基于交互行为和连接分析的社交网络社团检测

李 鹏 李英乐 王 凯 何赞园 李 星 常振超
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 社交网络的迅猛发展极大地方便了人们的日常生活、工作和学习,但也带来了大量复杂的交互行为和连接模式。如何有效地综合分析网络中的交互信息和网络节点之间存在的连接信息,进而完成高效的社团检测,是在当前网络多维属性的复杂背景下进行网络分析所面临的关键难题。基于此,从有效融合两类不同的异质信息研究出发,提出了一种基于交互行为和连接分析的社交网络社团检测(CDUILS)方法。该方法基于两类信息能够从不同的角度反映网络同一个社团归属的假设,采用联合非负矩阵分解架构,以迭代更新的方式,同时利用两类信息进行社团结果的获取。在真实网络数据集上的实验表明,与已有方法相比,所提方法能够有效融合两类信息进行社团检测,取得了更好的社团划分质量。

关键词 交互信息,非负矩阵分解,社交网络,社团检测
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.07.035

Community Detection Based on User Interaction and Link Analysis in Social Networks

LI Peng LI Ying-le WANG Kai HE Zan-yuan LI Xing CHANG Zhen-chao
(National Digital Switching System Engineering & Technological Research Center, Zhengzhou 450002, China)

Abstract With the rapid development of social media network, the user is also more convenient to participate in social networking, which also brings a large number of complex interaction and connection mode. How to effectively analysis the interactive information and the connection information between network nodes to complete the efficient community detection is the key problem faced by current network analysis. Based on this, this paper put forward a kind of social network community detection method(CDUILS) based on the interaction behavior and link analysis. In this method, the interaction information between nodes is used as the cooperative learning of the community. The non negative matrix factorization is used to analyze the two types of information sources by the way of iterative update, and the community results can be obtained with two kinds of information retrieval. Experiments on real data sets show that the proposed method can effectively utilize the interaction behavior to guide the community division and have better quality of community division.

Keywords Interaction information, Non-negative matrix factorization, Social network, Community detection

1 引言

作为最基本的组织结构,社团存在于人类社会所构成的各种网络中,如社团可以是科学家合作网络中处于同一个研究领域的研究群组、引文网络中处于同一个课题的论文、社交网络中具备相似爱好的兴趣部落等[1]。借鉴经典的图论方法进行社团划分,网络中的基本单元(用户、个体)可用图中的节点进行表示,而网络中节点之间的连接关系可用图中的边进行描述。

在线社交网络(如 Facebook、Twitter、微博、微信,等)的蓬勃发展使得网络中的用户可以更便捷地发布和获取其感兴趣的内容信息,用户之间存在大量的交互行为(评论、转发、点赞等),这些交互信息从侧面反映了人们之间关系的紧密程度。对于具备更为亲密的关系的群体,其网络交互行为也更为紧密。当前社交网络普遍存在弱连接现象,且在数据采集过程中,数据的残缺导致仅从连接结构出发难以获取精确的社团结构。在线社会网络的社团结构不仅体现在节点之间的连边关系上,还更多地体现在交互行为上[2-3]。经典的社团检测方法仅从网络的连接关系出发进行分析,已经取得了较大的研究进展;针对这种具备更多交互行为的网络社团检测问

到稿日期:2016-05-30 返修日期:2016-08-27 本文受国家自然科学基金创新群体项目(61521003),国家重点基础研究发展计划资助项目(2012CB315901,2012CB315905),国家科技支撑计划(2014BAH30B01)资助。

李 鹏(1978-),男,硕士,工程师,主要研究方向为通信与信息系统;李英乐(1985-),男,硕士,助理研究员,主要研究方向为通信与信息系统;王 凯(1980-),男,博士生,副研究员,主要研究方向为通信与信息系统;何赞园(1975-),男,硕士,高级工程师,主要研究方向为通信与信息系统;李 星(1987-),男,博士生,助理研究员,主要研究方向为通信与信息系统;常振超(1987-),男,博士生,主要研究方向为通信与信息系统,E-mail:changzhenchao1987@126.com(通信作者)。

题的研究，目前尚处于起步阶段[2]。

本文从节点的交互信息能够为社团的精确划分提供有效指导的角度出发，将交互行为分析和网络的连接关系分析进行融合，对在线社交网络中的社团检测问题展开研究。基于上述分析，本文提出了一种基于交互行为和连接分析的社团检测方法(Community Detection based on User Interaction and Link Similarity，CDUILS)。该方法将网络的连接信息和节点之间的交互信息进行融合分析，分别构建了连接相似度矩阵和交互相似度矩阵，通过联合矩阵分解架构，完成对社团归属的逼近。本文所提方法能够有效利用不同信息源之间的相互促进作用，融合两类不同信息源进行联合分析，为当前具备复杂交互行为的网络社团检测提供了新的解决思路。

