# 社区环境下基于节点交互和主题的影响力计算模型

王大刚[1],钟 锦[1,2],吴 昊[1]
(1. 合肥师范学院计算机学院,安徽合肥 230601;2. 中国科学技术大学计算机学院,安徽合肥 230600)

**摘 要:** 为解决现有算法对社交网络节点影响力计算准确度不高的问题,本文整合节点不同维度信息,综合考虑节点在多个主题社区上的主题分布向量,提出一种新的节点影响力计算模型. 模型首先将主题相关性作为先验信息;然后利用混合隶属度随机块(Mixed Membership Stochastic Block)模型表达节点间的交互关系,用主题模型学习主题内容;最后结合全局拓扑关系迭代计算节点的全局影响力. 本文选取社交网络数据,以 P@ N、MAP 等作为评价指标同现有主流算法进行比较. 实验结果显示,本文算法有效提升了影响力节点识别的准确度和排名的有效性.

**关键词:** 主题; 影响力; 混合隶属度随机块; 先验
**中图分类号:** TP311 **文献标识码:** A **文章编号:** 0372-2112 (2020)03-0582-08
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.03.023

# Influence Calculation Model Based on Node Interaction and Topic in Community Environment

WANG Da-gang[1], ZHONG Jin[1,2], WU Hao[1]
(1. *School of Computer Science and Technology, Hefei Normal University, Hefei, Anhui* 230601, *China*;
2. *School of Computer Science and Technology, University of Science and Technology of China Hefei, Anhui* 230600, *China*)

**Abstract:** To solve the accuracy problems of the existing algorithms in calculating the influence of social network nodes, by integrating different dimension information of nodes, and considering the topic distribution vector of nodes on multiple communities, a new model is proposed. It first regards the correlation between topics as the prior information, then uses the mixed membership stochastic block (MMSB) model to express the interaction among nodes, learns topic contents using topic model, and finally, iteratively calculates the global influence of nodes with global topological relationship. We select data from social networks, use P@ N, MAP, etc., as the evaluation indicators, and compare the proposed algorithm with the existing mainstream algorithms. The experimental results show that our algorithm significantly improves the identification accuracy of influential nodes and the validity of ranking.

**Key words:** topic; influence; mixed membership stochastic block; prior

## 1 引言

社交网络的分析和研究中,影响力计算具有非常重要的地位. 影响力计算模型中,PageRank 算法是比较有代表性的方法,PageRank 将网页看成是个有向图结构,利用图的整体拓扑结构,用链接关系衡量网页影响力大小. 基于用户的 PageRank 分析用户影响力,扩展出许多基于拓扑结构的模型,比较典型的算法有 LeaderRank[1] 算法,模型从利用拓扑结构出发,从整体拓扑结构层面刻画节点的影响力. Li[2] 等人在 LeaderRank 基础上,考虑中心节点的权威性,通过简单加权对多种信息源进行整合,结合 PageRank 计算节点的影响力. 由于 PageRank 的拓扑结构算法模型只考虑节点的链接属性,忽略了主题相关性,导致结果的相关性和主题性减少. 但是社会网络不仅有用户间的链接关系,还有用户发布的各种基于主题的内容,而主题内容是影响力传

收稿日期:2019-03-06;修回日期:2019-07-03;责任编辑:梅志强
基金项目:安徽省自然科学基金(No. 1708085QF157);安徽省高校优秀青年人才支持计划(No. gxyq2017050);安徽省教育教学委托研究项目(No. 2018jyxm1470);国家大学生创新创业项目(No. 201914098034)

播的载体,同时也是背后的本质的机理.

本文在社交节点主题内容的基础上,考虑主题的社区相关性,把节点在社区中的交互关系和节点的主题属性结合,设计贝叶斯模型,动态生成用户的主题分布向量,在主题分布向量基础上,重新定义节点相似性,计算出转移概率,最后结合网络拓扑属性,计算模型的基于主题的全局影响力. 本文的主要贡献包括:

(1)将主题的相关性信息引入模型先验,利用相关性表达主题间影响力的相互影响.

(2)利用交互关系和主题内容学习节点的主题分布向量.

(3)提出统一的概率框架表达社区节点间的交互关系、主题分布,利用数据学习节点的联合后验分布概率. 同时把学习到的关系与拓扑结构结合,共同表达基于主题的全局影响力.

## 2　相关工作

在模型中结合用户的信息内容有助于对影响力计算做更精确的分析. 从用户主题的角度看,社会网络中人们都在不同的话题空间交流和通信. 研究者发现在不同的主题上影响力是不同的. 利用机器学习技术度量基于话题的影响力成为比较有代表性的研究方向. Barbieri[3] 等证明用户主题偏好可以影响用户的社交地位、权威和信任,因此提出基于主题的影响力级联模型 TIC (Topic based Independent Cascade),在此基础上进一步提出一种基于相似性的方法来研究基于主题的影响力最大化问题. Chen[4] 等提出线上线下分离策略来解决基于主题的影响力传播问题,其思想是在离线环节为每个主题筛选出备用节点集,然后在在线环节从备用节点集中选取种子节点,结果显示基于主题的信息传播模型的影响质量和影响效果均优于未考虑主题因素的信息传播模型. 文献[5]将网络主题基于邻接矩阵的网络边缘关系和 PageRank 结合计算社交网站节点间影响力,能够大幅提升用户排名的有效性. 这些基于 LDA( Latent Dirichlet Allocation ) 主题分析的影响力模型本质上都基于主题是独立性的假设,而且忽略用户行为对主题的影响.

