基于认知度的用户好友社团关系挖掘方法*

孟 聪，黄永峰，应励志
（清华大学 电子系，北京 100084）

摘 要：如何挖掘网络用户好友的社团关系是社会网络领域研究热点之一。人人网、Facebook 等网络的用户好友关系是通过用户注册信息来表征的，但对 BBS 和微博等网络用户来说，无法采用注册信息来表征好友关系。因此，针对 BBS 和微博等网络的用户间互动性这一特征，引入了认知度概念来描述用户发帖和回帖互动行为的联系紧密度，在此基础上提出了一种基于用户间认知度的用户好友社团关系挖掘算法；同时提出了一种好友社团关系的可视化呈现方法。该方法可以直观展现 BBS 用户友好群体分布和用户分类情况。上述方法在水木清华等高校 BBS 数据集进行了实验和验证。

关键词：BBS；认知度；社团关系
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2833-04
doi:10.3969/j.issn.1001-3695.2012.08.008

Community relationship mining method based on users' cognition degree

MENG Cong, HUANG Yong-feng, YING Li-zhi
(*Dept. of Electronics, Tsinghua University, Beijing* 100084, *China*)

Abstract: How to unearth relationships of network users is a hot spot of social network research fields. In some famous social network websites, such as RenRen.com and Facebook, relationship is characterized by users' registration information, but for users of BBS and micro-blog, the relationship can not characterize by registration information. Therefore, according to features of BBS users, this paper introduced the concept of cognition degree to describe the behavior of users' interactions. Furthermore, it proposed a community relationship mining algorithm based on users' cognition degree. At the same time, it put forward a visual presentation method. This method could intuitively show communities of BBS users and their classifications. Experiments on http://smth.edu.cn and other BBS systems show the method above has a good performance.

Key words: BBS; cognition degree; community relations

0 引言

电子布告栏系统(bulletin board system,BBS)是互联网最早的应用之一，也是典型的信息交流平台。随着计算机网络的发展，越来越多的人利用此系统来讨论热点焦点问题、发表自由意见，逐步发展成为一种不可或缺的交流和共享资源的场所。

传统 BBS 一般提供发帖、回帖、看帖等功能，且包含上百个版面(以水木清华为例，一共 367 个版面)，每天版面上都有大量新帖。由于版面众多，信息更新频率高，大部分用户只倾向于在自己感兴趣的版面中获取信息。这种以信息为中心、兴趣为导向的模式，在 BBS 中形成了一些具有小世界特性的群体[1,2]。另一方面，BBS 系统中的好友关系是非公开的，随着社交化网络的快速增长，用户对个性化、社会化的需求越发强烈，因此挖掘 BBS 中潜在的社交关系，不仅能够为 BBS 在社交化方面的演化和发展提供支持，同时能够为舆情监测等应用提供有力保障。

本文主要针对 BBS 数据的特点，提出一种基于用户间认知度的用户好友的社团关系挖掘算法，以校园最典型的 BBS 水木社区为实现载体，采用该算法对水木清华的数据进行处理、分析，建立了用户好友社团关系图。

1 社会网络关系分析相关研究

虚拟社会也被称为虚拟社区(virtual community)[3]，是指不同区域和阶层的人通过网络连接形成的有情感交流甚至物质交易的社会。

在虚拟社会研究领域，最有代表性的研究工作是 Freeman 的社会网络分析(SNA)模型。该模型用来分析社会成员之间的相互关系，探索成员间的相互影响规律，发现潜在的社会结构[4,5]。该模型定义了三种最常用基本概念[6]：连接度(degree)、边介数(betweenness)、紧密度(closeness)。社会被描述成一个复杂网络[7]模型[8]，节点 V 是信息载体也就是社会网络中的人，连接边 E 表示节点间信息传递也就是社会中人与人的交互关系。网络中节点与其他节点相连的个数称做连接度；节点的边介数被定义为网络中任意两个节点间的最短路径经过该节点的次数；网络中节点到其他所有节点最短路径的距

收稿日期：2011-12-14；**修回日期**：2012-02-02　　**基金项目**：清华大学自主科研计划资助项目(20111081023)

作者简介：孟聪(1985-)，男，辽宁沈阳人，工程硕士，主要研究方向为搜索引擎(83028735@qq.com)；黄永峰(1967-)，男，副教授，博士，主要研究方向为信息隐藏、搜索引擎等；应励志(1987-)，男，硕士研究生，主要研究方向为搜索引擎.