2 相关工作

当前针对社交网络中的社团检测，研究者们已经取得了大量的研究成果。按照所利用的网络信息的不同，可将这些方法分为两类：基于网络拓扑分析的社团检测和融合交互分析的社团检测方法。

基于网络拓扑分析的社团检测从网络的连接关系出发，被定义为图聚类问题，通常情况下很难取得最优解。基于网络拓扑的方法从社团的基本连接结构分析出发，通过连接相似度定义，如具备更多的邻居个数、具备节点群组的内部连接数目大于外部连接数目准则等，完成社团结构的判决，取得了大量的研究成果。按照所处理社团重叠性的不同，主要可分为基于重叠社团的研究和非重叠方法的研究，其中基于重叠社团的研究主要有边聚类的方法[4]、k 派系的 CFinder 方法[5]等，而针对非重叠方法的研究主要是从模块度角度出发，如 GN 算法[6]及其改进的衍生算法。

基于交互信息的社团检测方法从当前新型社交网络的基本属性出发，网络中不仅存在连接关系，而且错综多样的交互信息增加了对网络关系分析的复杂性，交互信息也能有效反映网络中节点之间关系的强度。针对此类问题，需要提出综合分析交互行为和连接关系的社团检测方法。Dev H 等人[7]同时考虑了节点之间以及节点的共同邻居之间的交互，借助于两类交互强度的衡量来决定节点对于群组的归属值，然后对节点对进行概率社团归属描述，进而完成层次的聚类过程。从融合两者信息的角度，许飞等人[2]提出了一种基于用户交互行为和相似度的社交网络社团发现方法，将用户之间的多维关联概括为交互行为和相似度，采用增加相似性惩罚因子的方法，以模块度为最终的优化目标。

通过综合分析现有方法可知，现有方法虽然从不同的角度提出了相应的解决方案，取得了较好的效果，但仍存在以下局限性：1)仅从拓扑的方法出发，难以有效反映当前社交网络中的用户之间的交互行为对社团结构的影响；2)现有综合分析交互和连接的方法大多首先研究两类不同信息对社团划分结果的影响，再进行线性叠加分析，无法有效反映二者在社团检测过程中的相互促进作用，无法从不同的角度同时对社团划分结果进行优化。

基于上述分析，本文从综合分析社交网络中的交互行为和拓扑结构两类信息出发，构建基于网络邻接矩阵和节点交互相似度矩阵的协同学习目标，采用联合矩阵分析的方法，最终获取原始社团结构的精确划分。

3 基于交互行为和连接分析的社团检测

基于交互行为和连接分析的社团检测针对网络节点的两类不同属性即连接关系和交互行为展开研究。当前社交网络通常具备弱连接现象，因此仅从连接进行分析的研究方法无法获取精确的社团结构。节点之间具备的海量复杂的交互模式能够有效解决这种信息残缺所带来的问题，具备大量交互(评论、转发)行为同时连接紧密的节点归属于同一个社团的可能性更大，因此融合二者信息进行联合分析的方法对此类问题的处理更有效。

3.1 基本定义及分析

网络图通常采用 $G(V,E)$ 来进行数学化表达，其中 V 为网络的节点集合，$V=\{v_1,v_2,\cdots,v_n\}$ 表示网络中的 n 个节点，E 为网络的连边集合，$E=\{e_1,e_2,\cdots,e_m\}$ 表示网络中的 m 条边，$C=\{c_1,c_2,\cdots,c_K\}$ 用于表示网络中的 K 个社团结构。

定义 1 邻接矩阵 $X\in\mathbb{R}^{n\times n}$，$n$ 为节点总数，矩阵元素的值为其中两个节点之间的连接关系，即当节点 v_i 和节点 v_j 之间有连接时，$X(i,j)=1$，反之 $X(i,j)=0$。

定义 2 社团归属矩阵 $H\in\mathbb{R}^{K\times n}$，$H$ 的第 j^{th} 列表示节点 v_j 在 K 个社团上的归属程度，刻画了原始信息的结构与特征。

定义 3 交互行为相似度矩阵 $S\in\mathbb{R}^{n\times n}$，$n$ 为节点总数，其中元素值的大小为两个节点之间的交互行为的相似度。

基于上述定义，本文联合分析的社团检测问题可以理解为：在获取了用于描述网络拓扑的邻接矩阵 X 之后，通过节点交互行为构造相似度矩阵 S，对这二者信息进行联合矩阵分解，以同时获取社团的归属矩阵 H。