Weng[6] 等人综合 Twitter 上用户关注网络的拓扑结构和用户之间的兴趣相似度,在 PageRank 的基础上提出 TwitterRank 模型以度量用户在不同话题上的影响力. 该模型将基于 LDA 主题模型计算出的相似度与结构相似度进行线性加权,计算出总体影响力. Dietz[7] 等人综合利用用户所发信息的文本内容和用户拓扑结构,分别提出了 Copycat Model( CM) 和 Citation Influence Model( CIM) 模型,通过 Gibbs 采样迭代计算文本内容的来源及其对应的概率,该模型把影响力作为一个概率分布的采样值,利用概率图模型学习出具体的值来衡量影响力. 这些算法只是把用户行为用于静态的拓扑结构研究. 为此文献[8]引入用户行为特征度量,通过区分用户转发行为挖掘微博中与主题相关的专家,提出概率生成模型 EMTM,可以对微博主题语义和主题中用户被转发概率分布同时进行建模,并且通过区分微博用户的“主题相关转发”和“跟随转发”两种转发行为,实现了微博中与主题相关的专家的挖掘发现. 模型认为在每个主题内对用户的点击、关注等行为服从独立的伯努利分布,分别对主题和行为建立概率分布,但模型没有考虑到行为和主题的相互作用. 文献[9]提出 UIRank 算法采用层次分析法对用户行为、内容主题等 4 项评价特征进行分析;文献[10]把用户的行为进行细化,定义四种用户关系,提出基于多关系的主题层次影响力模型 MultiRank;文献[11]基于 PageRank 算法,设计 MFP 模型从结构、行为、情感三个方面按照一定的权重结合,对用户的影响力进行综合度量. 这些算法利用线性加权进行多维特征融合,但由于各特征数据的维度、特征和作用等不同,权重系数其实很复杂、很难确定,且算法整体缺少动态的观点和可解释性.

## 3　基于交互关系和主题的全局影响力计算模型

### 3.1　利用交互关系进行影响力建模

上文分析看出用户间的影响力由于受到多方面因素的影响,计算非常复杂,需要考虑用户属性、用户之间交互活动以及网络整体拓扑的影响. 为此本文选择基于主题的相似度计算方法,从统计学意义上考虑用户间的主题相似性. 社区中影响力较大的人对被影响的人造成影响,会一定程度上改变被影响人的社区主题属性,当然被影响人也以一定概率接受其它主题兴趣;另一方面在某个社区主题内部有较高影响力的节点,如果该主题在整个社区中的规模很小,基于全局拓扑结构计算出来的值可能并不高. 所以基于局部性的主题分析与全局结构视图的拓扑结构融合在一起,会比较完整反映节点的全局影响力属性. 据此本文采用的概率框架将用户的主题因素、节点的交互关系和整个网络的拓扑结构统一到同一个模型中,算法分成两个阶段:第一个阶段利用改进的基于 LDA 的主题模型,计算出节点的主题分布; 第二个阶段利用主题分布向量计算节点间的转移概率,然后结合全局拓扑关系,借鉴 PageRank 的算法思想,利用整个网络图结构迭代计算节点的全局影响力.

模型首先考虑交互关系对主题的影响,并进一步构造结合节点间交互关系的主题分布向量. 研究表明用户之间的的关注、转发、评论等关系是用户相似性的强烈指

示. 客观世界中,每个人的后天兴趣构成可以看成是这个人的先天本性、周围的环境和朋友影响综合而形成的. 在模型中给先天本性赋予一个先验分布,节点与周围环境的交互用节点间链接关系来表达,每个用户的兴趣分布可以通过训练数据来学习,试图利用模型来反映随着用户链接交互不断发生,用户间的主题产生动态变化. 假设节点 $i$ 与节点 $j$ 发生链接关系,$i$ 是关注节点,$j$ 是被关注节点. 通过好友、关注等交互活动之后,节点 $j$ 以一定的概率接受节点 $i$ 的主题. 在这之后节点 $j$ 的主题内容由之前的自身属性和受影响获得的属性共同构成了节点 $j$ 的新的主题属性. 社区环境下节点 $i$ 和节点 $j$ 发生交互关系,节点 $j$ 的主题内容由之前的先验 $\alpha$ 和节点 $i$ 的主题 $\boldsymbol{\theta}_i$ 构成新的主题 $\boldsymbol{\theta}_j$,通过主题模型和概率图结构,节点 $j$ 的主题内容从主题向量中重新采样形成新的主题词汇 word. 其关系可以用图 1 表达.

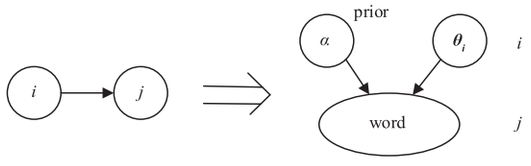


图1 节点交互与主题关系

假定 $i$ 是发送节点(社交网络中被关注节点),$j$ 是接受节点(关注节点). 如果 $j$ 关注 $i$,用 $e_{ij}$ 反映节点间的链接关系,对链接关系进行二值采样. 如果 bellnolli($e_{ij}$) 为 1,说明接受 $j$ 关注节点 $i$ 的兴趣推荐,受到其影响 $j$ 节点的的内容词汇 $w$ 通过 $i$ 节点的主题向量 $\boldsymbol{\theta}_i$ 采样,使用主题 $\boldsymbol{\theta}_i$ 与词汇分布向量 $\boldsymbol{\beta}$ 产生词汇 $w$;否则节点 $j$ 虽然关注了节点 $i$,但仍然坚持自己的兴趣爱好(没受到影响),内容词汇仍然由($\boldsymbol{\theta}_j$,$\boldsymbol{\beta}$)生成. 把这个过程与 LDA 模型结合,经过概率图迭代,得到文档-主题共现矩阵. 接下来针对每个用户节点对($i$,$j$),计算 $\cos(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)=\dfrac{\boldsymbol{\theta}_i\cdot\boldsymbol{\theta}_j}{|\boldsymbol{\theta}_i|\cdot|\boldsymbol{\theta}_j|}$,利用该余弦值作为衡量节点 $i$ 和节点 $j$ 的主题相似性大小. 上述过程即算法的第一阶段. 在第一个阶段的假设下,交互越多的节点,共享主题越多,节点的主题越相似,所以定义 $p(i,j)=\cos(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)$ 作为 $i$ 和 $j$ 在拓扑上的转移概率,这样主题越相似的节点,转移概率越大,转移概率越大的节点间,接下来链接的可能性越大. 然后利用网络的拓扑结构,定义 $r[i]$ 作为节点 $i$ 的影响力:$r[i]=\beta'\dfrac{1}{|N|}+(1-\beta')\sum_j r[i]p(i,j)$,其中 $N$ 是拓扑图上节点数,$\beta'$ 是初始权重. 在整个拓扑图上迭代求解得到影响力计算值.

