基于概率主题模型的社交网络层次化社区发现算法

毕 娟，秦志光

(电子科技大学计算机科学与工程学院 成都 611731)

【摘要】针对传统的社区发现算法大多基于网络拓扑结构寻找独立的社区结构，忽略了用户兴趣属性，并且不能有效地发现社区间的相关性和层次关系等问题。该文提出一种新型的基于PAM(pachinko allocation model)概率主题模型的层次化网络社区发现算法，综合考虑了用户的兴趣和用户的社交网络关系，在同一模型平台上实现层次化的社区结构发现和用户兴趣挖掘，并捕捉和揭示社区之间的关联性和重叠性等特征。模型采用Gibbs采样方法进行参数推导。在真实数据集上的实验结果验证了所提出算法的可行性和有效性。

关 键 词 层次化社区发现； LDA； 概率生成模型； 社交网络

中图分类号 TP391 **文献标志码** A doi:10.3969/j.issn.1001-0548.2014.06.018

Hierarchical Community Discovery for Social Networks Based on Probabilistic Topic Model

BI Juan and QIN Zhi-guang

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The traditional community discovery algorithms are generally based on the link structure of a given social network, they lack of consideration of user's interests and the hierarchical structure of community. In this paper, a novel PAM (Pachinko Allocation Model) probabilistic generative model is proposed to detect latent hierarchical communities based on the user interests and their social relationships. The joint model of topic modeling and community discovery can capture the correlation among multiple communities and their hierarchical structure. Experiments on real-world dataset have confirmed the feasibility and effectiveness of the proposed algorithm.

Key words hierarchical community discovery; LDA; probabilistic generative model ; social network

随着复杂网络研究的日益成熟以及web2.0技术的发展和普及，针对社交网络这一特殊的复杂网络的分析与挖掘的相关研究引起了国内外学者的高度关注。近年来，对社交网络的研究发现诸如国外的Facebook、Twitter，以及国内的微博、人人网等社交网站，其网络规模、用户行为模式、信息交互方式都不尽相同，但这些社交网络中普遍存在着小世界、无标度以及聚集等网络特性。除了这些基本的统计特性之外，社交网络中隐含着丰富的社区结构。如何在大规模的社交网络中挖掘出合适的社区结构，对于理解社交网络的结构和功能有着重要意义。在社会网络分析研究中，大多数研究热点集中于用户社会关系的挖掘，社会网络被表征为一个图，图中节点代表社会网络中的用户，节点之间链接表示社会网络中的关联关系，如朋友关系，Twitter中的关注，信息转发等。这种基于图论的研究方法忽略了大量由用户产生的信息内容，因此综合考虑用户兴趣和用户的社会关系为社交网络分析，特别是对社交网络上的社区发现算法提供了一个新的研究方法。

社区发现是社会网络分析最重要的问题之一，社区的一般定义是同一社区内部节点连接紧密，而社区与社区之间连接疏松[5]。传统的社区发现算法主要有两种：一种基于用户社会关系，利用图论的基本思想划分复杂网络来发现社区，主要包括图划分、层次聚类算法、谱聚类算法等。另一种方法基于用户的兴趣，按照兴趣的不同，将用户划分到不同的群组，从而形成以兴趣为主的主题社区，经典的方法包括了GT(group-topic)模型[9]，以及CUT(community-user-topic)模型[10]等。然而这些方法都有一个基本假设——一个用户只能属于一个社区。而在真实的社会网络中，许多节点通常同时属于多个社区，因此大多数社区都表现出一定的层次性结构，

收稿日期： 2013 – 12 – 21；修回日期： 2014 – 08 – 15
基金项目：国家高技术研究发展计划(2011AA010706)；国家自然科学基金(61133016)
作者简介：毕娟(1979 –)，女，博士生，主要从事数据挖掘、社交网络分析、信息安全等方面的研究.