离之和称为紧密度。连接度的高低说明该成员对网络影响的强弱;边介数高低说明该成员对交换信息、促进信息传播贡献的大小;紧密度的高低说明该成员在网络中扮演的角色重要与否。

另外,在网络数据中社区关系挖掘算法也是当前虚拟社会领域研究热点之一。为了在复杂网络中挖掘社团关系,许多研究工作提出了各自的聚类分析算法。例如,Kernighan-Lin 算法[9]是一种试探优化法,基本思想是提出一个为网络划分的增益函数 Q,定义 Q = 团体内边数 - 团体间边数,找出 Q 值最大的划分方法。

文献[10,11]提出了基于 Laplace 矩阵的谱平分法。n 个节点的无向图 $G=(V,E)$ 的 Laplace 矩阵是一个 $n\times n$ 的对称阵 L,元素 L_{ij} 表示节点 i 与节点 j 的连接关系,有连接值为 -1,否则为 0。谱平分法的理论基础是理论上不为 0 的特征值所对应的特征向量的各元素中,同一个团体内的节点对应的元素是近似相等的。考虑网络中仅存在团体社团这种特殊情况,这样除了最小特征值 0 以外,矩阵 L 其他特征值对应的特征向量总是包含正、负两种元素。当网络由两个团体构成时,可根据非 0 特征值对应的特征向量的元素对节点分类,正元素代表一个团体,负元素代表另一个团体。

另外,文献[12]也提出基于标准矩阵 $N=K^{-1}A$ 的谱平分算法,该算法是在传统谱分析基础上提出的。定义了平凡特征向量为行标准化转换得到矩阵 N 的最大特征值总为 1 时相应的特征向量。假设有 k 个团体数目,那么该矩阵有接近于 1 的第一非平凡特征值的数为 $k-1$,并且同一社团节点的元素应该相当接近,这 $k-1$ 个特征向量的元素呈阶梯状分布,数目刚好等于社团数目 k。因此,研究 $k-1$ 中的一个元素,便可通过元素相似性将节点分为相应的 k 个社团。

分析上述典型研究工作可以看出,传统方法利用图的相关概念,分析社交网络中用户的拓扑关系,但随着用户数量的增加,使得直观表述用户关系拓扑图存在一定的困难。本文采用社会网络分析模型的三个概念:连接度、边介数、紧密度,通过对 ID 回复关系、ID 间最短路径、ID 认知程度等 BBS 数据特征进行分析,提出了一种基于用户间认知度 BBS 社团关系挖掘算法。该算法在分析用户拓扑关系图的基础之上,以部分关键用户为基准,建立用户群体的向量空间,能直观展现 BBS 社团关系结构。

2 基于认知度的用户好友社团关系挖掘方法

2.1 用户间认知度的定义及计算方法

针对各类 SNS(social network site)即社交网站或社交网对好友定义的差异,比如 Facebook、人人网和腾讯 QQ 是基于用户注册信息来表征的,好友关系是公开透明的;而 BBS 和微博中的用户好友关系是不可见的。用户之间联系紧密程度的判定主要是基于用户间相互的发帖和回帖关系来表征的。

因此,本文基于 BBS 这一数据特征给出 BBS 用户间认知度定义。

定义 1 用户间认知度。通过用户间相互发贴和回帖关系来定量反映用户间联系紧密的程度。

本文中分析的 BBS 数据主要有两个来源:文集和讨论区。由于这两方面数据特征有所不同,因此处理过程存在一定的差异性。

首先,本文分析对讨论区数据处理的方法,认为 BBS 用户间的认知度是通过用户间发帖互动得以体现的,同时在不同主题下用户的发帖数目表征了该用户对此话题的参与程度。为了定量刻画用户间的认知度,下面引入了如下定义:

