软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2808−2823 [doi: 10.13328/j.cnki.jos.004720]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于 R-C 模型的微博用户社区发现*

周小平[1,2,3], 梁 循[1], 张海燕[1]

[1](中国人民大学 信息学院,北京 100872)

[2](北京建筑大学 电气与信息工程学院,北京 100044)

[3](北京市建筑安全监测工程技术研究中心,北京 100044)

通讯作者: 梁循, E-mail: xliang@ruc.edu.cn

摘 要: 在微博市场营销、个性化推荐等应用中,发现兴趣和网络结构双内聚的用户社区起着至关重要的作用.现阶段,绝大多数的用户社区发现算法往往将用户联系与用户内容相隔离,从而导致其社区发现结果不够合理,而少数综合用户联系和内容的用户社区发现算法较为复杂;LCA 算法是重叠社区发现算法中算法效率较高且社区质量较好的算法,然而,其在聚类时未考虑边的真实兴趣体现.针对这些问题,构建了以关注关系为网络节点、以关注关系之间是否有共同用户为关注关系潜在的边、以关注关系所关联用户的兴趣集的交集为关注关系的兴趣特征,构建微博网络 R-C 模型,并探讨了其进行微博用户社区发现的方法,分析了该方法的复杂度.最后,以新浪微博数据集为实验,对照节点 CNM 算法和 LCA 算法,从兴趣内聚和网络结构内聚两方面进行分析,发现该方法能够发现更好的微博用户社区.

关键词: 微博;社区发现;关注关系;重叠社区

中图法分类号: TP311

中文引用格式: 周小平,梁循,张海燕.基于 R-C 模型的微博用户社区发现.软件学报,2014,25(12):2808−2823. http://www.jos.org.cn/1000-9825/4720.htm

英文引用格式: Zhou XP, Liang X, Zhang HY. User community detection on micro-blog using R-C model. Ruan Jian Xue Bao/ Journal of Software, 2014,25(12):2808−2823 (in Chinese). http://www.jos.org.cn/1000-9825/4720.htm

User Community Detection on Micro-Blog Using R-C Model

ZHOU Xiao-Ping[1,2,3], LIANG Xun[1], ZHANG Hai-Yan[1]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Electrical & Information Engineering, Beijing University of Civil Engineering & Architecture, Beijing 100044, China)

[3](Beijing Engineering Research Center of Monitoring for Construction Safety, Beijing 100044, China)

Corresponding author: LIANG Xun, E-mail: xliang@ruc.edu.cn

Abstract: Detecting user communities with denser common interests and network structure plays an important role in target marketing and self-oriented services. User-Generated content and the relationship between the users are often separated in the current methods on community detection, which results in the unreasonable community structures. Though some methods tried to combine the two factors, they are complex. Link community algorithm (LCA) is an efficient state-of-art method on overlapping community discovery. However, LCA does not take into account the real interest characteristics when calculating the similarity between the links. To solve the issues on user community detection on Micro-blog, this paper proposes a R-C model which takes the user relationships as the network nodes, treats the intersection of the interest characteristics of the two users in a link as the link's interest characteristics, and makes the shared user between two links as the underlying link between the links. Also, the community detection method based on the R-C model is discussed,

* 基金项目: 国家自然科学基金(71271211); 北京市自然科学基金(4132067); 中国人民大学自然科学基金(10XNI029); 北京高等学校青年英才计划(21147514040)

收稿时间: 2013-11-13; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

and the complexity in clustering is analyzed. Finally, compared with node CNM and LCA, the method using R-C model is proved to be better in finding closer relationship and denser common interest user communities.

Key words: micro-blog; community detection; following relationship; overlap community

社区发现是指在社会网络中发现内聚的子群.社区发现是社会网络分析的重要问题,它有助于人们进一步认识、理解和掌握所研究的复杂网络对象,进而实现更深入的应用研究,例如个性化推荐[1]、朋友推荐[2]、大规模网络压缩求解[3]、异质网络分析[4]、社会网络演变[5]等.兴趣和网络结构双内聚的用户社区发现,是精准的市场营销和准确的个性化推荐服务等的重要研究内容[6−8].现实生活中,人们往往传播其所能接触到的感兴趣的信息.因此,好的用户社区发现应同时满足网络结构和兴趣双方面的内聚.网络结构是社区内部节点间信息传播的桥梁,兴趣是信息传播的原因.