挖掘隐藏在不同网络中层次化社区结构对进一步了解网络特性和用户行为有着重要的意义。

本文提出新型的基于PAM(pachinko allocation model)概率主题模型的层次化网络社区发现算法，综合考虑了用户的兴趣和用户的社交网络关系，在同一模型平台上实现层次化的社区结构发现和用户兴趣挖掘，并捕捉和揭示社区之间的关联性和重叠性等特征。以往的研究方法限制每个用户属于且仅属于一个社区，而本文提出的新型方法允许一个用户可以加入多个社区并且一个社区内的用户可以参与多个不同主题的讨论。

本文的主要贡献如下：

1) 综合考虑了社交网络上用户的社会关系(链接关系)和用户兴趣(文本信息的主题)，提出一种新型的基于概率生成模型的社交网络社区发现算法。该算法将社区和主题(用户兴趣)视为两个不同的，且相互依赖，互相促进的潜在变量。

2) 算法融入了PAM层次化思想，不仅可以捕捉到用户与用户之间的兴趣与社会关系的关联，同时可以挖掘出不同社区之间的层次关系。

3) 在真实数据集上的实验证明，提出的算法能够实现社交网络分析功能，且能够很好地发现链接紧密、主题明确的社交网络社区。

1　相关工作

1) 经典社区发现算法

基于图论算法的主要思想是首先构造社交网络的图模型，将其化分成若干个子集，要求各个子集之间所连接边数最少。经典的算法有Kernighan-Lin算法(简称KL算法)[3]，和谱平分算法[4]。KL算法为网络引入一个增益值Q，定义为社区内部的边数，表示两个社区之间的边数减去两个社区间的边数，然后基于贪婪算法寻找使Q值最大的划分方法。基于拉普拉斯图特征值的谱平分法是通过分析网络的拉普拉斯算子的特征向量完成社区发现。

基于用户兴趣的社区发现算法以计算用户兴趣的相似度为基准，将兴趣相同或相似的用户分到不同的群组，从而得到以兴趣为中心的社区结构。文献[10]提出了CUT(community-user-topic)模型，该模型考虑了社区与主题之间的联系，社区划分主要以成员的兴趣为标准。这类算法都是将网络划分成了独立无关联的社区结构，忽略了社区间可能存在的层次性关联。

2) 层次化社区发现

主要利用凝聚法或分裂法对网络进行聚类，从而得到社区的层次结构。最经典的方法是文献[5]提出的GN算法，该算法引入边介数来作为节点交互量的度量，首先计算每条边的边介数，除去边介数最大的边，重复这个过程直到网络中不再有边存在。文献[6]采用贪心算法将划分出的社区结果抽象为社区节点，建立社区关系图后进一步优化，从而得到层次性结构。这类算法都将用户严格划分到相应社区，每个用户属于且仅属于一个社区，这种“硬”划分方法不能充分反映真实的网络结构。CTM(correlated topic model)模型[18]是经典LDA主题模型的扩展，最初用于文档聚类问题，用于发现文档中潜在主题之间的相关性。将其扩展到解决社区发现问题，利用Logistic正态分布来提取隐含社区，并提取每对隐含社区之间的关联性。但CTM模型只能描述成对社区之间的相关性，无法揭示多个社区之间的层次性和重叠性特征。

2　新型层次化网络社区发现

给定有向带权网络$G=(U,E,X,W)$，其中网络节点即用户集合为U，网络边集合为E，$e_{ij}\in E$表示用户u_i与用户u_j之间的社会关系，X是边权值集合，$x_{ij}\in X$代表边e_{ij}上的权重，该权重值用来衡量用户之间的关系强度，表示为u_i与u_j之间交互的次数。W代表用户生成的文本信息。比如微博中，x_{ij}表示u_i@u_j或u_i转发u_j微博的次数。因此，一个用户u_i的社会关系(social relationship, SR)可以表征为如下形式：$\mathrm{SR}(u_i)=\{(v_{i1},x_{i1}),(v_{i2},x_{i2}),\cdots,(v_{im_i},x_{im_i})\}$，其中，$v_{ij}\in U$表示$u_i$第$j$个邻居用户。

2.1　算法描述

该算法将传统意义上的社区划分成超社区(super-community)和普通社区(regular- community)，分别由c^s和c^r表示。整个网络结构表征为社区的混合分布，超社区由普通社区混合而成，而普通社区则由所有用户的社交关系和兴趣随机组合。在该层次结构上，第一层为根节点代表整个网络；第二层有s_1个超社区$C^s=\{c_1^s,c_2^s,\cdots,c_{s_1}^s\}$；第三层有$s_2$个普通社区$C^r=\{c_1^r,c_2^r,\cdots,c_{s_2}^r\}$；最底层是用户社会关系节点和主题兴趣。根节点与所有超社区相关联，每个超社区与所有普通社区相关联，捕捉各个普通社区之间的关联关系，而普通社区与所有链接和主题相关联，揭示用户之间的社交关系以及兴趣的相似性。