非负矩阵分解(Nonnegative Matrix Factorization，NMF)最早被用于对图像进行压缩[8-9]，通过对高维原始数据进行分解获取低维的特征表达。非负矩阵分解要求矩阵中的值为非负的，从社团检测的角度而言，当对网络结构进行处理时，由于网络节点之间的连接是非负的，因此用于网络描述的矩阵也是非负的。真实网络的连接通常具备稀疏特征，比较适合采用非负矩阵分解的方式进行社团挖掘。原始矩阵经 NMF 分解后可以得到基向量矩阵 W 和归属矩阵 H，分别用于描述降维后的社区特征和节点在某个具体社团内的隶属程度。

经典的非负矩阵架构的社团检测过程的定义如下，假设某个网络 $G(V,E)$ 的邻接矩阵为 $X\in\mathbb{R}^{n\times n}$，则基于 NMF 的社团检测问题为：通过寻找最大近似原始网络数据 X 的两个低秩因子矩阵 W 和 H 来完成社区检测的过程。当采用欧几里德距离时，优化目标函数 $O^l(E)$ 为：

$$\min O^l(E)=\min_{W,H}\|X-WH\|_F^2 \quad \text{s.t. } W\geqslant 0, H\geqslant 0 \tag{1}$$

其中，$\|\cdot\|_F$ 为距离度量函数，称为 Frobenius 范数，用于描述分解后的适量空间与原始空间的逼近程度。$W\in\mathbb{R}^{n\times K}$ 和

$H\in\mathbb{R}^{K\times n}$分别是分解之后得到的关于模式节点的基矩阵和归属矩阵。n 表示网络中的节点个数 K 表示节点在空间降维后的维数即聚类数目,在社团检测中为网络 G 中的社团划分个数。

基于交互行为和连接分析的社团检测以及联合学习逼近同一个目标函数的思想,采用联合矩阵分解的架构[10],主要目标是通过定义联合非负矩阵的分解目标函数将不同类的信息进行综合分析,以获取同一个结果。其假设:根据不同的信息进行学习,能够获取同一个社团归属矩阵 H。本文所提算法综合分析了通过拓扑结构获取的原始网络邻接矩阵和通过交互行为获取的相似性矩阵,有效增强了社交网络中网络社团分析的可行性。本文所提算法的基本处理架构如图 1 所示。

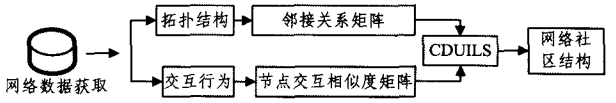

图 1　基于交互行为和连接的社团检测算法

3.2　交互行为紧密度矩阵的获取

网络的连接关系可以直接获取,而针对具备大量交互信息的社交网络进行研究时,需要准确地描述这种交互行为的强度。本文借鉴文献[7]的交互紧密度,构造交互行为所产生的节点之间的相似度矩阵。

交互紧密度矩阵 S 中的任意元素 S_{ij} 的定义为:

$$S_{ij}=R_{ij}*F_{ij} \tag{2}$$

其中:

$$R_{ij}=\begin{cases}\dfrac{\min(W_{ij},W_{ji})+1}{\max(W_{ij},W_{ji})+1}, & W_{ij}>0 \text{ or } W_{ji}>0\\ 0, & \text{else}\end{cases} \tag{3}$$

$$F_{ij}=\begin{cases}\sqrt{(w_{ij}+1)*(w_{ji}+1)}, & W_{ij}>0 \text{ or } W_{ji}>0\\ 0, & \text{else}\end{cases} \tag{4}$$

其中,W_{ij} 和 W_{ji} 分别表示社交网络中两个用户 i 和 j 之间发生交互的次数(没有方向性限制),R_{ij} 表示节点之间的相互程度,F_{ij} 表示节点之间交互的频度。借助于式(2)量化用户间的紧密相似度,进而构造基于用户行为的紧密相似度矩阵 S。

