一种面向语义重叠社区发现的 Block 场取样算法

辛 宇 [1] 杨 静 [1] 谢志强 [2]

摘 要 语义社会网络 (Semantic social network, SSN) 是一种包含信息节点及社会关系构成的新型复杂网络. 传统语义社会网络分析算法在进行社区挖掘时, 需要预先设定社区个数且无法发现重叠社区. 针对这一问题, 提出一种面向语义重叠社区发现的 block 场采样算法, 该算法首先以 LDA (Latent dirichlet allocation) 模型为语义分析模型, 建立了以取样节点为核心节点的 block 场 BAT (Block-author-topic) 模型; 其次, 根据节点的语义分析结果, 建立可度量 block 区域的语义凝聚力方法, 实现了语义信息的可度量化; 最后, 以节点的语义凝聚力为输入, 改进了重叠社区发现的标签传播算法 (Label propagation algorithm, LPA) 及可评价语义社区的 SQ 度量模型, 并通过实验分析, 验证了本文算法及 SQ 度量模型的有效性及可行性.

关键词 语义社会网络, 重叠社区, LDA 模型, 社区发现

引用格式 辛宇, 杨静, 谢志强. 一种面向语义重叠社区发现的 Block 场取样算法. 自动化学报, 2015, **41**(2): 362−375

DOI 10.16383/j.aas.2015.c140136

An Overlapping Community Structure Detecting Algorithm in Semantic Social Network Based on Block Field

XIN Yu[1] YANG Jing[1] XIE Zhi-Qiang[2]

Abstract The semantic social network (SSN) is a new kind of complex networks consisting of the node content and topological relationship. The traditional community detection algorithms need to preset the number of the communities and could not detect the overlapping communities. To solve this problem, an overlapping community structure detecting algorithm in semantic social network based on the block field is proposed. Firstly, it takes the latent dirichlet allocation (LDA) model as the semantic analyzing model, establishing the block-author-topic (BAT) model with the sampling node as the core node. Secondly, it suggests the measurement of the semantic cohesion of the block field, depending on the analysis of SSN, to achieve the evaluation of semantic information. Finally, it improves the label propagation algorithm (LPA) which could detect the overlapping communities, with the semantic cohesion as input, and designs the SQ measurement modularity for semantic measuring. The efficiency and feasibility of the algorithm and the semantic modularity are verified via experimental analysis.

Key words Semantic social network, overlapping community, latent dirichlet allocation (LDA), community detection

Citation Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping community structure detecting algorithm in semantic social network based on block field. *Acta Automatica Sinica*, 2015, **41**(2): 362−375

随着网络通信的发展, 电子社交网络, 如 Facebook、Twitter 等, 已成为人们日常生活中不可分割的社交渠道. 为丰富用户的 web 社区生活, 各社交网站推出了 “社区推荐” 及 “好友圈” 服务. 由此而生的社区划分及社区推荐算法, 已成为社会网络数据挖掘研究的热点. 从社区划分算法的研究内容方面, 可分为三个阶段: 硬社区划分、重叠社区划分及语义社区划分. 其中硬社区划分和重叠社区划分研究的出发点在于根据社会网络中节点的关系属性划分关系紧密 “社交群落”, 该领域早期的研究为硬社区划分, 即将社会网络拆分为若干个不相交的网络[1]. 代表算法如 GN[2]、FN[3] 算法. 随着社会网络应用的发展, 社区结构开始出现彼此包含的关系, 为此, Palla 等提出了具有重叠 (Overlapping) 特性的社区结构, 并设计了面向重叠社区发现的 CPM 算法[4]. 此后, 许多经典算法孕育而生, 如 EAGLE[5]、LFM[6]、COPRA[7]、UEOC[8]、蚁群算法[9]、拓扑势算法[10] 等. 近几年来, 社区划分又派生出新的方法, 如遗传算法[11−12]、广义网络社区挖掘算法[13] 等.