本文提出的算法可以看作是经典主题模型LDA的扩展，算法的贝叶斯网络图如图1所示，算法中所使用的参数符号如表1所示。Dir(.)表示狄利克雷分布，Mult(.)表示多项式分布。用户兴趣及社区成员身份的生成过程如下所示：

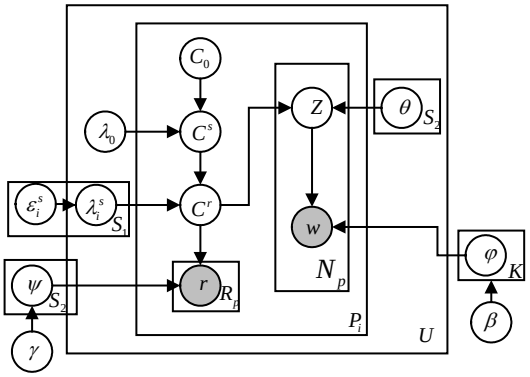

图1 贝叶斯网络图

表1 符号定义说明

符号	定义
K	主题集合 $\{z_1, z_2, \cdots, z_K\}$
C_0	根节点，总主题，主题集合K中的一个特殊主题
C^s	上层社区集合 $\{c_1^s, c_2^s, \cdots, c_{s_1}^s\}$
C^r	普通社区集合 $\{c_1^r, c_2^r, \cdots, c_{s_2}^r\}$
U	用户总数
V	所有特征词构成的词汇表
P_i	用户 u_i 发布的微博集合
R_i	与用户 u_i 有关联的用户集合
w_i	用户 u_i 发布的有单词 $w_i = \{w_{i1}, w_{i1}, \cdots, w_{iN_i}\}$
$\theta_{u,c}$	主题在用户u和社区c上的多项式分布
ψ_c	用户在社区c上的多项式分布
λ_0	上层社区在根节点上的多项式分布
λ_i^s	普通社区在上层社区i上的多项式分
φ_z	单词在主题z上的多项式分布

1) 对每个主题 $z \in \{1,2,\text{L}\ ,K\}$：
生成主题z的词多项式分布 $\varphi_z \sim \text{Dir}(\beta)$

2) 对每个普通社区 c^r， $i \in \{1,2,\text{L}\ ,s_2\}$：
生成该普通社区的链接多项式分布 $\psi_{c_i^r} \sim \text{Dir}(\gamma)$

3) 对每个用户 $u_i \in U$：

① 对根节点 c_0：生成根节点上的超社区多项式分布 $\lambda_0 \sim \text{Dir}(\varepsilon_0)$

② 对每个超社区 c^s， $i \in \{1,2,\text{L}\ ,s_1\}$：生成超社区上对应的普通社区多项式分布 $\lambda^s_{u_i,c_i^s} \sim \text{Dir}(\varepsilon_i^s)$

③ 对每个普通社区 $c^r, i \in \{1,2,\text{L}\ ,s_2\}$：生成该普通社区内部的主题多项式分布 $\theta_{u_i,c_i^r} \sim \text{Dir}(\alpha)$

④ 对用户 u_i 发布的每一条信息：$p \in \{1,2,\text{L}\ ,\ p_i\}$

Ⅰ 选择一个上层社区 $c_p^s \sim \text{Mult}(\lambda_0)$

Ⅱ 选择一个子社区 $c_p^r \sim \text{Mult}(\lambda_{u_i,c_p^s})$

Ⅲ 对每一条链接 $r \in \{1,2,\text{L}\ ,R_p\}$：生成当前信息的接受者 $r_p \sim \text{Mult}(\psi_{c_p^r})$

Ⅳ 对当前信息 p 中的每个单词：$w_j, j \in \{1,2,\text{L}\ ,N_p\}$

选择一个主题 $z_{p,j} \sim \text{Mult}(\theta_{u_i,c_p^r})$，生成当前单词 $w_j \sim \text{Mult}(\varphi_{z_{p,j}})$

