虚拟社区中的社团结构研究与分析

朱永真[1],夏正友[1],卜 湛[1],刘新建[2]
(1.南京航空航天大学 信息科学与技术学院,江苏 南京 210016;
2.东南大学 信息科学与工程学院,江苏 南京 210096)

摘 要:针对虚拟社区中成员身份不真实、成员之间关系模糊、社团结构未知的特点,提出基于空间和时间对虚拟社区进行社团划分的两种算法,最终在线了社团内部成员之间的交互关系和社团之间的关联关系。然后利用共同信息比较法和D函数比较法分析了两种划分下的社团结构,并利用其相似性和相异性对虚拟社区中的社团进行了评价。实验结果表明,不同的划分标准会影响到最终的社团结构,且划分得到的社团之间不是绝对的彼此独立,而是具有重叠关联性。

关键词:虚拟社区;社团结构;社团划分;共同信息比较法;D函数比较法
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2011)01-0046-04

Research and Analysis on Community Structure in Virtual Community

ZHU Yong-zhen[1],XIA Zheng-you[1],BU Zhan[1],LIU Xin-jian[2]
(1. College of Information and Techn., Nanjing University of Aeronautics and Astronautics, Nanjing 210016,China;
2. College of Information Science and Engineering, Southeast University,Nanjing 210096,China)

Abstract:;Since the characteristics of members is not true, the membership of virtual community is fuzzy, the characteristics of its association structure is unknown, two division algorithms of the virtual community associations division on space and time are presented. The relations between members and the interaction between communities are also presented at last. Comparative law based on common information and .) function is applied to analyze the two division of societies, with the similarities and diversity to evaluate the virtual community of societies. Experimental results show that the different standards will affect the final structure of societies, and the societies is not absolutely in complete isolation from one another, but in association with overlapping.

Key words:virtual community; community structure; community partitions; common information comparative law; D function comparative law

0 引 言

一个社团通常由具有相似特征或共同爱好的成员组成,社团内部各节点的连接紧密,而社团之间的连接却比较松散[1,2]。揭示网络中的社团结构,对于了解网络结构与分析网络特性有着重要的价值,不仅如此,社团结构分析在生物学、物理学、计算机图形学和社会学中也有着广泛的应用[3,4]。

笔者以在线虚拟社区为依托,利用真实而丰富的社区数据[5]作为研究对象,对虚拟社区的成员交互关系和社团结构展开研究。虚拟社区中成员的身份是虚拟的、不确定的[6],社团结构是未知的[7],这给研究带来了一定困难。鉴于此,文中提出基于空间和时间对虚拟社区中社团划分的两种算法,考虑到现实中的网络大多不是绝对的孤立,文中社团的定义允许重叠。最终划分出了社区的社团结构,获取了每个社团的具体成员,并在线了内部成员之间的交互关系和社团之间的关联关系。同时由于虚拟社区中社团结构的未知性,无法与真实的社团结构相比来判断划分的正确性,因此文中利用共同信息比较法和D函数比较法分析了两种划分下的社团结构,最后利用其相似性和相异性对划分结果进行了评价。

1 虚拟社区的网络数据

文中以天涯社区中的“国际观察”版(http://www.tianya.cn/publicforum/articleslist/0/worldlook.shtml)为数据源,获取了2005-2009五年用户的行为数据。首先通过广度优先的搜索策略,获取所有帖子的链接信息,再利用网页爬虫技术获取每一个帖子的内容。最终获取了321027条用户信息,96073条主题

收稿日期:2010-05-17;**修回日期**:2010-08-24
基金项目:南京航空航天大学基本科研业务费专项基金资助项目(100456Y1012)
作者简介:朱永真(1984-),女,河南开封人,硕士研究生,研究方向为虚拟空间网络与信息安全;夏正友,副教授,研究方向为计算机网络与人工智能。

信息,5241338 条回复关系信息。

实验表明,主题数据具有极大的异质性,即只有少数的主题引起了较大部分用户的兴趣,这部分主题的出现使大量用户之间有了联系。具体的说,143 个的主题(0.15%)被回复次数大于 1000,而 5110 个主题(5.32%)的被回复次数不超过 20,94.53% 的主题被回复次数(21-1000)居中。成员之间的交互关系直接依赖于主题,即 A 成员与 B 成员之间有联系,并不直接因为 B 是谁,而是因为 B 所发表的主题引起了 A 的兴趣,进而对其进行讨论,或者 B 讨论了 A 所发表的主题。为此,我们建立了主题数据库,描述用户参与每个主题的情况,并抽取出发布主题的作者用户及其回复者用户。实验结果显示,用户参与主题的总数量一般不超过 20,即用户一般只参与讨论特定的主题。假设一个主题对应一个兴趣爱好,则可认为参与讨论相同主题的用户具有相同的兴趣爱好,基于此,就可以从一定程度上分析社区中的社团结构[8]。