### 3.2 主题内容相关性建模

考虑在建立模型中,需要给用户节点先天兴趣一个初始分布. 基于 LDA 和 MMSB[12] 的模型先验都假定建立在 Dirichlet 分布基础上,Dirichlet 过程假定数据具有简单的性质,如可交换性或者条件独立等. 但客观世界中的数据往往具有非常复杂的结构关系和依赖关系,比如影视娱乐主题和体育主题的相关性要强于娱乐和教育科研主题的相关性,一个在体育社区有影响力的人相比教育科研主题,更有可能在影视娱乐主题具有一定的影响力. 对于关系数据而言,对象可以承载多个潜在角色或影响其与其它人的关系的集群成员资格,关系数据的"混合成员方法"允许我们描述扮演多个角色的对象之间的交互. 所以在考虑实际数据关系时必须把数据间的依赖性考虑在内,相关主题模型 CTM[13](Corelation Topic Model)由美国卡内基梅隆大学 M. Blei 等人 2005 年提出的模型,是对主题模型进行扩充,用于文本的关联主题挖掘. CTM 不使用 Dirichlet,而是从多元高斯中抽取一个实值的随机向量,然后将其映射到单纯形以获得多项式参数. 本文借鉴相关主题模型中的思想,把主题间的相关性引入到主题社区的发现过程,本文认为将这种"主题社区内关联"作为先验信息并入影响力模型是很重要的. 本文通过贝叶网络的先验把相关性带入到影响力模型,模型能够捕获基于主题影响力计算时的主题间相关性,并在此基础上开发变分推理算法对参数求解. 主题向量要从多元正态分布的映射中来,所以定义 $\boldsymbol{\eta}_d$ 从高斯均值为 $\boldsymbol{\mu}$ 和协方差为 $\boldsymbol{\Sigma}$ 的正态分布采样,向量 $\boldsymbol{\eta}_i\propto N(\boldsymbol{\mu},\boldsymbol{\Sigma})$,对 $\boldsymbol{\eta}_d$ 进行从自然参数到平均参数的映射来产生主题向量 $\boldsymbol{\theta}_i=f(\boldsymbol{\eta}_{i,k})=\dfrac{\exp\{\boldsymbol{\eta}\}}{\sum_k\exp\{\boldsymbol{\eta}_k\}}$,进而方便对 $\boldsymbol{\theta}_i$ 进行多项式采样 $Z\propto\mathrm{mul}(\boldsymbol{\theta})$,然后定义 $\boldsymbol{\beta}$ 为词分布向量,抽样出主题词汇 $w\propto\boldsymbol{\beta}_Z$ 和节点的链接指示 $s_{ij}$(链接从 $i$ 节点发出到节点 $j$ 的成员指示)、$r_{ij}$(链接从 $j$ 节点发出到节点 $i$ 的成员指示). 这样通过相关主题可以模拟节点的潜在属性与每个子群主题分布间的关系,多元正态分布参数由 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 构成,$\boldsymbol{\mu}$ 是 $k$ 维多元高斯均值,$\boldsymbol{\Sigma}$ 是 $k*k$ 协方差,协方差矩阵 $\boldsymbol{\Sigma}$ 一定程度上诱导了社区内容不同主题间的相关性.

### 3.3 ICMNT 模型

社交网络中朋友圈的关注、微博的转发、评论、点赞等都视为网络中节点间存在的交互关系. 模型把交互关系对应到概率图中的节点间存在边的关系,定义 $e_{ij}$ 表示节点的交互,($s_{ij}$,$r_{ji}$)分别为节点间的链接指示,定义交互 $e_{ij}$ 取决于社区成员指示变量 $s_{ij}$ 和 $r_{ij}$ 的交互价值. 交互的价值取决于两个相应的社区 $k$ 和 $l$ 的兼容性. 假定($s_{ij}=k$,$r_{ij}=l$),交互的价值由脚色匹配矩阵 $\boldsymbol{W}_{k,l}$ 决定. 模型中 $\boldsymbol{W}$ 矩阵是节点间的匹配关系矩阵,$\boldsymbol{W}_{k,l}\propto\mathrm{beta}(\lambda_1,\lambda_2)$,假设 $\boldsymbol{W}$ 矩阵一个具体的实现如表 1 所

示,如果节点 $i$ 来自潜在主题社区 2,即 $s_{ij}=2$,节点 $j$ 来自潜在主题社区 3,即 $r_{ji}=3$,那么根据表 1,$e_{ij}$是我们观察到的节点间交互,服从 bernolli($\boldsymbol{W}_{s_{ij},r_{ji}}$)分布.

$$\text{bernolli}(\boldsymbol{W}_{s_{ij},r_{ji}})=\text{bernolli}(\boldsymbol{W}_{2,3})=\text{bernolli}(0.2).$$