3.3　基于交互行为和连接分析的社团检测

根据 3.1 节的描述,当从网络拓扑结构进行考虑时,社团检测等效于式(1)中的目标逼近问题,即获取用于描述原始网络的社团归属矩阵 H。通过 3.2 节的定义,引入交互密度矩阵 S 时,采用非负矩阵分解架构获取同一个社团归属矩阵 H 以及另一个用于描述特征的基矩阵 $W_2\in\mathbb{R}^{n\times K}$,将其等效为对交互行为目标的 $O^a(N)$ 最小优化目标逼近。在本文的研究问题中,需要对连接相似度目标 $O^l(E)$ 和交互行为目标 $O^a(N)$ 进行综合分析,因此本文的联合优化问题定义如下:

$$\begin{aligned}&\min O^l(E)+O^a(N)\\&=\min_{W,H}\|X-W_1H\|_F^2+\|S-W_2H\|_F^2+\\&\lambda\sum_{j=1}^{n}\|H(:,j)\|_1^2+\zeta(\|W_1\|_F^2+\|W_2\|_F^2)\\&\text{s.t. } 0\leqslant W_1H\leqslant 1_{n\times n}\\&\quad H\geqslant 0\end{aligned} \tag{5}$$

其中,参数 λ 是对逼近程度和非负归属矩阵 H 进行均衡化的惩罚因子;ζ 通过控制基矩阵 W_1 和 W_2 中元素值的大小来合理化这两个矩阵中元素值对优化结果的影响。本文采用 L_1 正则化的方式处理 H 中的列向量,保证了节点隶属于少量的社团。

在式(5)中对该联合矩阵中的各项进行求解时,需要保证算法的全局优化可求解性,即将函数的非凸性转化为凸函数进行求解。通常的做法是在求解某个变量时确保其他变量是保持不变。因此,分别选择不同的变量单独进行求解,以获取其局部的最小值,具求解过程如下。

当选择 W_1 和 W_2 为首先固定的两个变量时,求解 H 的过程如定理 1 所示。

定理 1　式(5)中将基矩阵 W_1 和 W_2 进行固定时,对 H 的逼近过程等效于式(6)的优化目标过程。

$$\begin{aligned}&\min_{W,H}\|B-CH\|_F^2\\&\text{s.t. } DH\leqslant E\end{aligned} \tag{6}$$

其中,B,C,D,E 的定义如下:

$$\begin{aligned}B&=(X^{\mathrm{T}},S^{\mathrm{T}},0_{n\times 1})^{\mathrm{T}}\\C&=(W_1^{\mathrm{T}},W_2^{\mathrm{T}},\sqrt{\lambda}1_{K\times 1})^{\mathrm{T}}\\D&=(-1_K,W_1^{\mathrm{T}},-W_1^{\mathrm{T}})^{\mathrm{T}}\\E&=(-0_{n\times K},1_{n\times n}^{\mathrm{T}},-0_{n\times n})^{\mathrm{T}}\end{aligned} \tag{7}$$

证明:当 W_1 和 W_2 确定时,式(5)中最后的正则化项 $\zeta(\|W_1\|_F^2+\|W_2\|_F^2)$ 为常数。由于 H 中的元素具备非负限制,$\sum_{j=1}^{n}\|H(:,j)\|_1^2=\|1_{1\times K}H\|_2^2$,式(5)中的优化过程可以等效为:

$$\begin{aligned}&\|X-W_1H\|_F^2+\|S-W_2H\|_F^2+\lambda\|1_{1\times K}H\|_2^2\\&=\|(X^{\mathrm{T}},S^{\mathrm{T}},0_{n\times 1})^{\mathrm{T}}-(W_1^{\mathrm{T}},W_2^{\mathrm{T}},\sqrt{\lambda}1_{K\times 1})^{\mathrm{T}}H\|_F^2\\&=\|B-CH\|_F^2\end{aligned} \tag{8}$$

同时,式(5)中的约束条件也可以表示为:

$$(-1_K,W_1^{\mathrm{T}},-W_1^{\mathrm{T}})H\leqslant(-0_{n\times K},1_{n\times n}^{\mathrm{T}},-0_{n\times n})^{\mathrm{T}} \tag{9}$$

定理 1 的证明过程到此结束。

与之相似,当把基矩阵 W_2 和归属矩阵 H 进行固定以求解 W_1 时,可以证明如下定理。

定理 2　当把基矩阵 W_2 和归属矩阵 H 固定时,式(5)中对 W_1 的求解过程为式(10)的优化目标。

$$\begin{aligned}&\min_{W,H}\|B-CW_1^{\mathrm{T}}\|_F^2\\&\text{s.t. } DW_1^{\mathrm{T}}\leqslant E\end{aligned} \tag{10}$$

其中,B,C,D,E 的定义为:

$$\begin{aligned}B&=(X,0_{K\times n})^{\mathrm{T}}\\C&=(H,\sqrt{\zeta}1_K)^{\mathrm{T}}\\D&=(H,-H)^{\mathrm{T}}\\E&=(1_{n\times n},-0_{n\times n})^{\mathrm{T}}\end{aligned} \tag{11}$$

