基于 WB-MMSB 模型的微博网络社区发现

徐建民[1] 武晓波[1] 吴树芳[2] 粟武林[1]
(河北大学数学与计算机学院 保定 071002)[1] (河北大学管理学院 保定 071002)[2]

摘 要 提出了一个用于微博网络社区发现的模型 WB-MMSB,该模型考虑了微博网络中节点存在的单向关系,节点的社区隶属度从链入主题隶属度和链出主题隶属度两个方面表示。用指数族分布和平均场变分推理方法推导了模型中各变量的表示,并用 SVI 算法计算模型涉及的参数。实验在新浪微博数据集上进行,采用归一化互信息和困惑度进行评估,结果表明,WB-MMSB 模型的社区发现能力优于 aMMSB 模型,并且其收敛速度快于 aMMSB 模型。

关键词 微博网络,社区发现,混合隶属度随机块模型,重叠社区
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.014

Community Detection for Micro-blog Network Based on WB-MMSB Model

XU Jian-min[1] WU Xiao-bo[1] WU Shu-fang[2] SU Wu-lin[1]
(College of Mathematics and Computer, Hebei University, Baoding 071002, China)[1]
(College of Management, Hebei University, Baoding 071002, China)[2]

Abstract Considering the nodes of Mico-blog network have single direction relations, a new model WB-MMSB was put forward for community detection, which uses directed edges to embody the direction relations of nodes, and two aspects link-in and link-out are used to quantify the community membership of nodes. Exponential family distribution and mean-field variational inference method were used to inference the representations of variables in this model, and SVI algorithm was used to compute relating parameters. Experiments adopted Sina-Weibo dataset and NMI to testify the performance of WB-MMSB. The results indicate that the community detection ability of WB-MMSB model is better than aMMSB model, and the convergence rate of WB-MMSB model is faster than aMMSB model.

Keywords Micro-blog network, Community detection, Mixed membership stochastic block model, Overlapping communities

1 引言

现实生活中的复杂网络系统,如生物网络、文献引用网络、电力网络和万维网等,它们普遍存在着一种形如社区的拓扑结构,这种结构通常表现为社区内部节点连接紧密而社区间节点连接松散。社区发现[1]是指在某一特定的集合中,根据个体间的相互关系,将集合划分为若干个社区的过程。自 2002 年 Girvan 和 Newman[2]提出社区发现问题以来,社区发现已被广泛应用于恐怖组织识别、蛋白质相互作用网络分析、Web 社区挖掘、搜索引擎等诸多领域[3-5]。识别网络中的社区有助于深入认识网络的结构和功能。

根据是否允许一个节点同时属于多个社区,社区发现方法主要分为两大类:传统的社区发现和重叠社区发现[4,5]。传统的社区发现假设一个节点只属于一个社区,主要根据社区内节点链接稠密和社区间链接稀疏的特征,并借助于网络的平均路长、聚类系数、度分布等刻画网络结构的属性来划分社区,主要有谱方法、模块度方法、层次聚类算法、标签传递算法(label propagation)、边预测方法(edge prediction)等[3,20,22,23]。传统社区发现方法的假设与现实生活中的情况相矛盾,现实生活中,一个节点可能以不同的角色与其他节点交往,不同角色属于不同的社区,那么该节点属于多个社区并链接向不同的社区节点。为此,2005 年 Palla[7]等对传统的社区发现方法进行扩展,允许一个节点隶属于多个社区,开辟了重叠社区发现方法研究的新领域。重叠社区发现方法有团渗方法 CPM(clique percolation method)、线图方法、局部扩展方法[3]和基于统计推理的方法等。

微博是 2008 年出现的一种新兴社交网络,已成为越来越多国内外学者研究的热点。微博用户之间相互联系组成一个关系紧密、结构复杂的社会网络,微博网络中的社区是由用户基于相同或者相似的兴趣爱好分享与交流信息形成的一种小团体。微博网络也是一种网络,已有的社区发现方法也可以用于微博网络社区的发现,例如文献[8]基于图摘要的方法对微博用户进行聚类。但微博网络又区别于普通网络,它具有用户节点规模大、节点的度分布不均匀、弱社交网络[8]等特

到稿日期:2014-04-17 返修日期:2014-07-15 本文受中国博士后科学基金项目(20070420700),河北省自然科学基金项目(F2011201146)资助。
徐建民(1966—),男,博士,教授,主要研究领域为信息检索、不确定信息处理,E-mail:hbuxjm@hbu.edu.cn;**武晓波**(1986—),男,硕士生,主要研究领域为信息检索、社区发现;**吴树芳**(1979—),女,博士生,主要研究领域为信息检索、话题跟踪与检测;**粟武林**(1987—)男,硕士生,主要研究领域为信息检索与信息系统。

点。传统的社区发现方法不能发现重叠社区，算法的计算复杂度会因为网络节点的增多而变大，因此将传统的社区发现方法用于微博网络存在局限性。但是基于统计推理的社区发现方法能够用于重叠社区的发现并且适用于大规模节点的网络。基于统计推理的社区发现模型主要有混合模型 NMM (Newman's Mixture Model)、混合隶属度模型 MMM(Mixed Membership Model)、混合隶属度随机块模型 MMSB(Mixed Membership Stochastic Block model)[10,11] 和 aMMSB 模型 (assortative Mixed Membership Stochastic Block model)。