该生成算法的图模型如图1所示，首先模型定义一个四层有向层次结构。第一层为根节点， s_1 个超社区位于第二层， s_2 个普通社区分布于第三层，叶子节点由V个单词和K个主题组成。任意的普通社区之间如果存在关联关系，其共同的父亲节点–超社区可以学习并表现这种相关性。

对于每一个用户，生成该用户的每条信息取决于该用户所属的普通社区， λ_{u,c_i^s} 表示用户u属于普通社区i的概率分布，而每条信息的主题，即用户的兴趣则是由该用户及其所属的社区所决定， θ_{u,c_i^r} 表示用户u所属的社区i中各个主题分布情况，通过该参数在同一个模型平台上同时进行主题挖掘和社区发现。

为了模型的推理更为简单，本文给定了超社区数 s_1，普通社区数 s_2 和主题数K，并令W表示文本信息中所有单词的总数，R_i表示社交网络中所有与用户u_i链接的总数。

因此，给定用户、超社区、普通社区、主题、单词和链接的联合概率分布公式如下：

$$P(W,R,U,Z,C^s,C^r,\Lambda,\Theta,\Phi,\Psi \mid \alpha,\beta,\varepsilon,\gamma) \propto$$
$$P(W \mid Z;\Phi)P(Z \mid C^r,U;\Theta)P(R \mid C^r;\Psi)\times$$
$$P(C^r \mid C^s,U,\Lambda^s,)P(C^s \mid C_0;\lambda_0)P(\Theta \mid \alpha)P(\Phi \mid \beta)\times$$
$$P(\Psi \mid \gamma)\prod_{i=1}^{S} P(\lambda_i^S \mid \varepsilon_i^S)P(\lambda_0) =$$
$$\int\prod_{i=1}^{U}\prod_{n=1}^{N} P(w_{i,n} \mid \varphi_{z_{i,n}})\prod_{z=1}^{K} P(\varphi_z \mid \beta)d\Phi \times$$
$$\int\prod_{i=1}^{U}\prod_{n=1}^{N_p} P(z_{i,n} \mid \theta_{i,c_i^r})\prod_{c=1}^{s_2} P(\theta_{i,c} \mid \alpha)d\Theta \times$$

$$\int\prod_{i=1}^{U}\prod_{j=1}^{R_p}P(r_{i,j}\mid\varphi_{c_{i,j}})\prod_{c=1}^{s_2}p(\psi_c\mid\gamma)d\Psi\times$$

$$\int\prod_{i=1}^{s_1}P(\lambda_i^s\mid\varepsilon_i^s)\prod_{p=1}^{P}\sum_{c^s c^r}(P(c_p^s\mid\lambda_0)P(c_p^r\mid\lambda_{c_p^s}^s))d\Lambda$$

2.2　参数估计

文中采用Gibbs采样方法来近似估计模型参数。首先根据已观测变量和其他变量，为每个用户更新超社区和普通社区的成员身份；根据用户的社区成员身份，即该用户所属的社区，更新用户的兴趣主题。具体Gibbs采样算法如下：

$$P(c_p^s=c_i^s,c_p^r=c_j^r\mid c_{-p}^s,c_{-p}^r,U,R,Z)\propto$$

$$P(r,c_p^s,c_p^r\mid c_{-p}^s,c_{-p}^r,R_{-p}U,\mid Z)=$$

$$\frac{n_{-p(0,u)}^{(i)}+\varepsilon_0}{n_{-p(0,u)}^{(\cdot)}+s_1\varepsilon_0}\times\frac{n_{-p(u,i)}^{(j)}+\varepsilon_{i,j}^s}{n_{-p(u,i)}^{(\cdot)}+\sum_{j=1}^{s_2}\varepsilon_{i,j}^s}\times\frac{\prod_{r\in R_p}(n_{-p(u,j)}^{(r)}+\gamma)}{\prod_{i=0}^{R_p-1}\left(n_{-p(u,j)}^{(\cdot)}+i+E\gamma\right)}\times$$

$$\frac{\prod_{z=1}^{K}\Gamma(e_{p,z,u}+n_{-p(c_p^r,u_p)}^{(z)}+\alpha)}{\Gamma\left(\sum_{z=1}^{K}\left(e_{p,z,u}+n_{-p(c_p^r,u_p)}^{(z)}\right)\right)+K\alpha}$$