**表 1　匹配关系向量**

| | | | |
|---|---|---|---|
| 0.2 | … | 0.91 | 0.1 |
| 0.3 | … | 0.2 | 0.5 |
| … | … | … | … |
| 0.4 | 0.2 | 0.01 | 0.2 |

社区中节点比喻成文档,由隐含特征导致的相关因素决定文档在社区中的主题. 模型由引入相关性的 LDA 建模主题,构建相关主题向量. 模型中交互关系定义为二进制变量 $e_{ij}$,每对交互节点一个. 这些二进制变量分配取决于生成每个构成文档的主题. 由于这种依赖性,文档的内容在统计上与它们之间的交互链接关系相关,利概率图模型学习主题和交互关系的联合后验分布概率,每个节点的混合成员都取决于节点的主题内容以及交互的关系,反过来其成员资格相似的节点将更有可能在模型下交互. 模型在 MMSB 块和 LDA 的基础上,把节点的主题内容和节点的交互结合在一起,引入多元正态分布作为相关性,利用概率图模型统一起来,利用概率关系学习节点的联合后验 pr($e_{ij}$,$w_1$,$w_2$|rest),通过数据训练模型得到特征最优化的主题向量,设计如图 2 所示的概率图模型,本文称为 ICMNT 算法(Influences Computing Model Based on Node Interaction and Topic). 图 2 中 $\boldsymbol{\theta}_j$ 对应的主题词汇 $w_2$,对交互关系 $e_{ij}$进行 bernolli 采样,如果 $e_{ij}=0$,从 $\boldsymbol{\theta}_j$ 采样词汇标志 $Z_{j,k}$,说明节点以一定概率接收主题 $\boldsymbol{\theta}_j$;或者 $e_{ij}=1$,则从 $\boldsymbol{\theta}_i$ 采样,表示节点虽然和其它节点发生交互,但维持原来的主题不变.

图 2 的概率图模型对应似然函数:

$$\text{pr}(e_{ij},w_1,w_2|\lambda_1,\lambda_2,\gamma,\mu,\boldsymbol{\Sigma}) \tag{1}$$

模型包含的隐变量包含:$(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},\boldsymbol{W},Z_{ik},Z_{jk})$,由于先验从正态高斯采样,与多项式分布不共轭,所以利用变分 EM 对模型中的隐变量进行推理.

### 3.4　模型参数的推理

由生成模型得到隐变量在给定观测条件下的最大联合后验分布表示如式(2):

$$\text{pr}(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},\boldsymbol{W},Z_{ik},Z_{jk}|\text{rest}) \tag{2}$$

式(2)没明确的解析解,又因 $\boldsymbol{\theta}$ 是从一个高维高斯中采样的向量,由于多维高斯空间的稀疏性,利用 MCMC 方法不太现实,因此只能利用变分 EM 对参数进行近似求解,变分贝叶斯算法是利用简单的 $q$ 分布去逼近复杂的 $p$ 分布的计算过程. 假设 $p(X,Z)$是联合分布,在相关文献中已证明等式 $\ln(q_i(Z_i))=\text{E}_{i\neq j}[\ln(p(X,Z))]$成立的情况下,可以用对数似然函数的下界 $L$ 来近似式(1). 设:

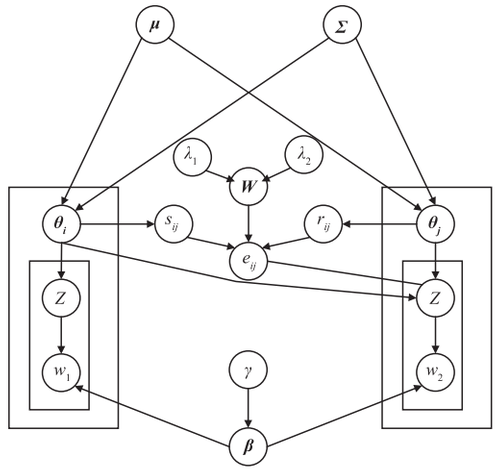


图2　ICMNT概率图模型

$$\begin{aligned}&q(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},\boldsymbol{W},Z_{ik},Z_{jk})\\&=q(W|\rho)*q(\boldsymbol{\beta}|\rho')*q(\boldsymbol{\theta}_i|\lambda,\nu^2)\\&\quad*q(\boldsymbol{\theta}_i|\lambda',\nu'^2)*q(z'(s_{ij},r_{ji})|\delta_{i->j},\delta_{j->i})\\&\quad*q(Z_{ik}|\varphi_{ik})*q(Z_{jk}|\varphi'_{jk})\end{aligned} \tag{3}$$

$q$ 分布初始化为 Gamma( )分布. 利用 $q$ 分布对复杂分布 $p$ 进行解耦,需要进行变分推导近似逼近最大似然的下界 $L$:

$$\begin{aligned}L=&\text{E}[\ln p(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},W,Z_{ik},Z_{jk},e_{ij},w_1,w_2)]\\&-\text{E}[\ln q(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},W,Z_{ik},Z_{jk})]\end{aligned} \tag{4}$$

其中:

$$\begin{aligned}&\text{E}[\ln p(\boldsymbol{\beta},\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,z'(s_{ij},r_{ji})_{\forall 1\leqslant i\leqslant j\leqslant N},W,Z_{ik},Z_{jk},e_{ij},w_1,w_2)]\\&=\text{E}[\ln p(\boldsymbol{\theta}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})]+\text{E}[\ln p(\boldsymbol{\theta}_j|\boldsymbol{\mu},\boldsymbol{\Sigma})]\\&\quad+\text{E}[\ln p(s_{ij}|\boldsymbol{\theta}_i)]+\text{E}[\ln p(r_{ji}|\boldsymbol{\theta}_j)]\\&\quad+\text{E}[\ln p(e_{ij}|(s_{ij},r_{ji}),W)]+\text{E}[\ln p(Z_{ik}|\boldsymbol{\theta}_i)]\\&\quad+\text{E}[\ln p(Z_{jk}|\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)]+\text{E}[\ln p(\boldsymbol{\beta}|\gamma)]\\&\quad+\text{E}[\ln p(w_1|\boldsymbol{\beta},Z_{ik})]+\text{E}[\ln p(w_2|\boldsymbol{\beta},Z_{jk})]\end{aligned} \tag{5}$$