混合隶属度随机块模型 MMSB 和 aMMSB 模型，均用概率图解释了节点之间链接的生成，用于无向网络的社区发现，而微博网络是弱关系(单向关注关系)的有向网络，并且网络中的节点呈现异构性。本文在 aMMSB 模型的基础之上，结合微博网络的特点提出了带有向边的 WB-MMSB 模型，并在真实微博数据上验证了模型的有效性。

2 相关工作

2008 年 Airoldi 和 Blei[11] 等人将混合隶属度模型和随机块模型相结合，提出了 MMSB 模型，它可以用来识别重叠社区，时间复杂度为 $O(KN^2)$，适用于小型无向网络。2013 年 K. Gopalan 和 M. Blei[9] 等人利用随机变分推理算法 SVI (Stochastic Varitional Inference)[12] 改进了 MMSB 模型(以下简称 aMMSB 模型)，使之能够用于大规模节点的网络发现社区。本文微博网络社区发现是基于 aMMSB 模型，下面介绍 aMMSB 模型的相关基础知识。

aMMSB 模型用 $G=(V,R)$ 表示由节点和边组成的无向网络，其中 $V=\{1,2,\cdots,N\}$ 表示节点集合，R 表示无向边的集合，$y_{ij}\in R$ 表示节点对(i,j)之间存在边或者不存在边。

假定网络 G 中有 K 个社区，$\theta_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{iK})$ 为节点 i 在社区上的隶属度向量，其中 θ_{iK} 表示节点 i 隶属于社区 K 的概率，且 $\sum_{k=1}^{K}\theta_{ik}=1$。结合图 1，为了产生一个网络，对于网络中的节点对(i,j)选择社区指示器 $z_{i\to j}$ 和 $z_{i\leftarrow j}$，社区指示器指向 K 个社区中的一个。当网络中两个节点指示器都等于 k(即 $z_{i\to j}=z_{i\leftarrow j}=k$)时，$\beta_k$ 表示这两个节点链接的概率，即社区 k 的强度。图 1 是网络中无向边生成的概率图，假设节点对(i,j)有边存在即 $y_{ij}=1$；反之 $y_{ij}=0$。图中 y_{ij} 为可观测量，(θ,z,β)为未知的潜在变量，(α,η)为参数。

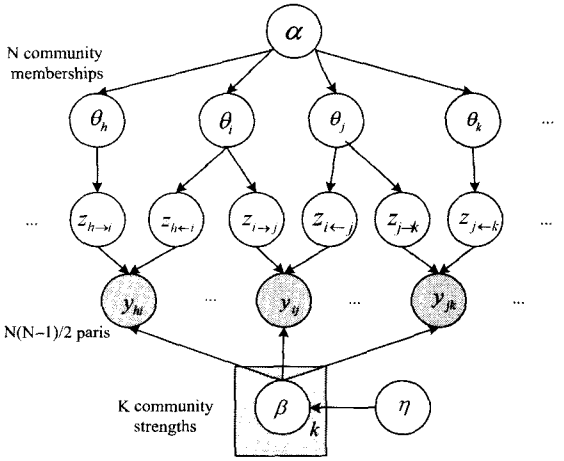

图 1 aMMSB 模型的概率生成图

根据以上 aMMSB 模型的描述，结合图 1，无向网络的生成步骤如下：

(1)对每个节点 $i\in V$，从狄利克雷分布 $Dir(\alpha)$ 上取样 K 维社区隶属度向量 θ_i；

(2)对节点对$(i,j)\in N\times N$ 且 $i<j$：

从多项式分布 $Mult(\theta_i)$ 上为节点 i 取样社区指示器 $z_{i\to j}$；

从多项式分布 $Mult(\theta_j)$ 上为节点 j 取样社区指示器 $z_{i\leftarrow j}$。

从如下的概率分布取样链接：

$$p(y_{ij}=1\mid z_{i\to j},z_{i\leftarrow j})=\begin{cases}\beta_{z_{i\to j}}, & \text{if } z_{i\to j}=z_{i\leftarrow j}\\ \varepsilon, & \text{if } z_{i\to j}\neq z_{i\leftarrow j}\end{cases}$$

aMMSB 模型能够处理大规模节点的真实网络，并且可以识别重叠社区。但该模型没有考虑边的指向，同等看待边的有向性和无向性，故将其用于微博网络的社区发现存在一定的不足。

图 2 是从微博网络中抽取的一个子集。假定这个子集包含 3 个社区：影视娱乐社区、科技社区和经济社区，其中科技社区和经济社区的一些微博用户关注了影视明星姚晨。aMMSB 模型中当两个节点对之间的社区指示器相同时，根据节点之间边的后验概率来计算节点的社区隶属度。由图 2 可知，姚晨很多的粉丝是科技社区和经济社区，不考虑边的指向，那么姚晨也隶属于科技社区和经济社区，现实生活中我们可知姚晨只隶属于影视娱乐社区。而在微博网络中考虑节点之间的方向性，节点指示器也按照方向只考虑一个社区指示器。图 2 中姚晨集中关注的用户是影视娱乐明星(例如舒淇、李冰冰、伊能静等)，那么姚晨应该隶属于影视娱乐社区；同时经济社区和科技社区的一些用户关注了姚晨，这些人也可能隶属于影视娱乐社区，只是他们在影视娱乐社区的隶属度比较低，而他们更多地关注是本社区内的用户，因此他们分别在科技社区和经济社区的隶属度较大。