定义 2 在 BBS 中,设主题帖集合 T_{u_i} 为某用户 u_i 参与讨论的 n 个不同的主题帖集合,其中 T_i 是用户 u_i 参与的第 i 个主题($i\in N,0<i<n$)的回帖数,则 $T_{u_i}=\{T_1,T_2,\cdots,T_n\}$。

定义 3 设 $C_{(u_iu_j)|T_i}$ 为任意两个用户 u_i、u_j 同时参与某话题 T_i 的讨论时,u_i 对 u_j 的回复数。

在以上定义的基础之上,本文提出两个假设:

假设 1 u_i 对 u_j 回复数 $C_{(u_iu_j)|T_i}$ 表征在话题 T_i 下 u_i 对 u_j 的感兴趣程度。

假设 2 任意两个用户 u_i、u_j 同时参与多个话题讨论时,u_i 对 u_j 总回复数 $c_{(u_iu_j)}$ 代表 u_i 对 u_j 的感兴趣程度。合并公式为

$$c_{(u_iu_j)}=F(c_{(u_iu_j)|T_1},c_{(u_iu_j)|T_2},\cdots,c_{(u_iu_j)|T_n}) \tag{1}$$

式(1)所示的函数为 $F(x_1,x_2,\cdots,x_n)$ 非线性函数,随着 n 的增大,$F(x_1,x_2,\cdots,x_n)$ 的增长速度越快,其物理意义为在多个主题帖下,用户 u_i 都与用户 u_j 进行发回帖讨论,代表这两个用户有共同的兴趣方向,增强 u_i 为 u_j 的单向好友的可能性。

本文在处理讨论区信息时,选取 K 个帖子进行处理,相应得到用户集合 $U=\{u_1,u_2,\cdots,u_N\}$,分析此 N 个用户的发回帖关系,可以得到一个 $N\times N$ 维的用户关系矩阵 M_T。

其次,分析数据的另一来源——文集。文集属于作者个人空间,用户留言的回复关系在页面中无法直接体现。为便于分析和处理,本文不考虑用户在文集评论中的相互回复关系,而是采用以文集作者为中心的分析方法,通过提取用户在某个文集不同文章下的留言数目,并利用 F 函数进行合并,获得用户间的认知度。

定义 4 设 P_{ui} 为某用户 u_i 的文集,由 u_i 个人撰写的文章构成,则 $P_{u_i}=\{p_{u_{i1}},p_{u_{i2}},\cdots,p_{u_{in}}\}$,$p_{u_{ik}}$ 代表用户 u_i 在文集 P_{u_i} 中发表的第 k 篇文章。

定义 5 设 $c'_{u_j|p_{u_{ik}}}$ 为某用户 u_j 在用户 u_i 的文集 P_{u_i} 中某篇文章 $p_{u_{ik}}$ 的留言数。

接下来在已获取的用户集合 U 的范围内进行如下统计:

a)统计用户 u_j 对用户 u_i 文集 P_{u_i} 中各篇文章的评论数,代入式(1)的 F 函数可得 $c'_{u_iu_j}$,即 $c'_{u_iu_j}=F(c'_{u_j|p_{ui1}},c'_{u_j|p_{ui2}},\cdots,c'_{u_j|p_{uim}})$。

b)根据步骤 a)处理方法,获得的用户 u_i 对用户 u_j 文集 P_{u_j} 中各篇文章的评论数 $c'_{u_ju_i}$。

c)处理用户集合 U 中所有用户的个人文集,获得关系矩阵 M_P。

最后,本文结合 $c_{(u_iu_j)}$ 的定义以及 F 函数的物理意义,得到式(2)。

$$c_{(u_iu_j)}=F(c_{(u_iu_j)|T_1},c_{(u_iu_j)|T_2},\cdots,c_{(u_iu_j)|T_n})=(c_{(u_iu_j)|T_1}+c_{(u_iu_j)|T_2}+\cdots+c_{(u_iu_j)|T_n})\times\alpha^n \tag{2}$$