得到似然下界 $L$ 之后,目标就是极大化这个下界,从而得到后验概率分布的近似分布 $q$(E-step),然后固定变分参数,极大化超参数(M-step). 在(E-step)阶段,依次对每个用于解耦的隐变量参数求偏导并令结果等于零,依次对参数的偏导等式求解. 在(M-step)阶段,基于整个训练集中全部节点求和,分别从 $L$ 提取包含参数的项,得到更新后的超参数更新公式.

$$\boldsymbol{\mu}=\frac{1}{D}\sum_{d=1}^{D}\boldsymbol{\mu}_d$$

$$\boldsymbol{\Sigma}=\frac{1}{D}\Big(\sum_{d=1}^{D}\text{diag}(\boldsymbol{\Sigma})^2+\sum_{d=1}^{D}(\boldsymbol{\mu}_d-\boldsymbol{\mu})(\boldsymbol{\mu}_d-\boldsymbol{\mu})^{\text{T}}$$

$$\lambda_1=\lambda_2=\frac{\sum_{ij}(1-e_{ij})*\big(\sum_h\delta_{i->h}\delta_{j->h}\big)}{\sum_{ij}\sum_h\delta_{i->h}\delta_{j->h}}$$

$$\gamma_{ik} \propto \sum_{d=1}^{D}\sum_{n=1}^{N}(\varphi_{dn,i}w^{k}_{1.dn} + \varphi'_{dn,i}w^{k}_{2,dn})$$

其中 $w^{k}_{1,dn}$ 表示 $d$ 文档中的 $n$ 字是词汇表中的 $w^{k}_{1}$ 字. 本文算法的时间开销主要来源于两大部分:变分求解参数和拓扑迭代. 变分求参数又分为 E-step 和 M-step 两个阶段,考虑 ICMNT 算法中的 $D$ 个社区中 $N$ 个节点产生 $N(N-1)/2$ 个交互的节点对,需要对 $N^2$ 数量级的节点对处理,每对节点是关于 $K$ 主题上的更新,主题向量有 $V$ 个词汇,针对每对节点和主题 $K$,变分计算 $\varphi$ 需要时间复杂度 $O(DN^2K)$,计算主题向量 $\boldsymbol{\theta}$ 和词汇向量 $\boldsymbol{\beta}$ 的复杂度共 $O(N^2D+DKV)$,保持数据的预期对数似然的平均变化小于 0.001% 时,我们停止在网络训练集上进行训练,假设每一次迭代收敛平均需要常数 $L$,那么 E-step 阶段时间复杂度 $O(LDN^2+LDN^2K)$. 在隐变量收敛后,M-step 最大化阶段需要对整个数据集全部节点的超参数求最大,通过梯度下降法迭代经过 $T$ 次收敛,需要复杂度 $O(TDK+TDV)$. 得到概率图参数后,根据社区拓扑结构,需要对 $N$ 个节点迭代,利用稀疏矩阵的特征值分解计算方法,每次迭代可以在线性时间 $O(MN)$ 时间完成,$N$ 个节点需要 $O(MN^2)$. 因为 $K \ll V$,所以本文算法的总复杂度为:$O(LDN^2K)+O(TDV)+O(MN^2)$.

## 4 实验结果与实验分析

### 4.1 模型效用评估

#### 4.1.1 模型先验

本实验利用人工合成数据研究模型先验参数的有效性. 实验使用社区节点个数 $n=30$,因此社区间的交互是 30×30 构成的不对称二元矩阵. 设置参数使得 30 个节点被划分为 4 个子组,每个子组分别具有 16、8、4、2 个节点. 生成的合成数据形成一个块对角矩阵,通过设置二元矩阵中某些异常值来体现节点交互时的多属性特点. 对相互作用的节点对采用 5 倍交叉验证,算法中协方差矩阵的和角色匹配矩阵初始向量设置为:

$$\boldsymbol{\Sigma}=\begin{pmatrix}0.91 & 0.05 & 0.1 & 0\\ 0 & 0.92 & 0.2 & 0\\ 0.05 & 0.1 & 0.90 & 0\\ 0.05 & 0.05 & 0 & 0.90\end{pmatrix}$$

$$\boldsymbol{W}=\begin{pmatrix}0.96 & 0 & 0 & 0.05\\ 0 & 0.96 & 0.05 & 0\\ 0 & 0.05 & 0.96 & 0\\ 0.05 & 0.05 & 0.05 & 0.96\end{pmatrix}$$

实验分别选择 MMSB、LDA、EMTM 同本文的 ICMNT 模型在社区内节点间关系的发现上进行对比. LDA 模型的工作假设每个节点都有一个潜在变量来直接指示其社区成员资格,由社区的单一分布决定. 社区的主题分布建立在独立同分布的基础上,然而在许多社交网络环境中,这种表示可能无法很好地捕获节点之间的复杂交互,某个节点可以扮演多个角色身份. 图 3 反应的是节点交互关系的后验预测. 其中图 3($b$)是 LDA 模型在给定的数据基础上得到后验的训练结果,可以看到节点的所有属性来自社区,节点之间交互关系完全由所在的子群决定. EMTM 是一种基于主题模型的影响力计算模型,在潜在特征模型中考虑来自两个节点的所有关联社区,图 3($c$)反映的是 EMTM 算法的结果,可以看到节点具有多个属性,从概率上隶属于多个社区. 不同节点的交互,从概率上考虑节点其所属所有社区间的交互,图 3($c$)可以看到条状的分布图像,因为 EMTM 基于向量间的计算,虽然一定程度上体现了节点和社区间的一对多交互,但是仍然无法捕捉对随机点间的交互. MMSB 和 ICMNT 模型为了促进节点和社区之间的一对多关系,每个节点都有自己的"混合成员分布",然而 MMSB 模型针对于每对节点成员指标对是独立抽样的,一定程度上限制了成员指标的分配. 从图 3($d$)和 3($e$)可以看出,本模型由于在先验中引入了子群之间的相关性,同时在概率图中结合了 MMSB 模型建模交互关系,导致随机点检测和子群间的交互关系的发现上最终都优于 MMSB 模型.