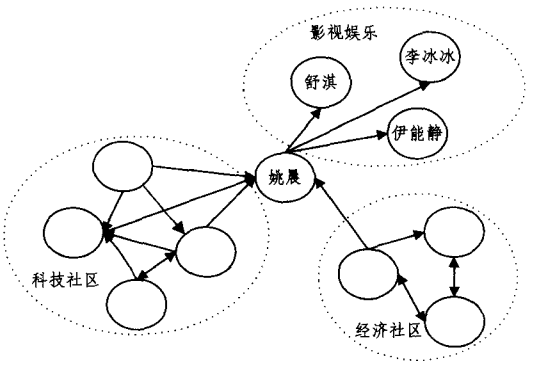

图 2 微博社区实例图

微博网络考虑节点之间的单向交互，能更准确地识别社区，同时提高计算节点社区隶属度的精度。为此，本文考虑节点之间的单方面关系，改进 aMMSB 模型，使之适用于微博网络的社区发现。

3 WB-MMSB 模型与参数推导

3.1 微博网络拓扑图

微博网络中节点呈现明显的异构特征，名人、媒体等节点被大量的普通节点关注，而名人、媒体等节点很少关注普通节

点。图 3 是从新浪微博网络中提取的一个典型关注关系的网络拓扑图(黑圈为名人节点,白圈为普通节点,带箭头的边为关注关系)。节点的关注关系是基于不同的兴趣,而这种兴趣往往表现为不同的主题社区。为了区分社交网络与图论中的术语,以下称关注关系为有向边,如图 3 中节点 i 有边指向节点 j,用 $y_{i\to j}$ 表示。$y_{i\to j}$ 对节点 i 称为链出,对节点 j 称为链入。节点 j 被大量的普通节点链入,反映了普通节点对其某些主题感兴趣;而节点 i 的链出边反映了节点 i 对其它节点某些主题感兴趣,因此一个节点的主题兴趣可以通过链入边和链出边来表现。

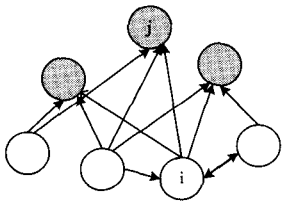

图 3 微博网络拓扑图

$y_{i\to j}$ 反映的是节点 i 对节点 j 单方面的指向关系,而 aMMSB 模型是基于节点间双向交互关系,本文在 aMMSB 模型的基础上基于节点之间的单方面关系,从两个方面考虑微博网络中节点的主题隶属度,提出 WeiBo-MMSB(简称 WB-MMSB)模型。

3.2 WB-MMSB 模型

微博网络中每个节点有链入边和链出边,每个节点从链入边和链出边两个方面表示节点在主题社区上隶属度的分布。节点 i 链出主题向量 $\theta_{i,\to}$ 表示其链出边在主题社区上的概率分布,即 $\theta_{i,\to}=(\theta_{i1,\to},\theta_{i2,\to},\cdots,\theta_{iK,\to})$,其中 $\theta_{iK,\to}$ 表示节点 i 的链出边属于主题社区 K 的概率,且有 $\sum_{k=1}^{K}\theta_{ik,\to}=1$。同样有节点 i 链入主题向量 $\theta_{i,\leftarrow}$ 表示其链入边在主题社区上的概率分布,即 $\theta_{i,\leftarrow}=(\theta_{i1,\leftarrow},\theta_{i2,\leftarrow},\cdots,\theta_{iK,\leftarrow})$,其中 $\theta_{iK,\leftarrow}$ 表示节点 i 的链入边属于主题社区 K 的概率,同样有 $\sum_{k=1}^{K}\theta_{ik,\leftarrow}=1$。

模型中节点间关系是有方向的,节点对 (i,j) 的社区指示器考虑单方向的情况,社区指示器 $z_{i\to j}\in\{1,2,\cdots,K\}$ 不再考虑节点指示器 $z_{i\leftarrow j}$。当节点 i 有向边指向节点 j 时,$y_{i\to j}=1$;无指向时,$y_{i\to j}=0$。由以上可知节点 i 链接向节点 j 生成有向边的联合条件概率为:

$$p(y_{i\to j}=1|\theta_{i,\to},\theta_{j,\leftarrow})=\sum_{k=1}^{K}\theta_{ik,\to}\cdot\theta_{jk,\leftarrow}\cdot\beta_k$$

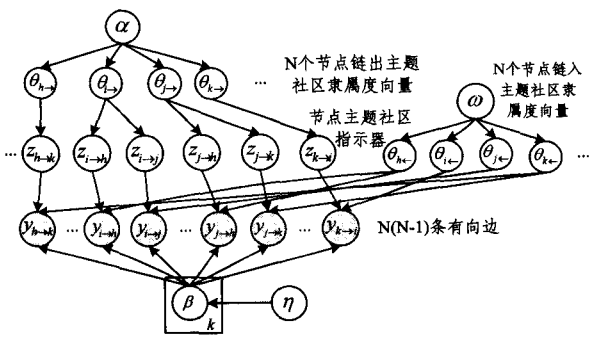

图 4 WB-MMSB 模型的概率生成图

结合图 4,有向网络生成的步骤如下:

(1)对于每个节点 $i\in V$,从狄利克雷分布 $Dir(\alpha)$ 上取样 K 维链出主题隶属度向量 $\theta_{i,\to}$;