其中:α 为调制系数,$\alpha>1$。在讨论区中,n 表示用户 u_i 回复用户 u_j 总主题帖数目;在文集中,n 代表用户 u_i 对用户 u_j 文集中文章的评论数。

以讨论区为例,假定总共处理 M 篇主题帖,通过处理已经获得各个主题帖下用户之间的回复关系矩阵 $T[M][N][N]$,根据 F 函数求对应的关系矩阵 $c[i][j]$,该矩阵初始化元素均为 0,算法的伪代码如下:

```
for i←0 to N
    for j ← 0 to N
      alpha = 1;
      for k←0 to N
        c[i][j] + = c[i][j];
        if T[k][i][j] >0 then
          alpha = alpha * α;//α 为常量
          c[i][j] = c[i][j] * alpha;
```

在表达用户间认知度上,讨论区和文集中的回复关系影响程度有一定差异。因此,在两者合并得出最终用户间认知度时,作如下处理:

$$m_{(u_iu_j)}=\tau c_{(u_iu_j)}+(1-\tau)c_{(u_iu_j)}'\quad 0\leqslant\tau\leqslant 1$$

其中:τ 代表权重,$m_{(u_iu_j)}$ 代表合并后的最终结果。

从宏观上看,合并过程就是对两个关系矩阵和的合并,得到表征用户之间的单向认知度的矩阵为

$$M_d=\tau M_T+(1-\tau)M_P$$

通过矩阵 M_d 的变换,可以将用户之间的单向联系转换为双向关系:

$$M=M_dM_d'$$

M 矩阵的值表征用户之间的认知度,其值越大,表示用户互相之间的联系越强。

本文在整个数据处理过程中,先通过爬虫获取目标 BBS 的网页文本,然后用训练好的模板抽取网页中的用户发帖及回复信息。由于 BBS 信息更新速度快,爬虫要具有高覆盖率和快速更新的能力;同时,网页存在一定不规则性,信息的抽取必须做到准确、全面。这就对系统的设计及算法实现提出了很高的要求。

2.2　基于认知度的社团关系挖掘算法

全体用户的分布是由若干个局部密集的社团构成的,社团代表了一群具有相同兴趣或偏好的用户团体。社团内部的用户联系相对紧密,而社团间的用户联系则相对稀疏,甚至有些完全没有联系。如何快速有效地挖掘出这些用户中的社团,是本文研究用户好友社团关系的核心。

在社会网络分析领域中,分级聚类问题和社团挖掘算法联系非常密切,可分为凝聚算法和分裂算法两类。算法主要基于各个节点的相似性或连接强度,将网络划为若干子群。凝聚算法的思想是找到相似度最高的节点对,往一个节点为 n 而边数为 0 的原始空网络中加边,任意节点可以终止,此时网络的组成就可认为是若干社团;相反,分裂算法的思想是找相似度最差的节点对,移除两点之间的边,重复上述过程,同样可以在任意节点终止,此时的网络也可以看成若干社团的集合。两种算法的最终目的都是为了发现网络中的密集区域。

本文提出的基于认知度的社团关系挖掘算法,就是将网络中的节点看做用户群中的单个用户,通过认知度来定量表示用户间的连接强度,在全部用户关系联通图中找出连接强度不在阈值区间的两个用户,移除此用户之间的边,本质就是求关系图中连通分支的问题。阈值范围可由用户根据查询的需求自行设定。本文基于认知度的挖掘算法采用广度优先搜索(breadth-first-search)遍历关系图来实现,具体算法流程如图 1 所示。