式中，$n_{-p,(b)}^{(a)}$代表除去当前文档p，实例$a\in A$被分配给或生成于$b\in B$的个(次)数。$n_{-p,(b)}^{(\cdot)}$代表除去当前文档p外，所有实例分配或生成于b的个(次)数。例如$n_{-p,(u,i)}^{(j)}$代表除去当前文档p，普通社区c_j^r出现在用户u所属的超社区c_i^s的次数。$n_{-p,(u,i)}^{(\cdot)}=\sum_{j=1}^{s_2}n_{-p(u,i)}^{(j)}$表示所有普通社区出现在超社区$c_i^s$的次数。$e_{p,z,u}$表示对于文档$p$，由用户$u$生成的主题$z$的次数。

$$P(z_{(p,i)}=z\mid Z_{-(p,i)},C^r,U,W)\propto$$

$$P(w_i=w,z_{(p,i)}=z\mid Z_{-(p,i)},U,W_{-i})\times$$

$$P(z_{(p,i)}=z\mid c_p^r=c,C_{-p}^r,U)=$$

$$\frac{n_{-(p,i)(c,u)}^{(z)}+\alpha}{n_{-(p,i)(c,u)}^{(\cdot)}+K\alpha}\times\frac{n_{-(p,i)z}^{(w)}+\beta}{n_{-(p,i)z}^{(\cdot)}+V\beta}$$

3　实　验

3.1　实验数据集

本文采用的数据集的原始数据来源于Twitter，采用自行开发的爬虫程序自动抓取了2012年7月至2012年10月间发布的数据和用户关系，该数据集收集了3 054名用户所发布的所有微博(tweets)，以及183 675个联系人。文本信息的预处理直接影响主题模型的准确性，通过去除停用(stopwords),非英语字符以及字数少于10的微博等工作，最终，选用的数据集包含了137 633个单词。

3.2　实验结果分析

1) 模型参数设置

Gibbs采样实验中需要确定社区个数和主题个数，实验中设置超社区个数为10个，普通社区个数为60个，主题个数为12个。后续实验将证明该设置结果可使算法达到最优。模型中假设根节点–超社区分布是服从超参数为ε_0的均匀Dirichlet分布，普通社区–用户分布服从超参数为β的均匀Dirichlet分布，分别设置ε_0=0.01，β=0.01。而超参数ε_i^s表示超社区c_i^s中各个普通社区所占比例，从而刻画了各个普通社区之间的关联关系，因此需要对每一个超社区c_i^s($1\leqslant i\leqslant s_1$)的Dirichlet参数$\varepsilon_i^s$进行学习。在每一轮Gibbs采样循环中，按如下算法更新ε_i^s：

$$\mu_{i,j}=\frac{1}{N_i}\times\sum_u\frac{n_{u,i}^{(j)}}{\sum_{j=1}^{s_2}n_{u,i}^{(j)}}$$

$$\sigma_{i,j}=\frac{1}{N_i}\times\sum_u\left(\frac{n_{u,i}^{(j)}}{\sum_{j=1}^{s_2}n_{u,i}^{(j)}}-\mu_{i,j}\right)^2$$

$$m_{i,j}=\frac{\mu_{i,j}\times(1-\mu_{i,j})}{\sigma_{i,j}}-1$$

$$\varepsilon_{i,j}^s=\frac{m_{i,j}}{\exp\left(\frac{\sum_{j=1}^{s_2}\log(m_{i,j})}{s_2-1}\right)}$$

2) 主题与社区挖掘结果分析

图2展示了该算法所挖掘的层次化的社区结构以及与各个社区相关联的主题信息。图中两个圆圈图例分别表示超社区1和超社区2，与其相连的子节点表示包含在该超社区内的普通社区，图中给出了每个普通社区内的出现概率最高的用户信息。最底层的叶子节点表示对应社区内的主题信息以及各个主题中出现概率最高的主题词汇。超社区1将普通社区1和普通社区25关联在一起，两个社区不仅在网络拓扑结构上有重叠用户节点，而且社区之间用户兴趣也有交集。比如用户节点“TechCrunch”以及“Klout”既是普通社区1的成员，同时也与普通社区25内的用户相连，这两个用户作为“桥梁”将两个普通社区联系起来，超社区1成功的捕捉到这一关联性，而传统的社区划分方法无法揭示这种层次