(2)对于节点对 (i,j) 且 $j\neq i$,

a)对于节点 j,从狄利克雷分布 $Dir(\omega)$ 上取样 K 维链入主题隶属度向量 $\theta_{j,\leftarrow}$;

b)从多项式分布 $Mult(\theta_{i,\to})$ 上为节点 i 取样社区指示器 $z_{i\to j}$;

c)节点对 (i,j) 生成有向边 $y_{i\to j}$ 的概率:

$$p(y_{i\to j}=1|z_{i\to j}=k)=\theta_{ik,\to}\cdot\theta_{jk,\leftarrow}\cdot\beta_k$$

概率分布 $\theta_{\to}$ 和 $\theta_{\leftarrow}$ 是参数分别为 α 和 ω 的狄利克雷分布,它们是 $N\times K$ 的矩阵,社区指示器 Z 是 $N\times N$ 的矩阵,β 是参数为 φ 的 Beta 分布,它是 $2\times K$ 的矩阵。

3.3 模型的参数推导

WB-MMSB 模型的潜在变量为 $(\theta_{\to},\theta_{\leftarrow},z,\beta)$,可观测量为 y,要计算的后验概率如式(1),式中分子为各变量联合概率分布,比较容易求解,但分母涉及到所有变量求和,不易求解。在贝叶斯模型后验概率不易求解时,一般用逼近算法求得近似解。近似求解方法主要有 MCMC(Markov Chain Monte Carlo)和变分贝叶斯方法。MCMC 计算后验概率时迭代考虑整个网络中所有的节点对,对于社区发现中后验概率的计算效率不高。为此本文使用随机变分推理算法 SVI,其每次迭代只需要随机抽取网络的一个子集,估计当前的梯度,然后计算这时的变量和参数,迭代计算各参数的速度快。SVI 算法是变分贝叶斯方法的一种,本文的近似求解利用 SVI 算法,其中 SVI 用到两个关键的方法:平均场变分推理(mean-field variational inference)和随机最优(stochastic optimization)。

$$p(\theta_{\to},\theta_{\leftarrow},\beta,z|y)=\frac{p(\theta_{\to},\theta_{\leftarrow},\beta,z,y)}{p(y)}\tag{1}$$

对于求解的后验概率 $p(\theta_{\to},\theta_{\leftarrow},\beta,z|y)$,在平均场家族定义一个含参数的潜在变量分布 $q(\theta_{\to},\theta_{\leftarrow},\beta,z)$,通过平均场变分推理逼近后验概率,后验概率和定义的含参分布之间的近似度用 KL(Kullback-Leibler)散度来衡量:

$$q^*(\theta_{\to},\theta_{\leftarrow},\beta,z)=\arg\min KL(q(\theta_{\to},\theta_{\leftarrow},\beta,z)\parallel p(\theta_{\to},\theta_{\leftarrow},\beta,z|y))\tag{2}$$

$q^*(\theta_{\to},\theta_{\leftarrow},\beta,z)$ 作为最优化的逼近后验概率,式(2)计算比较困难,进一步优化一个目标函数(式(3)),则将求 KL 的最小值转化为求 L 的最大值。

$$L=-KL(q(\theta_{\to},\theta_{\leftarrow},\beta,z)\parallel p(\theta_{\to},\theta_{\leftarrow},\beta,z|y))\tag{3}$$

在平均场变量家族[10],$q(\theta_{\to},\theta_{\leftarrow},\beta,z)$ 的各变量分布由各自的参数决定,与其他变量相互独立,设 $q(\theta_{\to},\theta_{\leftarrow},\beta,z)$ 的变量 $\theta_{\to},\theta_{\leftarrow},\beta,z$ 的各自参数分别为 γ,μ,λ,ϕ,在平均场家族 $q(\theta_{\to},\theta_{\leftarrow},\beta,z)$ 的分布形式为式(4)。

$$q(\theta_{\to},\theta_{\leftarrow},\beta,z)=\prod_{n=1}^{N}q(\theta_{n,\to}|\gamma_n)q(\theta_{n,\leftarrow}|\mu_n)\prod_{k=1}^{K}q(\beta_k|\lambda_k)\prod_{i\neq j}q(z_{i\to j}|\phi_{i\to j})\tag{4}$$

由平均场家族分布和变分推理可得目标函数式(3)转化为包含一个期望和熵的式(5),其中 $const$ 是常量,要求解的变量分布 $(\theta_{\to},\theta_{\leftarrow},z,\beta)$ 转化为关于参数 $(\gamma,\mu,\lambda,\phi)$ 的函数。

$$L(\gamma,\mu,\lambda,\phi)=E_q[\log p(\theta_{\to},\theta_{\leftarrow},\beta,z)]-H[q(\theta_{\to},\theta_{\leftarrow},\beta,z)]+const\tag{5}$$

3.4 模型的参数估计

SVI算法最优化式(5)，使用随机梯度上升算法，固定其它参数不变，迭代最优得到各个参数值。参数中全局参数为 γ,μ,λ，局部参数为 ϕ，为了区分全局参数与局部参数，用角标 g 和 l 分别表示参数是关于全局变量和局部变量。对于参数 λ，固定其他参数不变，则式(5)转化为一个关于 λ 的函数：

$$L(\lambda)=E_q[\log p(\beta|z,y,\varphi)]-E_q[\log q(\beta|\lambda)]+const \quad (6)$$

