挖掘。 文中提出基于 CN-M 的核心社团挖掘算法, 具体描述如下:

输入: 邮件网络图 $G(V, E, N)$;

输出: 邮件网络中的核心社团。

核心社团挖掘算法:

(1) 使用公式(2) ~ 公式(4) 分别计算各个节点连接度、中间度和紧密度。

(2) 使用公式(5) 计算邮件网络图 G 中所有节点的中心度 $G_c(k)$ 。

(3) 由所有节点中心度使用公式(6) 计算各个节点的加权中心度 $G_{cw}(k)$ 。

(4) 找到加权中心度最大的核心节点。

(5) 假设邮件网络图中每个独立的节点是一个社团, 并假设加权中心度最大的节点为核心社团。

(6) 合并与核心社团相连接的社团, 并计算模块度 Q 的值, 新的核心社团 = 核心社团 + 新加入的社团, 并保存 Q 的极大值。

(7) 终止条件: 核心社团的顶点数满足“150 法则”, 否则继续执行步骤(5)。

(8) 在 Q 的极大值之中, 找出最大的极值点。 最

大极值时的社团就是核心社团。

3 实验及结果

首先将邮件数据进行预处理,邮件数据有 592 个联系人地址和 9 079 对邮件往来关系,将处理后的数据存储在数据库中,根据这些数据来构建邮件网络图,然后使用基于 CN-M 的核心社团挖掘算法进行处理。通过对实验结果的分析可知:此算法的时间主要消耗在加权中心度的计算上,挖掘核心社团时要找到模块度最大的极值,对这组邮件数据进行核心社团挖掘得到模块度 Q 的极值分别为 0.176 和 0.325,在极值 $Q=0.176$ 时核心社团的节点数为 17,在极值 $Q=0.325$ 时核心社团节点数为 71。 Q 值在 0.325 时所得到的社团就是挖掘的核心社团,此时 Q 的值也在最佳范围内,得到的核心社团共有 71 个节点。核心节点数与模块度 Q 值变化关系如图 4 所示。

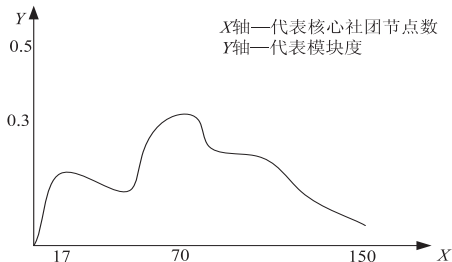


图 4 节点数和模块度变化曲线

4 结束语

随着社会网络分析研究的不断深入,将社会网络的分析方法应用于邮件网络中,来发现邮件网络中潜在的核心社团。文中首先使用 JavaMail 对邮件数据进行预处理,然后使用基于 CN-M 的核心挖掘算法进行核心社团的挖掘。实验结果表明该算法能很好地发现邮件网络中潜在的核心社团。

参考文献:

- [1] Girvan M, Newman M E J. Community structure in social and biological networks[J]. PNAS, 2002, 99(12): 7821-7826.
- [2] 段 丹, 郭绍忠, 李志博, 等. 基于邮件分类的敏感社团挖掘技术[J]. 计算机应用, 2007, 27(12): 3039-3041.
- [3] 刘 军. 社会网络分析导论[M]. 北京: 社会科学文献出版社, 2004.
- [4] Dwyer T, Hong S, Dirk K, et al. Visual analysis of network centralities[C]//Proceedings of the Asia Pacific symposium on information visualization. Tokyo: [s. n.], 2006: 189-197.
- [5] Newman M E J. A measure of betweenness centrality based on random walks[J]. Social Networks, 2005, 27(1): 39-54.
- [6] Bird C, Gourley A, Swaminathan A. Mining email social networks[C]//Proceedings of the 2006 international workshop on mining software repositories. [s. l.]: [s. n.], 2006: 137-143.
- [7] 社会网络分析与 SNS 网站[EB/OL]. 2005. <http://www.arcp.cn/portals/0/snaandsns.pdf>.
- [8] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.
- [9] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.
- [10] Dunbar R. Grooming, gossip, and the evolution of language[M]. US: Harvard University Press, 1998.
- [11] Marlow C A. Primates on Facebook[M]. [s. l.]: Economist, 2009.
- [12] Leskovec J, Lang K J. Statistical properties of community structure in large social and information networks[C]//Proceeding of the 17th international conference on World Wide Web. [s. l.]: [s. n.], 2008: 695-704.
- [13] Danon L, Duch J, Diaz-Guilera A, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics Theory and Experiment, 2005, 29(9): P09008.

(上接第 8 页)

- channel assignment for WLANs[J]. Mobile Computing and Communications Review, 2005, 9(3): 19-31.
- [10] Hsu Chih-Cheng, Liang Yian, Garcia J L, et al. Distributed flexible channel assignment in WLANs[C]//Proc of WCNC. Shanghai: IEEE, 2013: 493-498.
 - [11] 许国军, 沈连丰. 一种改进的 WLAN/WPAN 中自适应预测随机接入信道分配算法[J]. 应用科学学报, 2004, 22(4): 423-427.
 - [12] Chen J, Chen Y. AMNP: Ad Hoc multi-channel negotiation

protocol for multi-hop mobile wireless networks[C]//Proc of ICC. [s. l.]: IEEE, 2004: 3607-3612.

- [13] Hung W C, Leon-Garcia A. A dynamic multi-channel MAC for Ad-Hoc LAN[C]//Proc of 21st biennial symposium of communication. [s. l.]: [s. n.], 2002: 31-35.
- [14] 毛建兵, 毛玉明, 冷楚鹏, 等. 基于 802.11 的多信道 MAC 协议性能分析[J]. 计算机研究与发展, 2009, 46(10): 1651-1659.
- [15] 曹立明. 图论及其在计算机科学中的应用[M]. 北京: 清华大学出版社, 2002.

基于CN-M的邮件网络核心社团挖掘

作者：[胡天天](#)，[戴航](#)，[黄东旭](#)，[HU Tian-tian](#)，[DAI Hang](#)，[HUANG Dong-xu](#)

作者单位：[西北工业大学 自动化学院, 陕西 西安, 710072](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(11)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201411003.aspx