证明:当基矩阵 W_2 和 H 归属矩阵固定时,式(5)中 $\lambda\sum_{j=1}^{n}\|H(:,j)\|_1^2$ 和 $\zeta\|W_2\|_F^2$ 这两项为常数,而 $\|S-W_2H\|_F^2$ 中由于 S 项不包含 W_1 项的分解,因此求解 W_1 的目标函数转化为式(12)。

$$\| X-W_1H \|_F^2+\zeta \| W_1 \|_F^2$$
$$=\| (X,0_{K\times n})^{\mathrm{T}}-(H,\sqrt{\zeta}1_K)^{\mathrm{T}}W_1^{\mathrm{T}} \|_F^2$$
$$=\| B-CW_1^{\mathrm{T}} \|_F^2 \tag{12}$$

与之前定理 1 的证明过程相似,约束条件也相应地改变。

定理 2 证明完毕。

由定理 2 可知,当固定归属矩阵 H 时,W_1 的计算过程与 W_2 和 S 无关。

同样,当 W_1 和 H 确定时,由于式(5)中的约束条件与 W_2 无关,除去常数项,式(5)可转化为如下无约束的优化问题。

$$W_2=SH^{\mathrm{T}}(HH^{\mathrm{T}}+\zeta 1_K)^{-1} \tag{13}$$

由定理 1 和定理 2 可知,对 W_1 和 H 的优化过程等效于如下带约束的目标函数逼近过程。

$$\min_X \| B-CX \|_F^2$$
$$\text{s. t. } DX\leqslant E \tag{14}$$

由式(14)可知,本文所提联合矩阵分解问题等效于最小均方误差问题,也就是凸二次规划的有效集方法[11]。算法的输入部分为设定好的 4 个具体参数(B,C,D,E),算法 1 对归属矩阵 H 的更新准则进行了描述,求解过程等效于经典的非负矩阵分解过程。

算法 1 归属矩阵 H 的更新过程

输入:邻接矩阵 X,紧密度矩阵 S,恒定参数值 W_1,W_2 和 λ

输出:归属矩阵 H

1. 由式(7)获取并计算相应的参数矩阵 B,C,D,E;
2. for i=1→n do
3. 计算 H(:,i)←根据(X(:,i),B,C,D(:,i))
4. end for

与之相似,本文更新 W_1 的方法在算法 2 中进行描述。根据定理 2 的分析过程,算法 2 中 W_2 和 S 为固定参数,因此其描述如下所示。

算法 2 基矩阵 W_1 的更新过程

输入:邻接矩阵 X,恒定参数值 H 和 ζ

输出:基矩阵 W_1

1. 由式(11)获取并计算相应的参数 B,C,D,E;
2. for i=1→n do
3. 计算 W_1^{T}(:,i)←根据(X(:,i),B,C,D(:,i))
4. end for

基于 H 和 W_1 的更新算法 1 和算法 2,可以对式(5)的目标函数进行求解,联合矩阵求解的具体描述如算法 3 所示。

算法 3 联合矩阵分解算法

输入:连接矩阵 X,紧密度矩阵 S,恒定参数值 λ 和 ζ

输出:归属矩阵 H

1. 初始化 H,W_1 和 W_2
2. while Not convergent do
3. 获取并更新:W_2 by 式(13)
4. 获取并更新:H by 算法 1
5. 获取并更新:W_1 by 算法 2
6. end while

在算法 3 中,需要首先对原始矩阵进行初始化,然后固定其中两个变量,反复迭代运算 W_2,H 和 W_1,直到目标函数达到相应的收敛指标。

3.4 参数估计和算法复杂度分析

本文算法需要确定的参数主要有:1)在式(5)中,需要确定混合参数 λ 和 ζ 的值。这两个参数通常根据具体实验中数据集的分布决定,通过交叉验证方法来设定,本文中 λ 和 ζ 的值采用文献[2]所提供的方式进行设定。2) 在联合矩阵分解迭代运算中,需要已知 H 矩阵的行数目,即网络社团的数目 K。在已知社团数目的社团检测情况下,K 值已给出;在未知社团数目情况下,需要进行预估。已有多种预估计社团数目的方法,本文采用文献[10]所采用的谱分析方法,根据裴龙聚类(Perron clusters)的本征值获取聚类数目[12]。