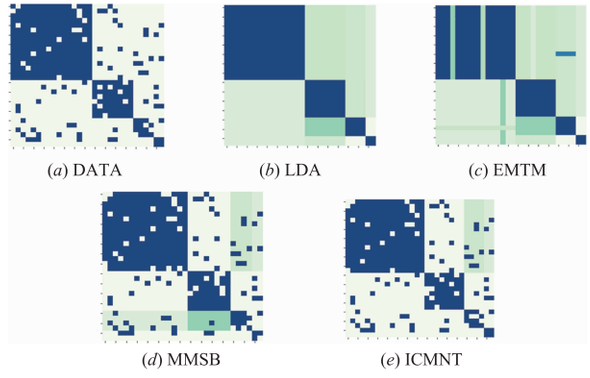


图3 训练模型的交互关系的后验

#### 4.1.2 有效性分析

实验从 stanford 官网下载 Twitter 2009 年 1 月第 1 周的数据,从中取出约 10% 的用户数据量,对于每个公开推文,可以获得以下信息:作者、时间、内容. 通过 Haewoon Kwak 等人提供的数据库获得用户的社交交互关系. 考虑到算法的时间复杂度,为降低问题规模,对关注或者被关注少 5 条的边进行过滤,仅包含具有一定影响值的边. 数据集包含用户数目 17069,推文数量 476553,链接关系 181611 对,用户标签数量 49293 个,从数据集中能够获得关注的内容总和,点评的内容数目. 分别从设定的主题中选择若干个起始种子用户,种子用户的类型包括知名人士、机构及普通用户,对关系网络进行拓展,从数据集中抽取 2000 个 Twitter 用户的数据. 设置主题个数等于 15,PageRank 迭代终止条件 <

0.001%,阻尼系数=0.85,

CM 算法参数设置为:$\alpha_\phi=0.01,\alpha_\gamma=1.0,\alpha_\theta=0.1$

CIM:$\alpha_\phi=0.01,\alpha_\theta=\alpha_\psi=0.1,\alpha_\gamma=1.0$.

ICMNT:$\lambda_1=\lambda_2=0.1,\boldsymbol{\mu}^{\mathrm{T}}=[0.1,0.1\cdots]$. 其中$\boldsymbol{\mu}$是由 15 维向量构成.

实验运行 ICMNT 算法得到推文中主题内的 TOP-5 影响力人物,主题内的关键字和发表该主题的最具影响力的推特人物列表,结果分别如表 2 和表 3 所示. 模型不但能够识别出主题内的影响力人物,还可识别出相应的主题,通过分析主题的关键词可以进一步理解社区.

**表 2　主题内的关键字**

| Topic #1 The slums | Topic #2 AIDS | Topic #3 Football | Topic #4 Rock |
|---|---|---|---|
| violence | world | Player | fashion |
| peace | HIV | Team member | character |
| Civilian | Prevention | Match | program |
| Brazil | Cure | Star | Film |
| … | … | … | … |

**表 3　top-5 影响力人物表**

| Topic#1 | Topic #2 | Topic #3 |
|---|---|---|
| 01. Brail | 01. Codename | 01. Eric |
| 02. Thunder_. . | 02. xeno225 | 02. Joshua |
| 03. mokey | 03. Glosi | 03. Korba1 |
| 04. Chanse | 04. David Heck | 04. Xinzhang |
| 05. Carole A . GOBLE | 05. Rudistudr | 05. PhilippeSmets |
| … | … | … |

利用文献[14]给出的相关性测量标准 $\tau$. $\tau$ 取值范围为[-1,1]. 如果两个列表完全相同,则 $\tau=1$; 如果一个列表与另一个列表相反,则 $\tau=-1$. 对于该范围内的其它值,$\tau$ 值越大意味着两个列表之间的一致性越高. 对上述算法分别得到的推荐结果列表进行分析,得到如表 4 和表 5 的比较结果.

**表 4　相关关系数值大小**

| | #1 | #2 | #3 | #4 |
|---|---|---|---|---|
| ICMNT VS CIM | 0.6810 | 0.6800 | 0.5899 | 0.5790 |
| ICMNT VS PageRank | 0.5534 | 0.5779 | 0.5560 | 0.5440 |
| CIM VS PageRank | 0.4887 | 0.4765 | 0.4600 | 0.4654 |
| CM VS PageRank | 0.4001 | 0.4200 | 0.4089 | 0.4500 |
| ICMNT VS CM | 0.6681 | 0.5800 | 0.6802 | 0.5800 |

**表 5　主题间的相似性**

| LDA | ICMNT | | | |
|---|---|---|---|---|
| | #1 | #2 | #3 | #4 |
| #1 | 0.881 | 0.281 | 0.081 | 0.123 |
| #2 | 0.153 | 0.851 | 0.077 | 0.181 |
| #3 | 0.188 | 0.276 | 0.760 | 0.656 |
| #4 | 0.156 | 0.145 | 0.170 | 0.761 |

CM 和 CIM 是基于主题的模型,PageRank 主要考虑节点的链接结构,ICMNT 同时考虑主题和链接结构. 从实验结果可以看到 $\tau\neq1$ 且 $\tau\geqslant0$,表 4 说明本文算法与其它算法产生的推荐列表存在正相关,但是可以看到 ICMNT 与其它算法相比,比它们相互之间存在更多的相关性,所有算法中 ICMNT 与 CIM 的相关性最高,说明对于本数据集,辅助数据库的精准性和完整性不够,导致主题的内容比节点间的关系贡献了更多信息. 表 5 是主题间影响力人物列表的相似值,从表 5 的结果来看,由于 ICMNT 采用了混合成员分布建模社区关系,相比 LDA 能够发现更准确的基于主题的社区分类.