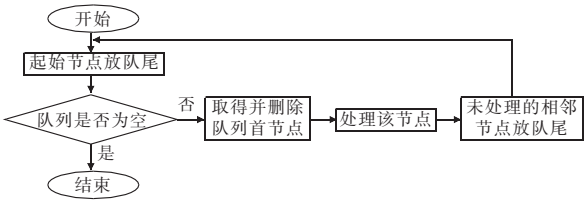

图 1　广度优先搜索流程

2.3　基于认知度社团关系挖掘的可视化方法

本文在社团关系挖掘算法的基础上获得仅包含密集区域的关系图后,为了让用户能直观地看出 BBS 用户好友的社团关系分布,展现 BBS 用户个体与个体、个体与社团和社团与社团的相互关系及影响,本文根据关系图矩阵在二维平面上描绘其点线关系图。具体步骤如下:

a)通过认知度算法计算任意两个用户的认知度 $m_{(u_iu_j)}$。

b)利用计算出的认知度算出相应用户的关系矩阵 M。

c)不考虑关系矩阵 M 对角线上的值,图中节点代表用户也就是关系矩阵中顶点,边的值代表用户间的认知度也就是关系矩阵中的边。选定阈值,作出关系图。

点线关系图会随着用户设定的阈值区间而变化,该图可直观展示相应的社团关系。

3　实验与分析

3.1　实验数据与设计

由于各大 BBS 的模式不一,实验分析只针对一个 BBS 用户群体,数据全部来自水木清华 BBS,采样 15 个版面。

实验系统设计如图 2 所示。该系统主要由信息采集、处理、分析三个部分组成。

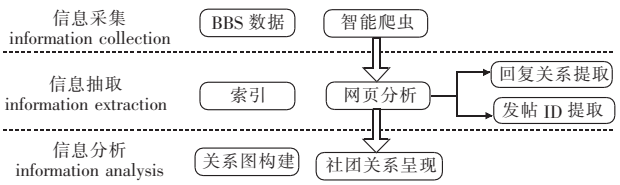

图 2　社团关系挖掘体系框架

a)信息采集(IC)。主要负责 BBS 用户数据的采集与存储,通过网络爬虫获取 BBS 数据。

b)信息处理(IP)。提取原始数据,网页去噪,提取网页内容,对采集的网页建立索引和存储并获取用户 ID 及其之间的回复关系。

c)信息分析(IA)。对预处理的数据进行相关算法的计算

和分析,获得用户关系矩阵 G,并提出用户空间的概念描述社区中的用户关系。

根据与用户交互的数量本研究将 BBS 用户分为活跃、较为活跃和普通用户三类[13]。通过得出的关系图可以直观地看出用户交互的情况,统计出与每个用户有交互关系的用户总个数,进而设置区间对所有用户进行分类。

3.2 实验结果与分析

在实验中,用 HTMLParser 进行解析,通过本文用户间认知度算法提取用户社团关系,用 Graphviz 得到一个局部的用户社团关系图如图 3 所示。

利用 Graphviz 生成图 3 的代码如下:

```
digraph 用户关系图{//图的类型和名字
size = "a,b";//图的大小
32--0[label = "83.4"];//"认知度"方向及数值
32--1[label = "7.6"];
…//省略其余相关用户认知度
119--178[label = "19.5"]}
```

如图 3 所示,节点数字代表用户 ID 编号,边的值代表两 ID 之间的认知度。图 3 直观地展现了用户之间友好社团分布,用户间认知度越高代表是好友的可能性越大,ID-32 与 ID-15 认知度为 39.2 大于其余 ID,所以 ID-32 与 ID-15 是好友的可能性最大;其次,可以直观地看出活跃用户和普通用户群体,从局部用户关系图中可以看出 ID-32 为活跃用户,ID-35、ID-9 为较为活跃用户,ID-39、ID-97 等为普通用户。综上所述,BBS 中用户存在一定的群体性特征,并通过几个关键用户联在一起。通过关系分析,能够直观地看出 BBS 中的用户社团关系,同时也能够发现用户社团之间的关联用户。

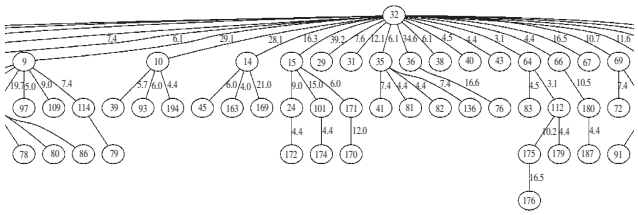

图 3 用户社团关系

本文将与用户交互的总个数划分为三个区间进行统计,交互个数小于 5 个时称之为普通用户,大于 5 小于 30 个时则称为较为活跃用户,大于 30 个的称为活跃用户。各类用户数占总用户数的比例如图 4 所示。