本文算法的复杂度包括 3 个部分:节点交互紧密度矩阵的获取、裴龙聚类特征值的求解过程和联合矩阵迭代分解所消耗的时间。当给定网络中的节点数目 n 时,由于紧密度矩阵的构造过程中需要对每两个节点的紧密程度进行计算,因此其算法复杂度与 $O(n^2)$ 呈线性关系。而在本征值求解过程中,按照裴龙聚类算法流程的最高运算复杂度部分进行描述,其运算复杂度为 $O(n^3)$。在最后的联合矩阵分解中,算法的复杂度不仅包含了矩阵乘运算,还与矩阵分解的迭代次数 l[13]有关,其算法复杂度为 $O(lmK^2)$,其中 m 为边的概述。对综合上面的分析过程可知,本文所设计的算法的复杂度为 $O(n^2)+O(n^3)+O(lmK^2)$。

4 实验

为了对算法的有效性展开分析,采用较为常见的真实社交网络数据集对所提方法进行了仿真验证。

4.1 实验数据集

本文针对已有文献中广泛应用的网络仿真数据库进行整理收集,对其详细介绍如下。

(1)Twitter 数据库[7]。参考文献[7],本文第一个数据集来源于真实的社交网络 twitter,节点为 twitter 用户,边为用户之间的关注与被关注关系。Twitter 为典型的社交网络,以 Twitter 用户为节点,以用户之间的互动操作为边,以朋友圈作为全局社团依据。对数据进行提炼之后,采集了以 10 个核心用户所构造的社团,其包含了 2150 个节点和 5429 条边。

(2)Amazon 购物网络[15]。该网络通过从 Amazon 网站爬虫得到,节点为网站上所销售的产品,如果两个商品同时被顾客购买,则认为这两个产品之间存在一条边。产品已有的分类准则为产品全局社团的依据。移除少于 3 个节点的无效社团结构,该数据集含 5000 个社团,共 334863 个节点和 925872 条边。

(3)Flickr 数据库[16]。Flickr 是一个图片共享网络,节点代表该网络中的用户,边代表不同用户之间的朋友圈关系。如果两个用户共享或者共同评论了同一个图片,那么这两个用户之间具备交互行为。该数据集使用 Flickr 的关键词群组搜索功能获取数据集,将最常见的标签作为查询条件。本文中将数据库的社团定义为共享相同图片的用户所标记的社交圈,由于共享同一个图片的用户交叉较多,社交圈出现了较大的重叠情况,该数据库包含了 100624 个社团,节点的个数为 16710,边的数目为 716063。

4.2　对比算法

为了合理地分析本文方法的有效性,本文实验的对比算法选择以下 3 种:1)只利用节点的拓扑结构相似度进行聚类的方法:Kmeans[16];2)只利用非负矩阵分解方法进行网络拓扑处理的方法:NMF[11];3)融合交互行为和连接的算法:CODICIL[14]和 Similarity-Louvain[12]。采用这 3 类方法进行对比的原因是:Kmeans 是经典聚类算法,能够有效地对具备相似度的节点进行聚类。NMF 方法在社团检测中广泛应用,也是本文主要采用的方法的基础,是一种无监督的聚类算法。而 CODICIL 是一种融合分析两类交互属性和拓扑属性的方法,本文中将其节点的属性值定义为节点对之间的交互行为的计算。Similarity-Louvain 则是最近两年基于交互行为和连接情况对社交网络进行聚类的方法。

4.3　评价指标

本文中的实验数据集是在全局背景已知的情况下的,即数据集的真实社团结构是可以获取的。因此,本文采用针对此类问题最为常见的测评指标即成对 F 测度和平均聚类纯度 ACP 来对算法的有效性进行验证和对比,其定义如下所示。

(1)成对 F-measure PWF(Pairwise F-measure)。成对 F 测度是准确率和召回率的折中表示。如式(15)、式(16)所示,G 代表真实情况下节点对至少处于一个类别中的情况,H 代表本文社团检测算法检测出的节点对至少处于同一个类别中的情况。

$$pr=\frac{|H\cap G|}{H} \tag{15}$$

$$rc=\frac{|H\cap G|}{G} \tag{16}$$

基于式(15)、式(16),成对 F-measure 为:

$$PWF=\frac{2\times pr\times rc}{pr+rc} \tag{17}$$

根据式(17),PWF 值越大,则所得到的社团聚类的结果质量越好。

(2)平均聚类纯度 ACP(Average Cluster Purity)。平均聚类纯度从社团检测的平均准确程度出发,其假设算法最终获取到了 K 个社团结果集,即 $C=\{C_1,\cdots,C_K\}$,针对得到的第 i 个社团 C_i,n_i 为某个社团中节点的个数,相应的网络节点数目为 $\{v_{1,i},\cdots,v_{n_i,i}\}$。设 $M_{l,i}$ 表明了节点 $v_{l,i}$ 对于社团的真实归属情况,则平均聚类系数的定义为:

$$ACP=\frac{1}{k}\sum_{i=1}^{k}\sum_{l=1}^{n_i}\frac{\delta(dom_i\in M_{l,i})}{n_i} \tag{18}$$