由于很难对相互之间的转发与评论数目进行准确统计,手动统计 3 个主题中影响力用户转发内容和评论内容,经验表明大概不到 1/10 的粉丝会持续关注,受到最终实际影响,所以生成转移概率时,随机抽取 1/10 的粉丝数和实际转发参与计算. 运行算法分别进行实验对比,表 6 所示为只考虑节点链接关系的 PageRank 与本文的 ICMNT 影响力计算排名在 3 个主题上的对比结果.

**表 6　影响力计算结果**

| Topic | Infulene twitter | ICMNT 排名 | PageRank 排名 | 1/10 粉丝数+转发数 | 消息内容个数 |
|---|---|---|---|---|---|
| #1 | 01 | 2 | 29 | 85 | 17 |
| | 02 | 13 | 25 | 105 | 28 |
| | 03 | 15 | 38 | 212 | 15 |
| | 04 | 19 | 40 | 46 | 17 |
| | 05 | 25 | 46 | 94 | 1 |
| #2 | 01 | 1 | 3 | 1027 | 19 |
| | 02 | 4 | 19 | 405 | 25 |
| | 03 | 7 | 15 | 512 | 17 |
| | 04 | 18 | 20 | 346 | 6 |
| | 05 | 20 | 21 | 294 | 3 |
| #3 | 01 | 14 | 14 | 188 | 7 |
| | 02 | 21 | 34 | 120 | — |
| | 03 | 25 | 45 | 30 | 4 |
| | 04 | 40 | 48 | 77 | 5 |
| | 05 | 45 | 50 | 6 | 1 |

表 6 中的 3 个主题 top-5 节点的影响力排名均在 PageRank 前 50 之内,一定程度上反应了本算法的有效性. 主题 2 内 top-5 节点的排序基本同主题内的关系数成正相关,说明主题 2 是当前比较流行的热点主题,参与主题讨论的人数较多,主题 2 内部节点 01、02 对主题转发的内容较多,其中 02 号节点的转移概率较大,计算出来的 ICMNT 的影响力比 PageRank 的值有明显的提高. 主题 1 内的 top-5 节点在主题内经过计算都具备较

高的主题内转移概率，整体上 top-5 节点在主题社区内相比 PageRank 有所提升，由于节点有比较多的内容输出，但是转发量不够，主题在整个社交数据集中是不太大众的社区讨论主题，社区规模较小，PageRank 中的整体排名比较靠后；可以解释在当前阶段主题还没有扩散开来，一旦在社区中传播开，整体影响力是比较大的. 所以通过 ICMNT 计算的社区主题影响比 PageRank 的排序更能够反应节点的全局影响力. 主题 3 内 top-5 节点由于微博输出内容不够，节点也没有足够的粉丝和转发数，主题 3 在社区内没有形一个比较有影响力的传播中心，相应节点在 PageRank 中计算的值都比较靠后.

### 4.2 模型性能分析

选取 PageRank、MFP、EMTM 同本文 ICMNT 算法在上文的 Twitter 数据集上进行模型的定量评估. 由于用户影响力研究缺少标准数据集，因此本文通过人工判定的结果去评价用户影响力的好坏，实验中由工作人员利用 Twitter 博可视化分析工具的自定义图表功能，结合用户所发全部帖子内容以及用户的活跃度、受关注度和影响力覆盖度 3 个指标来人为判定某用户是否为 TOP-5 节点，并对节点进行标定. 按照四个等级的分数(3、2、1 和 0)被手动标记为名人专长型、专长、一般专长，以及普通用户型. 在传统的推荐算法中，准确率只是考虑了推荐结果中物品个数，而没有考虑物品之间的排序. 本文的影响力排名推荐结果必然是有序的，主题内影响力大的用户排序越靠前越好，因此引入平均准确率 MAP 反应全局性能的综合指标. 选择相应算法与本文算法在 $P@N$、MAP 方面性能进行比较，前 $N$ 个结果中的准确率 P@N：

$$P@N = \frac{\text{前 } N \text{ 个结果中人工判定为影响力人物个数}}{N}$$

其中 AP 定义为平均准确率，MAP 定义在准确率的基础上考虑到影响力排名大小，所有主题所有用户平均准确率取平均值即为 MAP.

$$\text{MAP} = \frac{\sum_{u} \text{AP}(u)}{|U|} \quad \text{AP}(u) = \frac{\sum_{k=1}^{n_u} p@k(u)}{|R|}$$

$R$ 为主题内推荐列表长度，$n_u$ 表示主题 $u$ 内用户 推荐列表中所有排序后的影响力用户. MAP 对于排序位置敏感，其值越大，排序越靠前，算法整体性能越好.

图 4 是比较 $p@5$、$p@10$、$p@15$ 的实验结果. 对用户而言，往往仅仅关注前几个推荐的标签，所以 $N$ 较小时的推荐质量更重要，从图中看到 $p@5$ 效果最好，随着影响力人物候选集的增加，从多个特征维度得到的交集集合元素变少，各算法结果整体出现正确率下降趋势，在 4 组算法的比较中看出 ICMNT 比采用 PageRank 算法的精度分别提高了 10.5%、6.5% 和 5.4%. 从图 5 上看出，利用单一的指标 MAP 来反映模型的综合性能，ICMNT 的平均准确率均高于其他算法.