2　基于两种方法划分虚拟社区

按照社团形成的过程,其划分思路可以分为四类:凝聚过程、分裂过程、搜索过程和其他过程。然而不管是分裂算法还是聚类算法,最终目的都是将网络划分为若干个互相分离的社团。但是,现实世界中很多网络并不存在绝对的彼此独立的社团结构;相反,它们由许多彼此重叠互相关联的社团组成,在这种情况下,很难单独地将这些社团划分出来。因此,文中利用两种不同的分类方法,从两个方面来分析这种互相重叠的社团结构。

2.1　基于空间的社团划分

虚拟社区中的用户行为一般包括:发布主题、浏览主题、回复主题,且不同的用户发布、浏览或回复的主题一般不同。文中主要研究基于主题回复的用户行为。假设不同类型主题对应不同的兴趣爱好[9],那么参与相同或相似类型主题的用户具有相同或者相似的兴趣爱好,则认为这些用户之间存在一定程度的相似性或"亲疏关系"。文中利用用户之间的"距离"来衡量这种相似性,所谓"距离"指的是欧氏距离[10],用户 i 和 j 之间的距离 d_{ij} 表示为:

$$d_{ij}=[(x_i-x_j)^T(x_i-x_j)^{1/2}]=[\sum(x_i-x_j)^2]^{1/2} \quad (1)$$

算法思想是:以回复者用户回复相同论题的数量作为衡量距离的标准,数量越多,用户之间"相对距离"就越近,则"相似度"就越高。具体实现步骤如下:

1)首先从数据库表"主题和回复者 PostandAnswer"中提取出所有的回复者用户信息,找出每一个用户参与讨论的所有主题 ID。

2)计算交互的每对回复者用户共同回复的主题数量。

3)将每个用户看成是一个独立的社团。

4)以主题 ID 和共同回复主题数量为参考,计算交互的每对用户之间的"距离",根据距离,计算交互的用户之间的相似性。

5)用户之间的相似性即是社团之间的相似性,合并最相似的类为一个新类。

6)重复 4)和 5) 至所有的社团都满足条件。

实验结果为:(1)得到了每个社团的内部成员,在线了成员之间的交互关系;(2)划分了社区的社团结构,在线了社团之间的关联关系。如图 1 为基于时间划分所得社团 7 的所有成员,其中,All_Follow_ID 列为成员的 ID。图 2 中社团之间的连接线表示社团之间有重叠,即社团之间不是绝对的孤立的。

Cases Classified to:

社团 7

All Author ID	All Follow ID	All Post ID
51172	632950	118556
51172	409198	124196
53874	925018	127094
51358	492780	129998
50102	614392	132501
54927	291801	132820
51091	270049	134217
51169	270991	144992

图 1　基于空间划分的社团内部成员

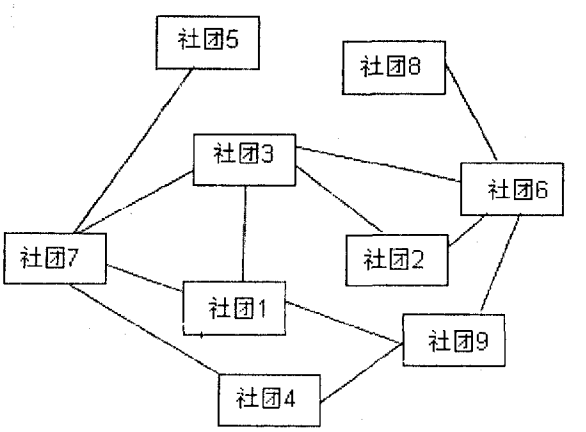

图 2　基于空间划分的社团之间的连接关系

2.2　基于时间的社团划分

虚拟社区中的用户一般在一个特定的时间参与讨论某一主题,若用户在相同时间段内参与讨论相同或者相似的主题,则认为用户之间具有相同或者相似的兴趣爱好,即用户之间存在相似性。与基于空间的相似性计算不同,基于时间的相似度计算是根据用户在相同时间段内参与同一主题的次数作为评估标准,数

越多,表示用户之间"相对距离"就越近,相似性就越高。

实现的具体步骤如下:

(1)从数据库表"主题和回复者 PostandAnswer"中提取出所有的回复者用户信息,找出每一个用户参与讨论的所有主题 ID,并记录下相应的参与时间。

(2)将较活跃的用户作为初始中心点,计算各个节点与中心点的相似性。

(3)根据相似性最高的原则进行节点合并,至合并"距离"的长短符合要求。

(4)重新计算各类的中心点,返回(3),至所有的社团都满足条件。

类似于2.1,实验结果为:1)获取每个社团内部的成员信息,在线了社团内部成员之间的交互关系;2)划分了社区的社团结构,在线了社团之间的关联关系。如图3显示了基于时间划分所得社团1的所有成员。图4中社团之间的连接线表示社团之间有关联。虽然不同的划分算法得到不同的社团结构,但是实验结果同样论证了虚拟社区中的社团之间具有一定的重叠性。