其中,若社团检测结果中的节点归属于真实归属社团,则 $\delta(\cdot)$ 取值为 1,反之其取值为 0。

同样,聚类纯度的值越大,则表明社团划分结果越接近真实网络中的社团归属情况。

4.4　实验结果

本文实验中需要首先确定混合参数 λ 和 ζ 的值。由于这两个值没有特定的推导方式,因此本文实验中借鉴文献[2]的数据验证结果,将这两个值设定为较小的正值,这里设定为 0.05。本文算法设计中,需要对非负矩阵分解的迭代停止条件进行设计,通常的设定方法为:在连续两次运算中,目标函数的差值小于 1e−6。同时,为了保证算法的准确性和合理性,实验均在相应的网络数据集上重复进行 10 次,并取各个度量指标的均值作为最终的结果。

为验证本文算法的有效性,在真实网络数据集上将本文提出的 CDUILS 算法与其他对比算法进行了仿真验证,实验结果分别如表 1 和图 2 所示。

表 1　5 种算法在 3 个网络数据集上的成对 F 测度

算法	Twitter	Flickr	Amazon
Kmeans	0.4186	0.3202	0.4673
NMF	0.3085	0.2620	0.4021
CODICIL	0.5202	0.4237	0.5069
Similarity-Louvain	0.5546	0.4309	0.5573
CDUILS	0.5788	0.4939	0.5665

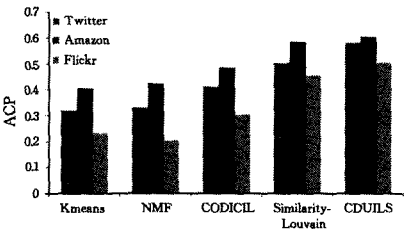

图 2　5 种算法在 3 个网络数据集上的平均聚类纯度

对 5 种算法在真实网络数据集上的实验结果进行分析可知:K-means 和 NMF 两种方法在两类真实网络数据集上划分社团的准确性均较低,原因在于这两类方法仅从单一信息处理的角度展开分析研究,无法有效融合网络的拓扑信息和网络的交互行为信息,但在真实的社交网络中节点的连接较为稀疏,仅从单一的角度出发难以获取较好的识别效果;而融合两类信息的 3 种方法均取得了更好的效果,这也验证了采用融合分析方法的有效性。针对 CODICIL,Similarity-Louvain,CDUILS 3 种融合分析算法,本文所提出的方法具备更好的社团检测效果,在 PWF 和 ACP 两项指标上都具有最优的结果。究其原因在于本文算法从不同的信息角度同时进行学习。从能够逼近同一个社团结构的角度分析,本文方法在同一个运算框架下利用不同的信息进行协同学习,有效利用了两种信息进行联合矩阵迭代分解,最终逼近同一个社团划分结果;而之前的算法仅将二者信息进行线性叠加,或者利用代价参数进行优化。因此,本文所提的融合模型更有效。综合以上分析可知,从单一信息角度出发,社团检测效果较差;而本文从融合角度进行研究,相比于已有的方法,取得了更好的社团检测结果。

结束语　本文从社交网络中丰富的交互行为对社团检测的影响出发,以联合矩阵分解为算法的分析基本架构,通过两类信息能够逼近同一个社团划分结果这个假设,对此类问题展开研究,提出了一种融合交互行为和连接分析的社团检测方法,并在真实网络数据集上对算法的有效性进行了仿真验证。随着社交网络日新月异的发展,越来越多的网络属性将会增加,这给当前社团检测提出了新的挑战,如何有效利用这些"海量、异质、多模"的多种属性数据,从融合分析的角度增强社团检测的效果和应对性,将是本文下一步的研究重点。

参 考 文 献

[1]　FORTUNATO S. Community detection in graphs [J]. Physics

Reports,2009,486(3-5):75-174.

[2] XU W,LIN B G,LIN S J,et al. Research on Community Detection Method for Social Networks Based on User Interaction and Similarity[J]. Netinfo Security,2015(7):77-83. (in Chinese)
许为,林柏钢,林思娟,等. 一种基于用户交互行为和相似度的社交网络社区现方法研究[J]. 信息网络安全,2015(7):77-83.

[3] TANG J,WANG X,LIU H. Integrating social media data for community detection[C]// Proceedings of the 2011 International Conference on Modeling and Mining Ubiquitous Social Media. Springer-Verlag,2011:1-20.