在语义社区划分方面, 其研究的出发点在于根据社会网络中节点语义信息内容 (如微博、社会标签

收稿日期 2014-03-14 录用日期 2014-08-12
Manuscript received March 14, 2014; accepted August 12, 2014
国家自然科学基金 (61370083, 61073043, 61073041, 61370086), 国家教育部博士点基金 (20112304110011, 20122304110012) 资助
Supported by National Natural Science Foundation of China (61370083, 61073043, 61073041, 61370086) and National Research Foundation for the Doctoral Program of Higher Education of China (20112304110011, 20122304110012)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001 2. 哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080
1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001 2. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080

等), 将具有相似信息内容的节点划为同一社区.

由于所划分的社区结构基于信息相似性, 其划分结果更能体现社区的凝聚性. 由于语义信息需要以文本分析为基础, 因此目前的语义社区划分算法大多以 LDA (Latent dirichlet allocation) 模型[14] 作为语义处理的核心模型. 根据 LDA 模型的应用方式算法可分为三类.

1) 关系语义信息的 LDA 分析, 此类算法以网络拓扑结构作为语义对象, 利用改进的 LDA 模型分析节点的语义相似性, 将 LDA 分析结果作为社区推荐及社区划分参数. Zhang 等提出了 SSN-LDA 算法, 将节点编号及关系作为语义信息内容, 将节点的关系相似性作为训练结果[15]. Henderson 等在 SSN-LDA 模型的基础上融入了 IRM (Infinite relational models)[16] 模型, 提出了 LDA-G 算法, 该算法有效地将 LDA 与图模型相结合, 在社区发现的基础上可进行社区间的链接预测[17]. 随后 Henderson 等在 LDA-G 的基础上加入了节点多元属性分析, 提出了 HCDF 算法, 增加了社区发现结果的稳定性[18]. Zhang 等也在 SSN-LDA 算法的基础上提了面向有权网络的 GWN-LDA 算法[19] 及面向层次划分的 HSN-PAM[20] 算法. 此类算法的优点是: 结构模型简单, 需要的信息量较少, 适合处理大规模数据. 缺点是: 此类算法所利用的语义信息并非文本信息, 所挖掘的社区不具有文本内容相关性, 属于利用语义分析的方法进行关系社区划分.

2) 关系 – 话题语义信息的 LDA 分析, 此类算法以节点的文本信息作为语义对象, 将相邻节点的文本信息作为先验信息, 使得 LDA 分析的语义相似性接近现实. 此类算法均以 AT 模型[21] 作为 LDA 分析的基本模型, 代表算法有 McCallum 等提出的 ART 模型, 该模型在 AT 模型的基础上加入了 recipient 关系采样, 将 AT 模型引入了语义社会网络分析领域[22]. 随后 McCallum 等在 ART 模型的基础上加入了角色分析过程, 提出了 RART 模型, 扩展了 ART 模型的在社会计算领域的应用[23]. Zhou 等在 AT 模型中加入了 user 分布取样, 提出了 CUT 模型[24]. Cha 等根据社交网络中跟帖人的 topic 信息抽取出树状关系模型, 并利用层次 LDA 算法对树状关系模型中的文本信息进行建模, 提出了 HLDA 语义社会网络分析模型, 该模型可有效处理论坛类 (非熟人关系) 网站的用户分类问题[25]. 此类算法的优点是: 在节点关系基础上结合了文本信息分析, 其划分的社区具有较高的内部相似性. 缺点是: 仅在文本取样时考虑了网络的关系特性, 缺少对网络局部社区特性的考虑, 使得划分的社区结果中出现不连通的现象.