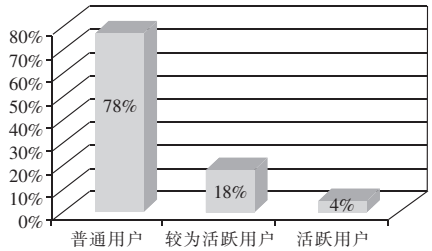

图 4 用户比例

在上述实验基础上,本研究还设计实现了相应系统,系统界面如图 5 所示。

该系统针对水木、未名等主要高校 BBS 论坛进行实时的分析处理,已处理 400 多万网页。

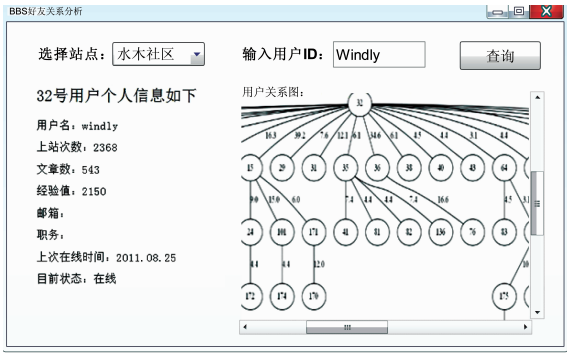

图 5 BBS 用户好友关系分析软件界面

4 结束语

本文提出一种基于用户间认知度好友的社团关系挖掘算法,并在此算法基础上研究一种好友的社团关系可视化方法,通过社团关系图所提供信息为 BBS 用户提供了好友社团发现等一系列好友服务。理论分析和实验结果证明本文提出的 BBS 好友的社团关系挖掘方法对用户友好社团发现有显著的效果。由于用来进行分析的数据主要是围绕用户之间的认知度,所以计算认知度的算法就显得尤为重要,其准确率和适用性有待进一步提高,不仅要从帖子数量反映用户之间的认知度,还可以加入帖子内容的语义分析对现有算法进行调整,使认知度更加准确,另外还有关系图中隐藏信息的挖掘,这些都是本研究未来研究的重点。

参考文献:

[1] MILGRAM S. The small-world problem[J]. Psychology Today, 1967,1(1):60-67.

[2] WATTS D J,STROGATZ S H. Collective dynamics of 'small world networks'[J]. Nature,1998,393(6684):440-442.

[3] JOHN H I I I,ARMSTRONG A G. 网络利益:通过虚拟社会扩大市场[M]. 王国瑞,译. 北京:新华出版社,1998.

[4] BAUMES J,GOLDBERG M,MAGDON-ISMAIL M,*et al*. Discovering hidden groups in communication networks[R]. [S. l.]:ISI,2004:378-389.

[5] WASSERMAN S,FUAST K. Social networks analysis:methods and applications[M]. Cambridge:Cambridge University Press,1994:27-35.

[6] FREEMAN L C. Centrality in social networks:conceptual clarification[J]. Social Networks,1979,1(3):215-240.

[7] 汪小帆,李翔,陈关荣. 复杂网络理论及其应[M]. 北京:清华大学出版社,2006:9-14.

[8] 荣波. 虚拟空间成员交互网络特性及潜在组织成员搜寻研究[D]. 南京:南京航空航天大学,2010.

[9] KERNIGHAN B W,LIN S. A efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal,1970,49(2):291-307.

[10] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal,1973,23(98):298-305.

[11] POTHEN A,SIMMON H,LIOU K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications,1990,11(3):430-452.

[12] CAPOCCI A,SERVEDIO V D P,CALDARELLI G,*et al*. Detecting communities in large networks[J]. Physica A:Statistical Mechanics and its Applications,2005,352(2-4):669-676.

[13] 彭小川,毛晓丹. BBS 群体特征的社会网络分析[J]. 青年研究,2004(4):39-44.