[4] TANG L,LIU H. Scalable Learning of Collective Behavior Based on Sparse Social Dimensions[C]// Proceeding of Acm Conference on Information & Knowledge Management (Cikm 09). 2009:1107-1116.

[5] GERGELY P,IMRE D,ILLÉS F,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature,2005,435(7043):814-818.

[6] GIRVAN M,NEWMAN M E. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2001,99(12):7821-7826.

[7] DEV H. A user interaction based community detection algorithm for online social networks[C]// ACM Sigmod International Conference on Management of Data. ACM,2014:1607-1608.

[8] LEE D D,SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature,1999,401(6755):788-791.

[9] WANG H,NIE F,HUANG H,et al. Nonnegative Matrix Tri-factorization Based High-Order Co-clustering and Its Fast Implementation[C]// 2011 11th IEEE International Conference on Data Mining. IEEE Computer Society,2011:774-783.

[10] CHANG Z C,CHEN H C,LIU Y,et al. Community detection based on joint matrix factorization in networks with node attributes[J]. Acta Physica Sinica,2015,64(21):0218901. (in Chinese)
常振超,陈鸿昶,刘阳,等. 基于联合矩阵分解的节点多属性网络社团检测[J]. 物理学报,2015,64(21):0218901.

[11] MAES C M. A regularization active-set method for sparse convex quadratic programming[M]. Stanford University,2010.

[12] WEBER M,RUNGSARITYOTIN W,SCHLIEP A. Perron Cluster Analysis and Its Connection to Graph Partitioning for Noisy Data:IB-Report 04-39[R]. 2004.

[13] JIN D,GABRYS B,DANG J. Combined node and link partitions method for finding overlapping communities in complex networks[J]. Scientific Reports,2015(5):08600.

[14] RUAN Y,FUHRY D,PARTHASARATHY S. Efficient Community Detection in Large Networks using Content and Links [C]// Proceedings of the 22nd international conference on World Wide Web. 2012:1089-1098.

[15] YANG J,LESKOVEC J. Defining and Evaluating Network Communities Based on Ground-Truth[J]. Knowledge and Information Systems,2012,42(1):745-754.

[16] KANUNGO T,MOUNT D M,NETANYAHU N S,et al. An Efficient k-Means Clustering Algorithm: Analysis and Implementation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2002,24(7):881-892.

(上接第 174 页)

[12] HULTEN G,SPENCER L,DOMINGOS P. Mining time-changing data streams[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2001:97-106.

[13] IKONOMOVSKA E,GAMA J,DŽEROSKI S. Learning model trees from evolving data streams[J]. Data Mining and Knowledge Discovery,2011,23(1):128-168.

[14] KEOGH E J,PAZZANI M J. Learning augmented Bayesian classifiers:A comparison of distribution-based and classification-based approaches[C]// Proceedings of the 7th Int'l Workshop on Artificial Intelligence and Statistics. San Francisco:Morgan Kaufmann Publishers,1999:225-230.

[15] LEWIS D D. Naive(Bayes) at forty:The independence assumption in information retrieval[J]. Machine Learning:ECML-98,1998,1398:4-15.

[16] MCCALLUM A,NIGAM K. A comparison of event models for Naive Bayes text classification[C]// Proceedings of AAAI-98 Workshop on 'Learning for Text Categorization',1998.

[17] LI H F,SHAN M K,LEE S Y. DSM-FI:An efficient algorithm for mining frequent itemsets in data streams[J]. Knowledge and Information Systems,2008,17(1):79-97.

[18] MERETAKIS D,WIJTHRICH B. Extending Naive Bayes classifiers using long itemsets[C]// Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. New York:ACM,2014:165-174.

[19] PFAHRINGER B,HOLMES G,KIRKBY R. New options for hoeffding trees[C]// Procee-dings of AI 2007:Advances in Artificial Intelligence. Heidelberg,Berlin:Springer-Verlag,2007:90-99.

[20] STREET W N,KIM Y S. A streaming ensemble algorithm (SEA) for large-scale classification[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,New York,2001:377-382.

[21] SCHLIMMER J C,GRANGER R H. Incremental learning from noisy data[J]. Machine Learning,1986,1(3):317-354.

[22] ZHANG P,GAO B J,ZHU X Q,et al. Enabling fast lazy learning for data streams[C]// Proceedings of 2011 IEEE 11th International Conference on Data Mining. New Jersey, USA:IEEE,2011:933-941.

[23] ZLIOBAITE I,BIFET A,PFAHRINGER B,et al. Active learning with evolving streaming data[J]. Lecture Note in Computer Science,2011,6913:597-612.