3) 社区 – 话题语义信息的 LDA 分析, 此类算法在关系 – 话题类算法的基础上加入了社区因素, 将 LDA 模型从邻接关系取样转向了局部区域取样, 有效避免了关系 – 话题类算法的局部区域不连通现象, 是成熟化的语义社区划分算法. 代表算法有 Wang 等提出的 GT 模型, 该模型是 ART 模型的扩展, 将 group 取样替代了 ART 模型的 recipient 取样[26]. 随后 Pathak 等论述了 recipient 取样的必要性, 并在 ART 模型的基础上加入了 community 取样, 提出了 CART 模型[27]. 近些年来, 话题 – 社区的关系成为 LDA 模型研究的重点, Mei 等将社区话题分布与社区模块度相结合, 提出了 TMN 模型并建立了话题 – 社区关系函数, 以指导社区的优化过程[28]. Sachan 等和 Yin 等分别从话题 – 社区分布和社区 – 话题分布角度, 在社区与话题间构建关联, 并将其引入了 LDA 模型, 分别提出了 TURCM[29−30] 及 LCTA 模型[31], 在增加社区划分结果的话题差异性的同时, 增加了社区划分结果的合理性. 此类算法的优点是: 语义社区划分准确性高. 缺点是: 模型复杂容易产生过拟合的现象, 由于 LDA 模型需要预先确定先验参数的维数, 因此, 所划分的社区个数需要预先设定, 且不同的预设社区个数所产生的社区划分结果差异较大.

上述算法的共同缺点在于需要预先设定社区个数且无法发现重叠社区, 由于语义社会网络是语义网络和社会网络的结合体, 是由信息节点及社会关系构成的新型复杂网络, 其宏观概念上具有社会网络的链接关系属性, 微观上每个节点具有语义信息属性. 因此, 语义社会网络的语义社区发现算法需要兼顾两方面条件: 1) 语义社区内部链接关系紧密; 2) 语义社区内部节点的语义信息相似度高. 为避免社区 – 话题 LDA 分析中预设社区个数的问题, 本文所设计的面向语义社会网络重叠社区发现算法, 建立了以节点为核心的 block 场取样方法, 以局部结构的启发式发现形式进行社区探测, 无需设定社区个数; 为实现重叠社区划分, 将语义量化分析结果作为标签传播算法的权重输入, 将标签归属不确定的节点作为重叠节点, 实现重叠社区的划分; 为度量语义重叠社区划分结果, 建立了可度量语义凝聚力的方法及评价语义社区划分结果的 SQ 度量模型. 由于社区发现算法以标签传播为模型仅以节点语义凝聚力作为输入, 避免了社区个数的预设问题. 最后通过实验, 分析了本文算法的参数选取及 SQ 度量模型的评价性能.

1 Block-author-topic 关系建模

1.1 LDA 关系表示

对于有代表性半监督语义社区发现算法, 如

AT、ART 及 HLDA 分别为 “点、面、放射” 的形式对语义网络中的节点进行文本取样, 其文本话题生成过程 (以节点 G_i, G_j 为例) 的差别如图 1 所示. 其中图 1 (a) 为 AT 模型的文本生成过程, 在全局文本分析时分别对节点 G_i, G_j 进行单独取样, 不考虑网络拓扑的相关性, 其文本生成过程是以 author 为单位; 图 1 (b) 为 ART 模型的文本生成过程, 由于 ART 模型以 recipient 作为 author 的桥梁, 在对节点 G_i 进行文本生成取样时, 加入了 G_i 相邻节点 G_1, G_2, G_j 的取样, 并在对节点 G_i 进行文本生成时, 加入了 G_j 相邻节点 G_3, G_4, G_5, G_i 的取样, 其文本生成过程是以 author 的中心区域为单位; 图 1 (c) 为 HLDA 模型的文本生成过程, 由于 HLDA 模型以分层的方式进行文本生成取样, 在对节点 G_j 进行文本生成取样时, 将与节点 G_j 直接相邻的节点 G_3, G_4, G_5, G_i 进行 2 次取样, 与节点 G_j 间接相邻的节点 G_3, G_4, G_5 进行 3 次取样, 其文本取样过程是以 author 的放射区域为单位.