得益于移动互联网的发展,微博用户规模及其社会影响力迅速增长.Twitter 有不少于 5 亿的注册用户,每月活跃用户为 2.3 亿,而日活跃用户为 1 亿,每天推文 5 亿次(http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city).新浪微博也拥有超过 5 亿的注册用户,每天有高达 4.62 千万的活跃用户和不少于 1 亿的微博(http://news.xinhuanet.com/ info/2013-02/21/c_132181760.htm).微博是现实社会的缩影,它为人们提供了巨量的有价值的研究数据.人们使用微博进行政治[9]、市场营销[10]等活动,微博已成为一个公认的发表意见与看法的平台[11].

目前,针对微博用户社区发现的方法大致可分为 3 种:

(1) 基于用户内容[12−16].将用户微博内容进行兴趣特征提取,然后,基于兴趣特征进行用户聚类.该类方法忽略了微博网络结构(关注关系)在信息传播中的桥梁作用;

(2) 基于用户联系[17−25].提取微博网络的关注或好友关系,将问题转化为图论等问题进行社区发现.该类方法没有考虑用户的兴趣特征,因此,无法证明其兴趣的内聚性;

(3) 综合方法[26,27].将微博内容和用户联系相结合,基于内容提取基于兴趣的用户社区,基于用户联系提取基于联系的用户社区,再采用某种方法将两个社区进行融合,形成兴趣和网络结构双内聚的用户社区.该类方法由于需要进行两次社区发现,且需要进行社区融合;因此,算法效率较低.

真实情况中,用户往往对多种兴趣感兴趣,每个用户都应根据其兴趣归属于多个用户社区.因此,用户社区实际上是一个重叠社区.LCA 算法[28]是目前较好的重叠社区发现算法,它以边为单元进行边聚类,从而根据边分属不同的社区,将节点划分到多个不同的社区.LCA 算法较好地平衡了社区发现中兴趣和网络结构双方面的因素,并且其聚类复杂度较低;然而,LCA 算法在边之间的相似度计算上忽略了边的真实兴趣特征.

关注关系是微博网络形成的基础,也是微博信息传播的纽带.关注双方往往因为某个或某几个共同兴趣而建立关注关系.因此,关注关系还体现了关注关系双方的共同兴趣特征.针对现有微博用户社区发现算法的不足,本文以关注关系作为聚类节点,根据用户微博内容提取关注关系的兴趣特征,构建微博网络 R-C 模型,进而根据关注关系的兴趣特征计算关注关系之间的相似性,将问题转化为加权无向网络社区发现问题,解决了现有算法考虑不周全或效率较低等问题.本文的学术贡献主要有:

(1) 提出了微博网络 R-C 模型,并探讨了其进行用户社区发现的方法;

(2) 分析了基于 R-C 模型进行社区发现聚类的时间复杂度;

(3) 以新浪微博为实例,对照节点 CNM 算法和 LCA 算法,分析了基于 R-C 模型的用户社区发现算法所发现的用户社区在兴趣和网络结构上都有更优的内聚性.

本文第 1 节介绍现阶段微博用户社区发现的相关文献.第 2 节详细描述微博网络 R-C 模型及其社区发现方法,并分析使用该方法进行聚类的时间复杂度.第 3 节以新浪微博数据为实验对象,对照节点 CNM 和 LCA 算法,从兴趣内聚和网络内聚两方面验证本文方法能够发现更好的微博用户社区.第 4 节对本文的工作进行总结.

1 相关工作

近几年,随着在线网络社区的发展,社区发现算法得到了广泛的研究.针对用户社区的发现,人们已经提出

了许多方法,这些方法主要可以分为 3 类:文本聚类法、网络结构法和综合法.

文本聚类法主要通过计算社区内节点的文本内容的相似性,根据相似性将文本内容相似的节点划分为社区.早在 1999 年,Kleinberg 等人提出了基于内容的网页聚类方法,即著名的 HITS 算法[12].主题模型是文本聚类法最典型的算法.2003 年,Blei 等人提出了 LDA 模型[13],认为文档是多个主题的概率分布.2004 年, Syeyvers 等人认为主题是多个关键词的概率分布,用户也以某种概率分布对多个主题感兴趣,并提出了 AT (author-topic)模型[14],用于发现用户、文档、主题和关键词之间的关系.2007 年,McCallum 等人基于发送-接受关系提出了 ART(author-recipient-topic)模型[15],用于聚类具有相似兴趣的用户.在 ART 模型的基础上,2008 年, Pathak 等人提出 CART(community-author-recipient-topic)模型[16].这些模型都忽略了用户之间显著的关注关系,从而可能导致社区发现结果的不合理.