特性。

3) Perplexity结果分析

本文采用Perplexity(困惑度)指标对实验结果进行度量。Perplexity是主题模型最常用的标准度量指标，用来衡量模型对新数据的预测能力。Perplexity值越小，表示模型的生成性能越好。

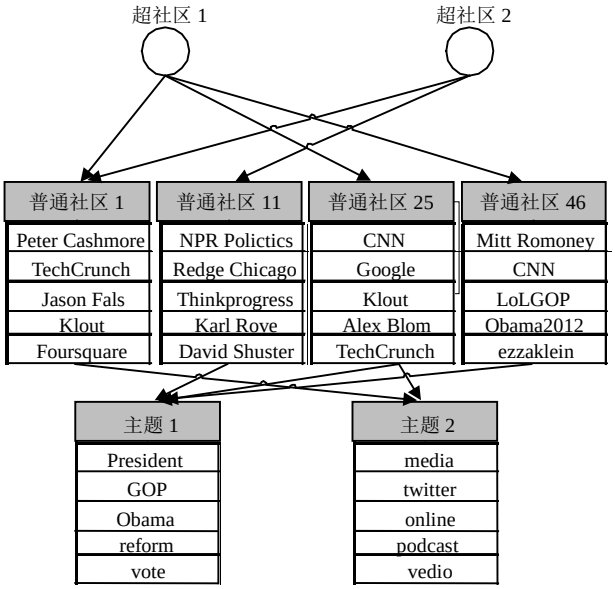

图2 算法挖掘出的社区与主题层次关系结构图

设整个实验数据集大小为 N_{total}，取75%的数据做为训练集，大小为 $N\{p_1, p_2, \cdots, p_N\}$，25%的数据做为测试集，表示为 $p_{N+1}, p_{N+2}, \cdots, p_{N_{\text{total}}}$。Perplexity的计算如下：

$$\text{Perplexity} = \exp\left(-\frac{\log P(p_{N+1}p_{N+2}\cdots p_{N_{\text{total}}} \mid p_1 p_2 \cdots p_N)}{N_{\text{total}} - N}\right)$$

$$P(p_{N+1}p_{N+2}\cdots p_{N_{\text{total}}} \mid p_1 p_2 \cdots p_N) = \prod_{i=N+1}^{N_{\text{total}}} P(p_i \mid p_1 p_2 \cdots p_N) P(p_i \mid p_1 p_2 \cdots p_N) = \prod_{i=1}^{N_{p_i}} \sum_{z \in K} \theta_{u_{p_i}, c_{p_i}^r, z} \times \varphi_{z, w_i} \times \prod_{r \in R_{p_i}} \sum_{c \in C^r} \lambda_{c_{p_i}^s, u_{p_i}, c}^s \times \psi_{c,r}$$

为了验证该算法的有效性，选取了CTM (correlated topic model)模型进行对比试验。图3给出了两个算法在给定不同(普通)社区个数的情况下产生的perplexity值，其中对于本文提出的算法，将超社区个数固定设置为10个。从图中可以看出，对网络进行粗粒度划分时，即(普通)社区个数较小，社区规模较大时，CTM模型的性能较好。但是随着(普通)社区个数增加，对网络进行更加细致化划分时，本文提出的算法性能更优。

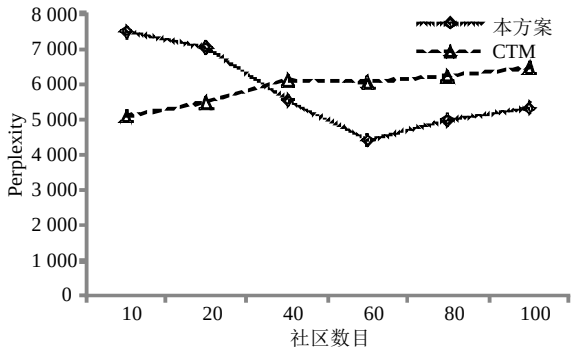

图3 不同社区数目对应的Perplexity性能对比

图4展示了在给定超社区和普通社区个数分别为10和60的情况下，本文算法的perplexity值随着主题个数增加的变化趋势。如图所示，算法在主题个数为12个时，perplexity值达到最小，由此可以确定最优的社区个数和主题个数。