ART 模型和 HLDA 模型是 AT 在社会网络中的应用, ART 模型的文本生成取样区域半径为 1, 导致以 author 为中心的区域规模较小, 文本生成取样结果仅代表直接邻接关系, 不具有社区的规模特性; HLDA 模型的文本生成取样过程是对放射区域的无权取样, 忽略了社会网络中距离对社区成员间的影响. 由于实现的社会网络中, 某一话题在传播过程中会产生信息的衰减, 即话题内容随着传播距离的增加而减少, 因此对话题的取样应以 author 为中心的有权关系区域 (block) 作为取样区域, 其中 block 中心位置权重应大于边缘权重. 为此, 本文设计了 BAT (Block-author-topic) 文本取样模型. 图 1 (d) 为 BAT 模型的文本生成取样过程, 在进行文本生成取样时以节点 G_j 作为取样中心, 与节点 G_j 直接相邻 (距离为 1) 的节点 G_i, G_3, G_4, G_5 作为 1-dis 取样点, 与节点 G_j 距离为 2 的节点 G_1, G_2 作为 2-dis 取样点, 并依据取样点与取样中心的距离分配取样权重. 不同于 HLDA, BAT 对取样的距离进行了加权, 其权重随距离的增加而减小, 从而将取样范围收敛为 block 区域. 因此, BAT 文本生成过程是以节点 G_j 为中心的全局区域 (block 区域) 为单位.

语义社会网络的语义信息体现为各节点的文本信息内容, 每个节点具有节点内部的局部语义信息, 各节点的信息集合构成网络总体语义信息. 本节对语义社会网络中的局部语义信息和总体语义信息的 BAT 建模过程进行描述, 所涉及到的数学符号如下:

1) G 表示全局网络, G_i 表示网络中的节点 i;

2) B_i 表示以节点 G_i 为中心的 block 区域;

3) x 表示集合 B_i 中抽取的一个元素;

4) N 表示语义社会网络中的关键字个数, N_i 表示节点 G_i 的关键字个数;

5) D 表示语义社会网络中的语料信息个数;

6) $\boldsymbol{w}$ 表示关键字的向量, w_i 为向量 $\boldsymbol{w}$ 中第 i 个关键字所对应的编号;

7) $\boldsymbol{z}$ 表示与关键字的向量 $\boldsymbol{w}$ 对应的话题编号向量, z_i 表示 w_i 所隶属的话题编号;

8) T 表示话题个数;

9) θ 表示话题分布概率;

10) ϕ 关键字的分布;

11) α 表示各节点的话题分布先验参数;

12) β 表示某一话题内部, 关键字分布的先验参数.

图 2 为 AT、HLDA、ART 及 BAT 模型的对比关系图.

语义社会网络的结构可看作节点间的关系拓扑, 节点间的作用力随距离的增加而减弱, 且语义社会网络中节点间语义信息借由邻接关系进行传播, 与 G_i 距离越远则与 G_i 的语义相关性越弱. 为描述节点间的作用力 (语义相关性), 可利用高斯数据场模型将节点间的作用力进行场势函数建模, block 区域 B_i 内节点 G_i 与其他节点的作用关系可用 block 场进行表达, 则节点 G_i 与 G_j 的作用力 (语义相关性) 可描述如下:

$$I_{i,j} = \exp\left(-\left(\frac{\mathrm{dis}\,(G_i, G_j)}{\sigma}\right)^2\right) \tag{1}$$