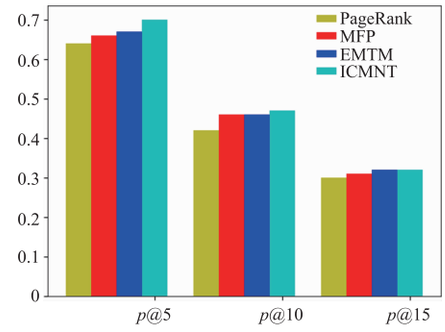


图4 不同算法的$P@N$性能表现

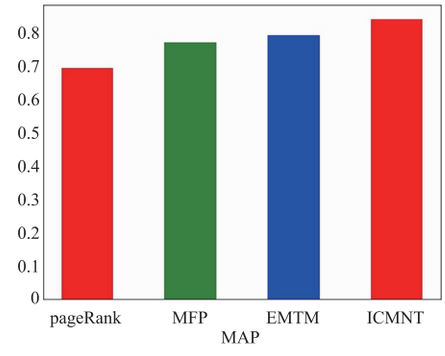


图5 平均准确率

## 5 结论

为进一步提高准确度，本文提出了影响力计算模型 ICMNT. ICMNT 模型利用相关性建模主题间的相互关系，利用节点的交互关系和主题内容重新学习节点的主题向量，设计统一的概率图框架，把节点的内容、交互关系统一起来，并结合社区的拓扑结构，计算节点基于主题的全局影响力，最后通过实验验证了模型的有效性. 进一步的工作需要增加数据量，验证数据在大规模，多样性的数据集上稳定性和精确度方面进行评价. 另外本文考虑的数据是静态的数据分析，未来需要进一步研究在线的，动态的数据演化规律对影响力的影响，并对现有计算算法作出改进，降低模型时间复杂度，找出更有效的适合大规模实用网络的计算算法.

### 参考文献

[1] Kleinberg JM. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[2] Li Q, Zhou T, Lü L, et al. Identifying influential spreaders by weighted LeaderRank[J]. Physica A: Statistical Mechanics and Its Applications, 2014, 404: 47-5.

[3] Barbieri N, Bonchi F, Manco G. Topic-aware social inf-lu-

ence propagation models[J]. Knowledge and information systems,2013,37(3):555 – 584.

[4] Chen W, Lin T, Yang C. Efficient topic-aware influence maximization using preprocessing[J]. The Computing Research Repository,2014,9197:1 – 13.

[5] H Zhao, X Xu, Y Song, et al. Ranking users in social networks with higher-order structures[A]. The Thirty-Second AAAI Conference on Artificial Intelligence [C]. Stroudsburg:ACL,2018. 1073 – 1083.

[6] Weng J, Lim EP, Jiang J, et al. Twitterrank:Finding topic-sensitive influential twitterers[A]. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining [C]. NewYork:ACM,2010. 261 – 270.

[7] Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences [A]. Proceedings of the 24th International Conference on Machine Learning [C]. Stroudsburg:ACL,2007. 233 – 240.

[8] 张腊梅,黄威靖,陈薇,等. EMTM:微博中与主题相关的专家挖掘方法[J]. 计算机研究与发展,2015,52(11):2517 – 2526.
ZHANG La-mei, HUANG Wei-qian, CHEN Wei, et al. EMTM:An expert mining method related to topics inmicroblog[J]. Computer Research and Development,2015,52(11):2517 – 2526. (in Chinese)

[9] 张仰森,郑 佳,唐安杰. 基于多特征融合的微博用户权威度定量评价方法[J]. 电子学报,2017,45(11),2800 – 2809
ZHANG Yang-sen,ZHENG Jia,TANG An-jie. Quantitative evaluation method of microblog user authority based on multi-feature fusion[J]. Acta Electronica Sinica,2017,45(11),2800 – 2809. (in Chinese)

[10] Ding ZY, Zhou B, Jia Y, et al. Topical influence analysis based on the multi-relational network in micr-oblogs[J]. Journal of Computer Research and Develop-ment,2013,50(10):2155 – 217.

[11] 曹玖新,陈高君,吴江林. 基于多维特征分析的社交网络意见领袖挖掘[J]. 电子学报,2016,44(4):898 – 90.
CAOJiu-xin, CHENG Gao-jun, Wu Jiang-lin. Social network opinion leader mining based on multidimensional feature analysis [J]. Acta Electronica Sinica, 2016, 44(4):898 – 90. (in Chinese)

[12] Jonathan Chang, David M Blei . Hierarchical relational models for document networks[J]. The Annals of Applied Statistics ,2010,4(1):124 – 150.

[13] David M Blei, John D Lafferty. Correlated topic models [A]. Proceedings of the 18th International Conference on Neural Information Processing System[C]. Canada:MIT Press,200. 147 – 154 .

[14] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. Paris:ACM,2009. 807 – 816.

## 作者简介

**王大刚**　男,1982 年 10 月出生,安徽肥东人,讲师,2008 年毕业于安徽大学获硕士学位,主要研究方向:数据挖掘,智能信息处理 ,Web 与数据库
E-mail:wdgx621@163.com

**钟　锦**　男,1973 年6 月出生,安徽舒城人,教授,分别于 2004 和 2008 在合肥工业大学和中国科学技术大学获硕士和博士学位,主要研究方向:人工智能,博弈论,智能信息处理

**吴　昊**　男 ,1983 年 5 月出生,安徽舒城人,副教授,研究方向:智能视频处理与分析