Cases Classified to:
社团 1

All Author ID	All Follow ID	All Post ID
6089	1135741	100387
3187	1150383	101583
2874	1046637	102847
2475	1230021	105411
4252	651434	110285
8200	1049008	110948
5376	702187	118651

图3 基于时间划分的社团内部成员

3 社团划分结果的评价与分析

不同的算法往往将同一个网络划分成不同的社团结构。对于社团结构已知的网络,与真实的社团比较可以得到划分的准确性;而对于结构未知的网络,多种划分结果的比较同样可以加深对网络结构的理解和网络特性的分析。

天涯社区中交互的用户之间无须考虑对方的年龄、住址、社会地位等状况,也无从知道对方的真实身份,因而用户之间的真实关系是模糊的,社区真实的社团结构是未知的。因此,无法利用真实的社团结构来判断划分的正确性。鉴于此,文中分别利用共同信息比较法和 D 函数比较法[11,12]两种方法,对划分得到的社团结构进行分析和评价。

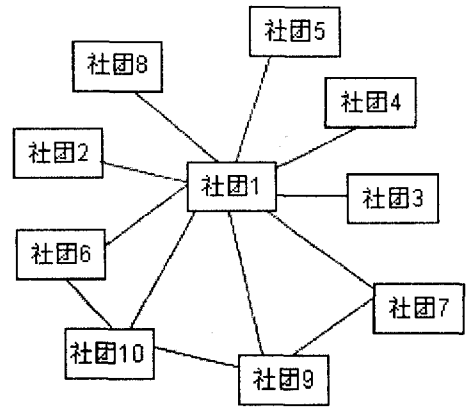

图4 基于时间划分的社团之间的连接关系

3.1 基于共同信息比较法评价社团划分结果

实现过程为:首先定义一个混乱矩阵 N,其中行元素为基于空间划分得出的社团成员的 ID,列元素为基于时间划分得出的社团成员 ID,矩阵 N 中的元素 N_{ij} 是既在基于空间划分的社团 i 中,又在基于时间划分得出的社团 j 中出现的顶点的个数。基于信息理论所得两种社团结构 A、B 的相似程度为:

$$I(A,B)=-\frac{2\sum_{i=1}^{cA}\sum_{j=1}^{cB}N_{ij}\log\left(\frac{N_{ij}N}{N_iN_{.j}}\right)}{\sum_{i=1}^{cA}N_i\log\left(\frac{N_i}{N}\right)+\sum_{j=1}^{cB}N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \quad (2)$$

其中,基于空间划分得出的社团的个数用 cA 表示,基于时间划分得出的社团个数用 cB 表示,N_i 为 N_i 的第 i 行的加总,$N_{.j}$ 为 N_{ij} 的第 j 列的加总。

如果两种划分结果所得社团结构完全一致,则 $I(A,B)$ 达到最大值1;当所得社团结构没有任何重叠时 $I(A,B)$ 达到最小值0。可见,$I(A,B)$ 值越大,两种划分之间的差异越小。计算得到两种社团结构的相似度为0.725。比较结果表明,不同的划分算法会影响最终的社团结构。

3.2 基于 D 函数比较法评价社团划分结果

两种划分结果差异性可以分解成社团对之间差异的总和,文中就是采用这一思路讨论两种划分结果间的差异性。首先将划分得到的社团视为集合,划分得到的社团结构就是一组集合,社团间的差异表现为集合中的不同元素。设 A、B 是任意两个集合,定义 $A\cap B$ 为两个集合的相似度,$(A\cap\bar{B})\cup(\bar{A}\cap B)$ 是两个集合的相异度。标准化后 A、B 的相似度(s)和相异度(d)为:

$$s=\frac{|A\cap B|}{|A\cup B|} \quad (3)$$

$$d=\frac{|(A\cap\bar{B})\cup(\bar{A}\cap B)|}{|A\cup B|} \quad (4)$$

对于基于空间和时间两种划分方法得到的社团结

构,采取的比较原则是:

1)建立两种社团结构之间的匹配关系:分别将两种社团结构中的社团进行对比,相似度最大的两个社团组成一对,然后根据相似度排序把各个社团配对。文中所得两种结构内部的社团数目不同,即,基于空间划分得到的社团数目为 9,而基于时间划分所得社团数目为 10,则将多出的社团与空社团配对。