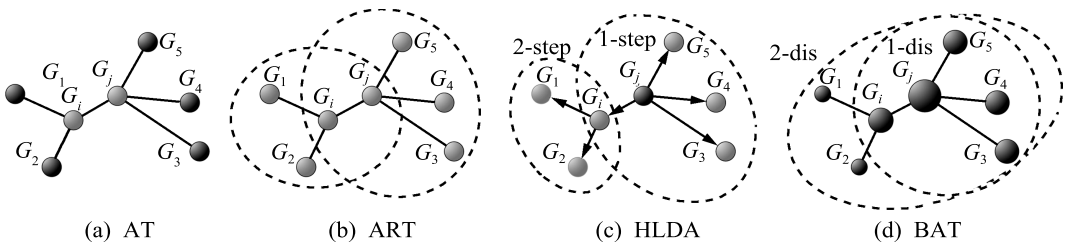


图 1 文本取样生成过程

Fig. 1 The sample proceeding of context

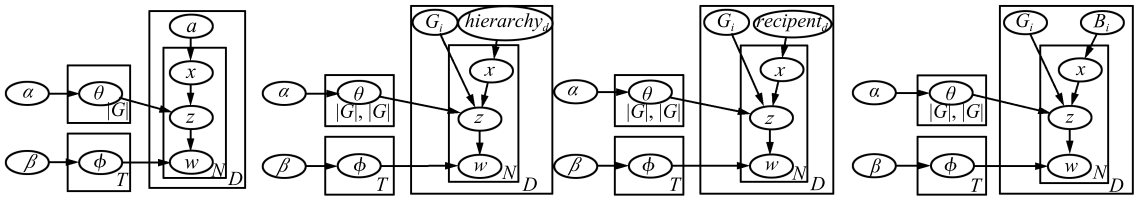


图 2　LDA 模型对比关系图

Fig. 2　The comparison of LDA models

其中, $\mathrm{dis}(G_i, G_j)$ 为节点 G_i 与节点 G_j 间的拓扑距离 (跳数), σ 为距离影响因子, 用于控制节点的作用力影响范围. 根据高斯函数的数学性质, 每个节点的影响范围近似为 $3\sigma/\sqrt{2}$ 跳的局部区域, 根据文献 [10] 可知 σ 的最优取值区间为 (1, 2). 因此根据式 (1) 可知节点 G_i 作用力的实际范围约为 3 跳内的局部区域. 语义相关性体现了节点间文本信息的相对近似性. 因此, 为提高文本信息取样的准确性, 在对以节点 G_i 为中心的 block 区域 B_i 进行随机扩展 author 取样时, 可按语义相关性 $I_{i,j}$ 比例作为扩展 author 选择的概率. 据此, author 的概率密度函数可表达为

$$\mathrm{Field}\left(x|B_i\right)=\frac{\exp\left(-\left(\frac{\mathrm{dis}\left(G_i,x\right)}{\sigma}\right)^2\right)}{\sum\limits_{j=1}^{|G|}\exp\left(-\left(\frac{\mathrm{dis}\left(G_i,G_j\right)}{\sigma}\right)^2\right)} \tag{2}$$

在 block 场中, B_i 的 BAT 模型的概率生成关系式如下:

1) $x|B_i\sim\mathrm{Field}\ (x|B_i)$: 表示从节点 G_i 为中心的 block 区域中, 按 block 场的取样概率函数, 选择一个元素 x 作为 G_i 的扩展 author;

2) $z|\theta\sim\mathrm{Multinomial}\ (\alpha)$: 表示从给定的 G_i 及选定的扩展 author 文本信息中, 抽取一个话题, 该话题服从以 θ 为参数的多项式分布;

3) $\theta|\alpha\sim\mathrm{Dirichlet}\ (\alpha)$: 表示 θ 服从以 α 为参数的狄利克雷分布;

4) $w|\phi\sim\mathrm{Multinomial}\ (\phi)$: 表示话题中的关键字 w 服从以 ϕ 参数的多项式分布;

5) $\phi|\beta\sim\mathrm{Dirichlet}\ (\beta)$: 表示 ϕ 服从以 β 为参数的狄利克雷分布.

$$\begin{aligned}&p\left(\theta,\varphi,x,z,w|\alpha,\beta,G_i,B_i\right)=\\&\quad\prod_{i=1}^{|G|}\prod_{j=1}^{|G|}p\left(\theta_{ij}|\alpha\right)\prod_{t=1}^{T}p\left(\varphi_t|\beta\right)\times\\&\quad\prod_{d=1}^{D}\prod_{n=1}^{N}\left(\mathrm{Field}\left(x_{dn}|B_i\right)p\left(z_{dn}|\theta_{B_i,x}\right)p\left(w_{dn}|\varphi_{dn}\right)\right)\end{aligned} \tag{3}$$

其中, x_{dn} 表示第 d 个语料信息中, 第 n 个关键字所隶属的 author 号; z_{dn} 表示 G_i 在第 d 个语料信息中, 第 n 个关键字所隶属的话题号; w_{dn} 表示 G_i 在第 d 个语料信息中, 第 n 个关键字在语义字典中的编号. $\theta_{Bi,x}$ 和 ϕ_{dn} 分别表示生成 d 个语料信息的第 n 个关键字时, 话题 z_{dn} 及字键字 w_{dn} 的出现概率. 对式 (3) 变量 θ 和 ϕ 求积分及对 x, z 求和, 得出 BAT 模型的 w 生成式 (4).