基于网络结构的社区发现算法是目前较为流行且研究较多的方法,这类方法根据用户之间的相互关系将社区网络划分为社区内联系紧密、社区之间联系稀疏的多个子社区.1970 年,Kernighan 和 Lin 针对图分割问题提出了 KL 算法[17],该算法应用于复杂网络社区发现,就是社区发现图分割法的典型算法.图分割法通过迭代的方式将图分解为最优的两个子图,反复处理,直至得到足够数目的子图.2002 年,Girvan 和 Newman 提出了 GN 算法[18],它通过反复识别和删除网络中边介数最大的连接,实现复杂网络聚类.GN 算法的复杂度较高,但它启发了人们对复杂网络社区发现的思路.2004 年,Newman 和 Girvan 提出的网络模块性评价函数——模块度 Q[19].Q 函数为社区内的实际连接数目与随机连接下社区内的期望连接数目之差,它描述了所发现社区的优劣.Q 值越大,社区结构越好.在此基础上,Newman 提出了基于局部搜索的快速复杂网络聚类算法,即快速 Newman 算法[20].快速 Newman 算法通过局部搜索,找到极大化的 Q 值,从而实现社区划分.同年,Newman 等人从算法复杂度的角度出发,通过引入模块度增量矩阵和堆结构,将快速 Newman 算法演进为了 CNM 算法[21].2005 年,Guimera 和 Amaral 以优化目标函数 Q 为目标,提出基于模拟退火(simulated annealing,简称 SA)算法的复杂网络聚类算法——GA 算法[22].SA 的引入,使得 GA 算法具有找到全局最优解的能力,因而,GA 算法具有很好的聚类精度.基于模块度优化的聚合方法是目前比较流行的社区发现算法,并被扩充到了加权网络社区发现[23]、有向网络社区发现[24]和重叠社区发现[25]等.虽然基于网络结构(用户关系)的社区发现算法能够对用户进行聚类,但由于其忽略了用户之间的共同兴趣特征,因此不能保证社区发现的兴趣内聚性.

针对上述两种社区发现在兴趣社区发现上的不足,2012 年,Zhang 等人[26]提出了将用户关系同用户内容进行结合,发现用户社区.他们采用 NMF 方法进行基于用户关系的社区发现,采用 AT 模型用于兴趣社区的发现,并在此基础上将两种社区发现结果进行融合,并在 Tweets 和 Delicious 上进行了验证.燕飞等人[27]首先对个人兴趣进行聚类,得到基于兴趣的行动者社区,然后使用社会网络拓扑结构信息对兴趣社区进行扩展,并在 Flickr 上进行了实验分析.这些方法虽然得到了较好的兴趣社区发现,并能将用户根据其兴趣划分到多个不同的社区,符合实际情况,但其算法逻辑较为复杂,而且复杂度较高.

真实世界中,社区结构大多数都是重叠且具有层次结构[28],微博用户往往具有多样化的兴趣特征,因此微博用户社区发现是重叠社区发现问题.CPM 算法[29]是目前流行的重叠社区算法,其在自然和社会学等领域[30,31]都有所应用,且被推广到了加权网络的重叠社区发现[32].然而 CPM 算法认为社区是强连通的簇,其对社区苛刻的定义使得在稀疏网络(如新浪微博用户联系网络[33,34])中社区发现效果较差.此外,CPM 算法需要指定 k 值,且复杂度较高,制约了 CPM 算法在大数据网络中的运用.2010 年,Ahn 等人提出了边社区概念及其算法——LCA 算法[28],并在生物网络、社会网络和其他代表性网络(哲学家关系网、单词关系网和 Amazon.com 产品联系网)上对照 CPM 算法、Infomap 算法[35]和快速 Newman 算法[20]验证了 LCA 算法能发现质量更好的重叠社区.

LCA 算法以边作为聚类节点,对边进行聚类,并根据边所属的社区,将节点划分到多个不同的社区.在一个具有 N 个节点的加权网络中,LCA 算法假定对于任意节点 i 都有属性向量 $\boldsymbol{a}_i=(\tilde{A}_{i1},...,\tilde{A}_{iN})$,且:

$$\tilde{A}_{ij}=\frac{1}{k_i}\sum_{i'\in n(i)}w_{ii'}\delta_{ij}+w_{ij},$$

其中,w_{ij} 为边 e_{ij} 的权重;$n(i)$为与节点 i 有连接关系的所有邻居节点集合;k_i 为集合 $n(i)$的元素数量;当 $i=j$ 时,$\delta_{ij}=1$,