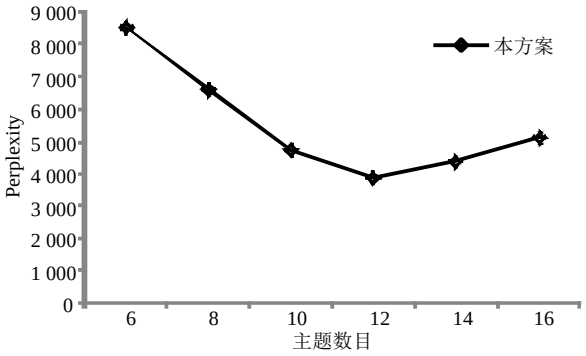

图4 不同主题数目对应的Perplexity性能对比

4 结 论

目前大多数的社区发现算法忽略了真实网络中用户关系和用户兴趣多样性的特点。针对现有的传统社区发现方法的一些缺陷，本文提出了基于PAM概率主题模型的层次化网络社区发现算法，综合考虑了用户的兴趣和用户的社交网络关系，在同一模型平台上实现层次化的社区结构发现和用户兴趣挖掘，并捕捉和揭示社区之间的关联性和重叠性等特征。在真实数据集上的实验结果表明，该算法能够较准确地发现网络中内部联接紧密、语义明确、主题鲜明的潜在社区，并且挖掘出社区之间的层次关联性，更加符合真实网络的社区特性。

未来工作将继续对该模型进行计算效率优化。同时还将着重于研究社区结构和用户兴趣之间随着时间是如何相互影响，共同发展，从而挖掘出社区动态演化特征。

参 考 文 献

[1] WASSERMAN S. Social network analysis: methods and

applications[M]. Oxford: Cambridge university press, 1994, 23-34.

[2] BLEI, DAVID, ANDREW Y, MICHAEL J. Latent dirichlet allocation[J]. The Journal of machine Learning research, 2003, 10(6): 993-1022.

[3] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 11(2): 291-307.

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

[5] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(20): 26-31.

[6] BLONDEL V D,GUILLAUME J L, LAMBIOTTE R. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10(3):108-120.

[7] LI W, MCCALLUM A. Pachinko allocation: DAG-structured mixture models of topic correlations[C]// Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM, 2006: 577-584.

[8] MAO X L, MING Z Y, CHUA T S, et al. SSHLDA: a semi-supervised hierarchical topic model[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics. Pennsylvania: ACM, 2012: 800-809.

[9] WANG X, MOHANTY N, MCCALLUM A. Group and topic discovery from relations and their attributes[J]. Advances in Neural Information Processing System, 2006, 18(1): 14-29.

[10] ZHOU D, MANAVOGLU E, LI J. Probabilistic models for discovering e-communities[C]//Proceedings of the 15th International Conference on World Wide Web. New York: ACM, 2006: 173-182.

[11] ZHANG H, QIU B, GILES C L, et al. An LDA-based community structure discovery approach for large-scale social networks[C]//Intelligence and Security Informatics. New Brunswick: IEEE, 2007: 200-207.

[12] HUANG H, HORNG C Y. Semantic clustering-based community detection in an evolving social network[C]// Genetic and Evolutionary Computing (ICGEC), Sixth International Conference on. Kitakyushu: IEEE, 2012: 91-94.

[13] RAO D, PAUL M, FINK C, et al. Hierarchical Bayesian models for latent attribute detection in social media[C]// Proceeding of the Fifth International Conference on Weblogs and Social Media. Barcelona: AAAI, 2011: 17-21.

[14] CHA Y, CHO J. Social-network analysis using topic models[C]//Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012: 565-574.

[15] ZHAO Z Y, FENG S Z, WANG Q, et al. Topic oriented community detection through social objects and link analysis in social networks[J]. Knowledge-Based Systems, 2012(26): 164-173.

[16] KIM J H, KIM D, KIM S, et al. Modeling topic hierarchies with the recursive Chinese restaurant process[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 783-792.

[17] CHANG J C, BLEI D M. Hierarchical relational models for document networks[J]. The Annals of Applied Statistics, 2010, 4(1): 124-150.

[18] BLEI D M, LAFFERTY J. Correlated topic models of science[J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[19] WANG C, BLEI D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 448-456.

编 辑 蒋 晓