$$\begin{aligned}&p\left(w|\alpha,\beta,G_i,B_i\right)=\\&\quad\iint\prod_{i=1}^{|G|}\prod_{j=1}^{|G|}p\left(\theta_{ij}|\alpha\right)\prod_{t=1}^{T}p\left(\varphi_t|\beta\right)\times\\&\quad\prod_{d=1}^{D}\prod_{n=1}^{N}\sum_{x_{dn}=1}^{|G|}\left(\mathrm{Field}\left(x_{dn}|B_i\right)\right)\times\\&\quad\sum_{z_{dn}=1}^{T}\left(p\left(z_{dn}|\theta_{B_i,x}\right)p\left(w_{dn}|\varphi_{dn}\right)\right)\mathrm{d}\varphi\mathrm{d}\theta\end{aligned} \tag{4}$$

经文献 [23] 的推导过程可知:

$$\begin{aligned}&P\left(x_{dn},z_{dn}|x_{-dn},z_{-dn},w,\alpha,\beta,G_i,B_i\right)\propto\\&\quad\frac{\alpha_{z_{dn}}+n_{i,x_{dn},z_{dn}}-1}{\sum\limits_{t=1}^{T}\left(\alpha_t+n_{i,x_{dn},t}\right)-1}\times\frac{\beta_{w_{dn}}+m_{z_{dn},x_{dn}}-1}{\sum\limits_{v=1}^{V}\left(\beta_v+m_{z_{dn},v}\right)-1}\end{aligned} \tag{5}$$

先验参数的估计为

$$\begin{aligned}\hat{\theta}_{ijz}&=\frac{\alpha_z+n_{i,x,z}}{\sum\limits_{t=1}^{T}\left(\alpha_t+n_{i,x,t}\right)}\\\hat{\varphi}_{tw}&=\frac{\beta_w+m_{t,w}}{\sum\limits_{v=1}^{V}\left(\beta_v+m_{t,v}\right)}\end{aligned} \tag{6}$$

其中, V 表示语义字典中关键字的个数; $n_{i,x,t}$ 表示从 G_i 的 block 区域 B_i 中随机选取了节点 x, 并在节点 x 的关键字中属于话题 t 的个数; $m_{t,v}$ 表示关键字 v 属于话题 t 的个数. 其中 θ_{ijz} 表示 G_i 与 author G_j 的混合文本中, 话题 z 出现的概率. Gibbs 取样过程如下.

算法 1. Gibbs 取样算法

initialize the node and topic assignments at random repeat

```
for d = 1 to D
  for i = 1 to N
    draw x_dn and z_dn from
      P(x_di, z_di | x_-di, z_-di, w, α, β, F)
    update n_{i,x_dn,z_dn} and m_{z_dn,w_dn}
  end for
end for
```

until the Markov chain reaches its equilibrium compute the posterior estimates of θ and ϕ.

2 语义凝聚力的量化映射

本文以 BAT 文本分析结果作为语义社会网络量化分析的输入. 为避免 LDA 语义分析模型中需要预先设定社区个数的问题, 需要在语义网络中建立节点的量化关系, 并依据量化关系建立启发式社区发现算法. 为此, 本文根据 BAT 模型的结果, 建立了网络链接的语义凝聚力度量.