式(4)给出了各自变量的参数，但不能确定变量分布的形式，而是指定它们是同样形式的指数族分布。由指数族分布[10]定义可知，式(6)右边的两个条件概率为：

$$p(\beta|z,y,\varphi)=h(\beta)\exp\{\eta_k^{\mathrm{T}}(z,y,\varphi)t(\beta)-a_g(\eta_g^{\mathrm{T}}(z,y,\varphi))\}$$

$$q(\beta|\lambda)=h(\beta)\exp\{\lambda^{\mathrm{T}}t(\beta)-a_g(\lambda)\}$$

指数族分布[10]具有性质 $E_q[t(\beta)]=\nabla_\lambda a_g(\lambda)$。根据指数族分布及其性质，式(6)可以变为

$$L(\lambda)=E_q[\eta_g^{\mathrm{T}}(z,y,\eta)]\nabla_\lambda a_g(\lambda)-\lambda^{\mathrm{T}}\nabla_\lambda a_g(\lambda)+a_g(\lambda)+const \quad (7)$$

式(7)是关于全局变量的参数 λ 的函数，为得到相应的更新梯度，对其两边求导：

$$\nabla_\lambda L(\lambda)=\{E_q[\eta_g^{\mathrm{T}}(z,y,\varphi)]-\lambda\}\nabla_\lambda^2 a_g(\lambda) \quad (8)$$

根据随机最优梯度上升算法，令 $\nabla_\lambda L(\lambda)=0$，得式(9)是一个随机变量和可观测值的期望，该式暗示要得到最优 λ，只需固定其他变量不变，更新参数 λ 即可。

$$\lambda_i=E_q[\eta_g(z_{i\to j},y_{i\to j},\varphi)] \quad (9)$$

同理，由指数分布族和随机最优梯度算法得到其它参数表达式，如下：

$$\gamma_{i,k}=E_q[\eta_g(z_{i\to j},\alpha)]$$

$$\mu_{j,k}=E_q[\eta_g(z_{i\to j},\omega)]$$

$$\phi_{i\to j,k}=E_q[\eta_l(\theta_{i,\to},\theta_{j,\leftarrow},\beta,y_{i\to j})]$$

3.4.1 参数估计算法

输入：社区数 K，包含节点之间关系的有向图 G。

输出：参数 γ,μ,λ。

1. 随机初始化全局参数 $\gamma=(\gamma_n)_{n=1}^N$，$\mu=(\mu_n)_{n=1}^N$，$\lambda=(\lambda_k)_{k=1}^K$；

2. 取样全局网络的子集 S；

3. 局部阶段：$\forall(i,j)\in S$，由式(10)计算局部参数 $\phi_{i\to j}$ 的最优解；

4. 全局阶段：$\forall(i,j)\in S$，计算节点 i 的链出主题向量随机梯度 $\partial\gamma_i^t$，更新 $\gamma_{i,k}^t\leftarrow\gamma_{i,k}^{t-1}+\rho_{t-1}\partial\gamma_{i,k}^{t-1}$，计算节点 j 的链入社区隶属度随机梯度 $\partial\mu_j^t$，更新 $\mu_{j,k}^t\leftarrow\mu_{j,k}^{t-1}+\rho_{t-1}\partial\mu_{j,k}^{t-1}$，对于每一个社区 k，计算社区强度的随机梯度 $\partial\lambda_k^t$，更新 $\gamma_{i,k}^t\leftarrow\gamma_{i,k}^{t-1}+\rho_{t-1}\partial\gamma_{i,k}^{t-1}$，$\rho_t=(\tau_0+t)^{-\kappa}$，$t\leftarrow t+1$；

5. 如果参数 γ,μ,λ 收敛，则结束；否则返回步骤 2。

在上述算法中，ρ_t 满足 $\sum_t\rho_t^2<\infty$ 且 $\sum_t\rho_t=\infty$，根据文献[12]设定 $\rho_t=(\tau_0+t)^{-\kappa}$，其中 $\kappa\in(0.5,1]$ 是遗忘率，τ_0 为延迟迭代权重。

3.4.2 参数估计算法中的梯度计算

全局阶段：包含 $O(N)$ 个节点的有向网络含有 $M=N(N-1)$ 个节点对，依照 $g(i,j)$ 分布随机抽取一对节点 (i,j)。根据文献[12]计算在第 t 次迭代中，全局变量的参数随机梯度如下：

$$\partial\gamma_{i,k}^t=\alpha_k+\frac{1}{g(i,j)}\sum_{(i,j)\in s_t}\phi_{i\to j,k}-\gamma_{i,k}^{t-1}$$

$$\partial\mu_{j,k}^t=\omega_k+\frac{1}{g(i,j)}\sum_{(i,j)\in s_t}\phi_{i\to j,k}-\mu_{j,k}^{t-1}$$

$$\partial\lambda_{k,0}^t=\varphi_k+\frac{1}{g(i,j)}\sum_{(i,j)\in s_t}\phi_{i\to j,k}\cdot y_{i\to j}-\lambda_k^{t-1}$$

$$\partial\lambda_{k,1}^t=\varphi_k+\frac{1}{g(i,j)}\sum_{(i,j)\in s_t}\phi_{i\to j,k}\cdot(1-y_{i\to j})-\lambda_k^{t-1}$$