2)根据配对计算每对社团的相异度。

3)综合每对社团的相异性,得到两种划分的相异度的数值:

$$D = \frac{\sum d_{xy}}{k} \tag{5}$$

其中,X, Y 为配对的社团;$k = 10$,为社团对的总数。

D 函数的取值范围是[0,1],值为 1 表示两种划分完全不同,取值为 0 表示两种划分完全相同,取值越大说明两种划分之间的差异越大。采用这种比较法得到 $D = 0.312$,即相似度为 0.688。较 3.1 的结果,社团结构间的差异较大。

4　结束语

针对虚拟社区用户行为特点和社团特点,基于时间和空间两种算法对虚拟社区中的成员关系和潜在的社团结构展开相关研究,划分结果表明不同的划分算法会影响到虚拟社区中的社团结构;并利用共同信息比较法和 D 函数比较法,对两种划分下的社团结构进行了分析和评价,验证了实验结果。

文中给出的算法允许社团之间的关联重叠,这具有一定的现实意义。然而,社团内部各个节点的连接比社团间节点的连接要紧密的多,只是一个定性的指标。所以今后的工作中,将从定量角度来更加精确地分析社团的结构。

参考文献:

[1] 汪小帆,李　翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006:162-193.

[2] Fortunato S, Castellano C. Community structure in graphs[J/OL]. Eprint arXiv, 2007, 0712:2716. [2009-03-10], http://ww. arXiv. org.

[3] Grivan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci, 2001, 99:7821-7826.

[4] Gleiser P, Danon L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6:565-573.

[5] 荣　波,夏正友,朱永真,等. BBS 在线网络及其成员交互特性研究[J]. 复杂系统与复杂性科学,2009,6(4):57-65.

[6] 荣　波,夏正友. 基于聚类的 BBS 成员交互网络特性研究[J]. 重庆科技学院学报,2009, 12(6):165-169.

[7] 范彦静,王化雨. 社会合作网络中社团结构的搜索算法研究[J]. 信息技术与信息化,2008,2:13-15.

[8] 王爱平,王占凤,陶嗣干,等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展,2010,20(4):105-108.

[9] 陈海强,程学旗,刘　悦. 基于用户兴趣寻找虚拟社区核心成员的方法[J]. 中文信息学报,2009,23(2):89-94.

[10] 范　明,孟小峰. 数据挖掘概念与技术[M]. 第 2 版. 北京:机械工业出版社,2007:85-100.

[11] Kuncheva L I, Hadjitodorov S T. Using diversity in cluster ensembles[C]//2004 IEEE International Conference Systems, Mqn and Cybemetics. [s. l.]:[s. n.],2004:1212-1219.

[12] Zhang P, L iM H, Wu J S, et al. The analysis and dissimilarity comparison of community structure[J]. Physica A, 2006, 367:577-585.

(上接第 37 页)

参考文献:

[1] Van den Bergh, Engelbrecht F. Effects of Swarm Size on Cooperative Particle Swarm Optimisers[C]//In:Proceedings of the Genetic and Evolutionary Computation Conference(GECCO). San Francisco, USA:[s. n.],2001.

[2] 李　丽,牛　奔. 粒子群优化算法[M]. 北京. 冶金工业出版社,2009:67-72.

[3] Liu B, Wang L, Jin Y-H, et al. Improved particle swarm optimization combined with chaos [J]. Chaos Solitons & Fractals, 2005, 25(21):1261-1271.

[4] 纪　震,廖惠连,吴青华. 粒子群算法及应用[M]. 北京. 科学出版社,2008:12-15.

[5] 于舒娟,张志涌. 含公零点 SIMO 信道 QPSK 序列盲检测[J]. 东南大学学报,2005,36(6):867-871.

[6] 祁　辉,熊　鹰,周树民. 基于粒子群算法的整数规划问题的求解算法[J]. 江汉大学学报(自然科学版),2009(3):14-18.

[7] 邹　毅,朱晓萍,王秀平. 一种基于混沌优化的混合粒子群算法[J]. 计算机技术与发展,2009,19(11):18-22.

[8] 崔海青,刘希玉. 基于粒子群算法的 RBF 网络参数优化算法[J]. 计算机技术与发展,2009,19(12):117-119.

[9] 刘　莉,张志涌. 基于 PSO 优化算法的 QPSK 信号盲分离[J]. 中南大学学报(自然科学版),2005(8):38-41.

[10] 张志涌,杨祖樱. MATLAB 教程[M]. 北京:北京航空航天大学出版社,2006.

[11] 张爱华,江中勤,张　华. 基于粒子群优化算法的分形图像压缩编码[J]. 计算机技术与发展,2010,20(2):21-24.

[12] 汪松泉,程家兴. 遗传算法和模拟退火算法求解 TSP 的性能分析[J]. 计算机技术与发展,2009,19(11):97-100.