从 BAT 模型的分析可知, 对 BAT 模型进行 Gibbs 迭代取样后, 可根据 Gibbs 取样过程计算出三维话题概率分布 θ_{iat}, 其中 i 表示节点维度, 共同有 $|G|$ 个不同元素; a 表示与节点 G_i 直接相邻的 author 维度, 共同有 $|G|$ 个不同元素; t 表示话题维度, 包含了 T (话题个数) 维向量, 根据第 1 章的 BAT 建模分析可知, 该 T 维向量表达了节点 G_i 的话题隶属度. 由于 BAT 的文本生成过程是以节点 G_i 为核心的 block 区域为单位, 三维话题概率分布 θ_{iat} 中 author 维度可作为节点维度的从属 (如图 1 (d) 中 G_j 与 G_i 及 $G_1 \sim G_5$ 的关系). 因此, 可以以节点维度作为 θ_{iat} 的主属性对 author 维度进行加和, 以消除 author 维度, 从而将三维话题概率分布 θ_{iat} 转化为二维 θ'_{it}. 即 $\theta'_{it} = \sum_{a=1}^{|G|} \theta_{iat}$. 二维话题概率分布矩阵 θ' 中的 i 行元素 θ'_i, 可看作以节点 G_i 为核心的 B_i 区域在语义社会网络 T 维话题空间中的坐标 m_i. 为衡量各 B_i 的语义凝聚力, 本文利用 PCA 主成分加权法, 将矩阵 θ' 各行向量的相关矩阵的 T 个特征值所构成的向量 $\mathbf{\Lambda} = \{\lambda_1, \cdots, \lambda_T\}$, 作为话题隶属度权重向量, 将 $\theta_{i,\cdot}$ (即 $\boldsymbol{m}_i$) 与 $\mathbf{\Lambda}$ 的内积作为 B_i 的语义凝聚力 W_i.

根据第 1 节的分析可知, 对扩展 author 的随机选择及 G_i 的混合文本取样过程与 G_i 的度数相关性较小, G_i 的话题隶属度更大程度取决于 G_i 与邻近节点的文本相似程度. 因此经 PCA 主成分加权法计算后的语义凝聚力 W_i 可代表区域 B_i 的紧致性. 根据式 (1) 可知节点 G_i 作用力的实际范围约为 3 跳内的局部区域, 因此, W_i 实际代表了半径为 3 的区域紧致性.

本文以清华大学 ArnetMiner 系统 QLSP (Quantifying link semantics-publication) 数据集的部分数据为例 (其中包含 108 篇论文 155 条引用关系). 本文算法分别在每篇论文的摘要中抽取 6 个关键字作为论文节点的语义信息, 以话题个数 T 为 5 进行 Gibbs 取样迭代 (所提取的话题如表 1 所示), 并利用 PCA 加权法计算各区域 B_i 的语义凝聚力, 其语义凝聚力拓扑图如图 3 所示, 其中节点的大小表达了语义凝聚力的大小.

表 1 QLSP 数据集的话题分组

Table 1 The topic groups of QLSP dataset

Topic	Group 1	Group 2	Group 3	Group 4	Group 5
word	protocol	words	graph	image	unify
	route	vocabulary	structure	vector	classifier
	topology	model	analysis	retrieval	semantic
	model	context	theory	mixture	matrix
	signal	candidate	engineering	detection	measure

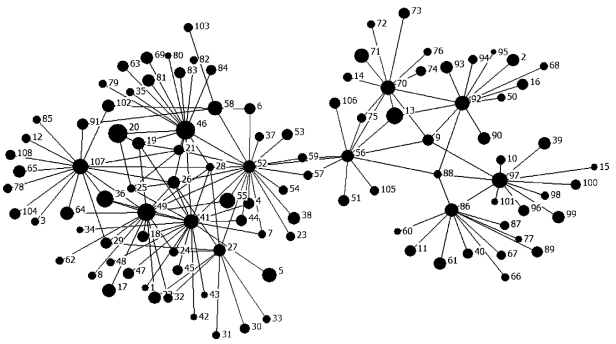


图 3 QLSP 数据集的语义凝聚力拓扑图

Fig. 3 The topology of the semantic cohesion of QLSP

3 基于语义凝聚力的标签传播算法

本文以 BAT 量化分析结果结合标签传播算法 (Label propagation algorithm, LPA), 设计了语义重叠社区发现的 BAT-LPA 算法. 根据 BAT 量化分析可知, 语义凝聚力 W_i 的大小表现了以 G_i 为核心的 B_i 区域的语义凝聚力大小. 为此, 可将语义凝聚力较大的 B_i 区域作为社区的局部块, 并可将彼此关系紧密的局部块聚合为社区, 从而实现社区发现. 文献 [10] 证明了 σ 的最优取值区间为 (1, 2). 根据