局部阶段：参数 $\phi_{i\to j}$ 最优表达式，如下：

$$\phi_{i\to j,k}^t|y_{i\to j}=1\propto\exp\{E_q[\log\theta_{i\to,k}]+E_q[\log\theta_{j\leftarrow,k}]+\phi_{i\to j,k}^{t-1}\cdot E_q[\log\beta_k]\} \quad (10)$$

$$\phi_{i\to j,k}^t|y_{i\to j}=0\propto\exp\{E_q[\log\theta_{i\to,k}]+E_q[\log\theta_{j\leftarrow,k}]+\phi_{i\to j,k}^{t-1}\cdot E_q[\log(1-\beta_k)]\}$$

4 实验分析

4.1 实验数据

目前对中文微博的研究还处于起步阶段，尚无标准的数据集，本文使用新浪微博开发平台提供的 API 接口收集了230000用户的基本信息，去除关注列表为空的无效数据，得到17139个有效用户的基本信息。每个用户信息包括：用户ID、屏幕名、性别、自我介绍、标签、地区、关注列表等。根据微博用户自己定义的标签，统计出16个主题社区：音乐、电影、美食、IT数码、体育、新闻传媒、母婴保健、手工艺设计、微博助手、政府官方微博、文艺漫画、电视台主持人、时尚、经济投资、电子商务平台和生活休闲。这16个主题社区可以整合为更少数目的社区。从17139个微博用户中挑选有标签的15030个微博用户作为实验数据，其中有120682条链接(含保留数据集 H)。实验数据包含两部分：训练数据集和保留数据集 H(held-out set)，训练数据集用于参数和变量的学习，保留数据集用于模型的测试比较，其中保留数据集 H 是从全部数据中随机抽取出5%的数据(同文献[9])，其中一半有链接，另外一半无链接。

4.2 实验设置与评测

为了便于比较，实验中参数设置统一。由于 K 是社区的数目，在真实的生活中均匀分配各社区强度的先验[9]，即狄利克雷分布超参数 $\alpha=\omega=\frac{1}{K}$。迭代步长 $\rho_t=(\tau_0+t)^{-\kappa}$，其中 $\kappa\in(0.5,1]$ 是遗忘率，用来控制旧信息被遗忘的速度，τ_0 为延迟迭代权重，用来控制每次迭代下降的大小，对于任意满足条件的 κ 和 τ_0，算法都会收敛于局部最优值，只是用它们的大小来控制收敛的速度。根据经验设置 $\tau_0=1024$，以确保其迭代步长在 1.0×10^{-3} 10.0×10^{-3} 之间。对于遗忘率 κ，实验中我们分别设置 $\kappa\in\{0.5,0.6,0.7,0.8,0.9,1.0\}$，发现其对算法的收敛影响不明显。文献[9]中社区发现的节点主题隶属度遗忘率 $\kappa=0.5$，社区强遗忘率 $\kappa=0.9$，为了便于比较，也设置为同样的值。参数估计中全局阶段使用均匀抽样，即 $g(i,j)=\frac{1}{N(N-1)}$。

本文采用两种评测指标，即归一化互信 NMI[16] 和算法收敛的困惑度(Perplexity)[17]。NMI 度量(式(11))是基于混淆矩阵 N，矩阵 N 中的行代表真实的社区，列代表发现的社区，N_{ij} 代表真实社区 i 中的节点出现在发现社区 j 中的节点数目。c_A 表示真实社区的数目，c_B 表示发现社区的数目，$N_{i\cdot}$

表示对第 i 行的所有元素求和，$N_{\cdot j}$ 表示对第 j 列的所有元素求和。当发现的社区与真实社区吻合时，$I(A,B)=1$；当整个网络属于一个社区时，$I(A,B)=0$。困惑度(Perplexity)用来度量模型的拟合程度，困惑度越低，效果越好。

$$I(A,B)=\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}N_{ij}\log\left(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}\right)}{\sum_{i=1}^{c_A}N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right)+\sum_{j=1}^{c_B}N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)} \quad (11)$$

$$perplexity(H)=\exp\left\{-\frac{\sum_{(i\to j)\in H}\log p(y_{i\to j}\mid y_{observed})}{|H|}\right\} \quad (12)$$

$$p(y_{i\to j}\mid y_{observed})\approx\sum_{z_{i\to j}}p(y_{i\to j}\mid z_{i\to j},\hat{\beta})p(z_{i\to j}\mid\hat{\theta}_{i,\to}) \quad (13)$$

4.3 实验比较与分析

针对本文提出的 WB-MMSB 模型和 aMMSB[9] 模型，分别进行了 3 次实验。实验 1 比较 aMMSB 模型改进前后在不同的社区数目 K 下的困惑度，实验 2 比较了两个模型的算法收敛速度，实验 3 比较不同 δ 值下的归一化互信息。

4.3.1 困惑度比较

由于现在还没有标准的微博数据集，实验中采集的数据集社区的数目不确定，且由于人工统计的 16 个主题社区和真实的情况还有一定的误差，因此比较了不同的社区数目下的困惑度(如图 5 所示)。首先用训练集学习模型中各个参数，然后在保留数据集 H 上计算 $Perplexity$。在 WB-MMSB 模型中，可以通过后验估计变量参数$(\gamma,\mu,\lambda,\phi)$点估计来计算社区隶属度$(\hat{\theta}_{i,\to},\hat{\theta}_{j,\leftarrow})$和社区强度$\hat{\beta}$，则通过式(13)预测一对节点之间的链接，aMMSB 模型使用文献[9]的计算公式，把实验数据的有向链接当为无向链接。

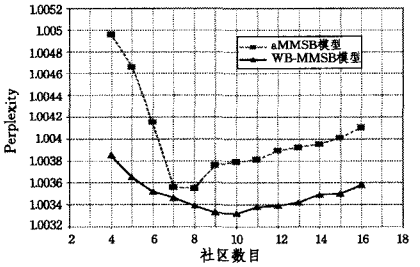

图 5 不同的社区数目下模型的困惑度比较

由图 5 可知，aMMSB 模型的困惑度随着实验中设定的社区数目的增加先变小然后变大，当实验中设定社区的数目 $K=8$ 时，困惑度最小，此时模型的拟合程度最好。而 WB-MMSB 模型的困惑度在 $K=10$ 时，困惑度最小。由图还知，不论社区数目设为多少，WB-MMSB 模型的困惑度都要比 aMMSB 模型的低。

4.3.2 收敛速度比较

基于 SVI 算法模型的收敛速度度量。为了使实验的数据拟合最佳，同时便于比较，设定社区的数目为 $K=8$，此时 aMMSB 模型困惑度最低。

实验使用保留数据集 H 的平均预测 log 似然函数值，当实验中的平均困惑度小于 0.001%或者平均困惑度不再增长时，停止测试，平均困惑度使用式(12)计算。实验中计算保留数据集 H 的全部数据困惑度需要一段时间，在这个过程中我们每隔 1 分钟统计平均困惑度，直到第 16 分钟程序停止计算，输出每个时刻的平均困惑度。

图 6 是两个模型收敛速度的度量，开始 3 分钟内 WB-MMSB 模型的困惑度大于 aMMSB 模型，从第 4 分钟开始 WB-MMSB 模型的困惑度下降速度快于 aMMSB 模型，并且在第 11 分钟 WB-MMSB 模型的困惑度已趋于收敛。实验表明，WB-MMSB 模型收敛速度快于 aMMSB 模型。

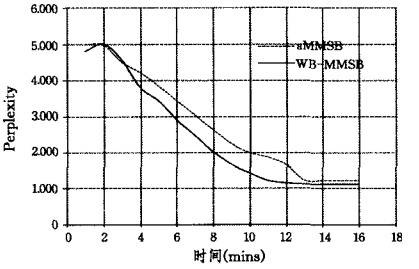

图 6 改进前后模型收敛速度的比较

4.3.3 互信息比较

为比较 aMMSB 模型改进前后社区发现的精确度，采用文献[16]中的归一化互信息(见式(11))进行评价。社区数目分别选取新旧模型困惑度最小时的社区数目，此时模型的拟合程度最好。WB-MMSB 模型对于节点 i 的社区隶属度计算综合节点链入主题隶属度和链出主题隶属度，即 $\theta_i=\sigma\theta_{i,\to}+(1-\sigma)\theta_{i,\leftarrow}$，$\sigma$ 是反映节点 i 链入主题和链出主题的权重，实验中设置 σ 的值为 0.2～0.9。当 $\theta_i\geqslant 0.001$ 时，规定节点 i 属于该社区。表 1 是选取社区数目 $K=8$ 和 $K=10$，且 WB-MMSB 模型选取不同的 σ 下新旧模型 NMI 值的比较。

表 1 不同社区数目下新旧模型互信息(NMI)的比较

社区数目	aMMSB	不同 σ 值下 WB-MMSB 性能							
		0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
K=8	0.6	0.58	0.65	0.63	0.76	0.82	0.79	0.83	0.85
K=10	0.72	0.75	0.73	0.78	0.83	0.89	0.9	0.88	0.87

由节点 i 的社区隶属度计算公式可知，σ 越小，θ_i 越趋近于 $\theta_{i,\leftarrow}$，节点的社区隶属度更多地由链入主题隶属度来衡量；反之，σ 越大，节点的社区隶属度更多地由链出主题隶属度来衡量。由表 1 可知，当社区数目 $K=8$，$\sigma=0.2$ 时，WB-MMSB 模型的 NMI 值比原模型要低，σ 取其它值时，WB-MMSB 模型的 NMI 值比原模型的要高，即社区发现的准确度更高；当社区数目设定 $K=10$ 时，不同的阈值 σ 下，WB-MMSB 模型的 NMI 值都比原模型高。以上表明微博网络中普通节点的社区隶属度更多地由链入主题隶属度来衡量，并且模型的拟合程度越好，社区发现的精确度越高。

结束语 本文提出了一种用于微博网络的社区发现模型，考虑了节点之间的指向，同时节点的社区隶属度从链入主题社区隶属度和链出主题社区隶属度两方面考虑。实验以归一化互信息为评价标准，在不同的链入主题和链出主题权重 σ 下与原模型比较，结果表明 σ 越大 NMI 值越高，从而提高了模型社区发现的性能。但本文还有一定的不足，如微博测试集选取的局限性，需在以后的工作中进一步完善。

参 考 文 献

[1] 丁连红，时鹏. 网络社区发现[M]. 北京：化学工业出版社，2008：1-138

[2] Girvan M, Newman M E J. Community structure in social and biological network[J]. Proceedings of National Academy of Sciences, 2002, 99(12): 7812-7826

[3] 杨博,刘大,Liu Ji-ming,等. 复杂网络聚类方法[J]. 软件学报,2009,20(1):54-66

[4] 程学旗,沈华伟. 复杂网络的社区结构[J]. 复杂系统与复杂性科学,2011,8(1):57-70

[5] 樊鹏翼,王晖,姜志宏,等. 微博网络测量研究[J]. 计算机研究与发展,2012,49(4):691-699

[6] 郭岩,白硕,杨志峰,等. 网络日志规模分析和用户兴趣挖掘[J]. 计算机学报,2005,28(9):1483-1496

[7] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[8] 陈克寒,韩盼盼,吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报,2013,36(2):349-358

[9] Gopalan P K, Blei D M. Efficient discovery of overlapping communities in massive network[J]. Proceedings of the National Academy of Science of the United States of American, 2013, 110 (36): 14534-14539

[10] Wainwright M J, Jordan M I. Graphical Models, Exponential Families, and Variational Inference[J]. Foundations and Trends in Machine Learning, 2008, 1(1/2): 1-305

[11] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic block models[J]. Journal of Machine Learning Research, 2008, 9: 1981-2014

[12] Hoffman M, Blei D M, Wang Chong, et al. Stochastic variational infernce[J]. Journal of Machine Learning Research, 2013, 14: 1303-1347

[13] Hastings M B. Community detection as an inference problem [J]. Physical Review E-PHYS REV E, 2006, 74(3)

[14] Gopalan P, Wang Chong, Blei D M. Modeling overlapping communities with node popularities[C]// Advances in Neural Information Processing Systems. 2013: 2850-2858

[15] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005(9): P09008

[16] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[17] 沈华伟,程学旗,陈海强,等. 基于信息瓶颈的社区发现[J]. 计算机学报,2008,31(4):677-686

[18] Robbins H, Monro S. A stochastic approximation method[J]. The Annals of Mathematica Statistics, 1951, 22(3): 400-407

[19] 杨楠,弓丹志,李饮,等. Web 社区发现技术综述[J]. 计算机研究与发展,2005,42(3):439-447

[20] 林友芳,王天宇,唐锐,等. 一种有效的社会网络社区发现模型和算法[J]. 计算机研究与发展,2012,49(2):337-345

[21] Gregory S. Find overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 103018

[22] Yan B, Gregory S. Detecting community structure in network using edge prediction methods[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012(9): P09008

[23] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large network[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(9): P10008

[24] He Dong-xiao, Liu Da-you, Zhang Wei-xiong, et al. Discovering link communities in complex networks by exploiting link dynamics[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012(10): P10015

(上接第 46 页)

[7] Wang H, Hu J D. Logistic-Map chaotic spread spectrum sequence[J]. ACTA Electronica Sinica, 1997, 25(1): 19-23

[8] Jessa M. The period of sequences generated by tent-like maps [J]. IEEE Trans. Circuits Syst. I, Fundam. Teory Appl, 2002, 49 (1): 84-88

[9] Kurian A P, Puthussery S, Htut S M. Performance enhancement of DS/CDMA system using chaotic complex spreading Sequence [J]. IEEE Transactions on Wireless Communications, 2005, 4 (3): 984-989

[10] Youssef M I, Zahara M, Emam E. Chaotic Sequences Implementations on Residue Number Spread Spectrum System[J]. International Journal of communications, 2008, 2(2): 143-154

[11] Ihan M, Philip, Andi S. Chaos Codes vs. Orthogonal Codes for CDMA[C]// ISSSTA 2010. Taichung, Taiwan, China, 2010: 189-193

[12] 张薇,谢红梅,王保平. 一种新型的分段 Logistic 混沌扩频通信算法[J]. 计算机科学,2013,40(1):59-62

[13] 王保平,李文康,吴成茂. 改进分段 Skew Tent 映射及其在扩频通信中应用[J]. 红外与激光工程,2013,42(10):2772-2777

[14] Nazila R, Siamak T. Performance comparison of chaotic spreading sequences generated by two different classes of chaotic systems in a chaos-based direct sequence-code division multiple access system[J]. IET Communications, 2013, 10(7): 1024-1031

[15] Zan L, C Jue-ping, C Yi-lin. Determining the Complexity of FH/SS Sequenceby Approximate Entropy[J]. IEEE Transactions on Communications, 2009, 57(3): 812-820

[16] 刘金梅,丘水生. 混沌伪随机序列复杂性的一种量度方法[J]. 计算机应用,2009,29(4):938-941

[17] Nguyen L. Self-encoded spread spectrum communications[C]// Proceedings IEEE MILCOM'99. Atlantic City, NJ, 1999: 182-186

[18] Wei M, Li Z L, Yin F. Analysis and Simulation of AR-SESS System Performance[C]// 2005 International Conference on Communications, Circuits and Systems. Chengdu, China, 2005: 160-164

[19] Duraisamy P, Nguyen L. Coded-sequence self-encoded spread spectrum communications[C]//Proceedings of the IEEE Global Telecommunications conference Hawaii. Honolulu, HI, 2009: 1-5

[20] 王福来. 基于复合符号混沌的伪随机数生成器及加密技术[J]. 物理学报,2011,60(11):191－197

[21] 张瀚,王秀峰,李朝晖,等. 一种基于混沌系统及 Henon 映射的快速图像加密算法[J]. 计算机研究与发展,2005,42(12):2137-2142