式 (1) σ 在区间 (1, 2) 内, 当节点距离 $dis(G_i, G_j) \in \{1, 2, 3\}$ 时, 节点 G_i 和 G_j 间的作用力占 $dis(G_i, G_j) \in \{1, 2, \cdots, \infty\}$ 的比率如图 4 所示. 其中说明了与 G_i 距离为 1 的节点是 G_i 的主要作用节点, 且距离为 3 时 G_i 与 G_i 间的作用力可忽略.

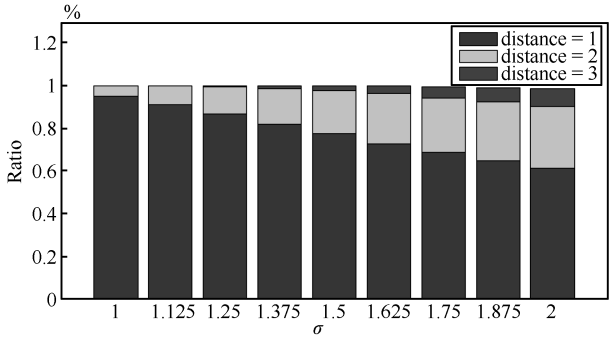


图 4　σ 对 distance 的影响

Fig. 4　The influence of σ to distance

LPA 是社区发现的经典算法, 文献 [32] 提出了 LPA 的基本框架 SLPA, 即节点间通过所建立的传播准则 (Speaking rule)、接收准则 (Listening rule, LR) 和终止准则 (Stop criterion, SC) 进行循环异步标签传播及标签更新, 当循环终止时, 将具有同一标签的节点划分为同一社区. 为实现语义社会网络的标签传播算法, 需要改进 LPA 算法的传播准则、接收准则和终止准则.

3.1　传播准则

网络中每个节点包含一个 $paris(c, b)$ 集合, 其中 c 为标签, b 为标签隶属系数 (Belonging coefficient, BC), 函数 $b_t(c, G_i)$ 表示 t 时刻节点 G_i 中标签 c 的标签隶属系数取值, 初始时节点 G_i 的集合 $paris_i(c, b) = \text{null}$. 当某一节点 G_i 作为传播节点时, 若 $paris_i(c, b)$ 为 null 则令 $paris_i(c, b) = (i, 1)$, 否则保持 $paris_i(c, b)$ 不变; 再将 $paris_i(c, b)$ 与扩散系数 E 相乘, 向与 G_i 距离为 1 及距离为 2 的节点传播. 由于标签传播与文本取样同样受距离的衰减影响, 为此本文根据式 (2) 所示的取样概率模型, 建立式 (7) 所示的扩散系数 E 的表达式.

$$E = \begin{cases} \dfrac{\exp\left(-\left(\dfrac{1}{\sigma}\right)^2\right)}{\displaystyle\sum_{i=1}^{\infty}\exp\left(-\left(\dfrac{i}{\sigma}\right)^2\right)}, & \text{distance} = 1 \\ \dfrac{\exp\left(-\left(\dfrac{2}{\sigma}\right)^2\right)}{\displaystyle\sum_{i=1}^{\infty}\exp\left(-\left(\dfrac{i}{\sigma}\right)^2\right)}, & \text{distance} = 2 \end{cases} \tag{7}$$

3.2　接收准则

接收准则是对 $paris$ (c, b) 集合进行标签筛选的规则, 经标签筛选后, 节点所具有的标签越多则节点隶属多个社区的概率越大. 当节点 G_i 在收到其邻居节点 $neighbor$ (G_i) 传来的标签时, $paris_i$ (c, b) 以标签作为主属性对标签隶属系数 b 进行归一化. 节点 G_i 中某一标签 c 的标签隶属系数的变化如下:

$$b_t(c, G_i) = \frac{b_{t-1}(c, G_i)}{\displaystyle\sum_{G_j \in G} b_{t-1}(G_j, G_i)} \tag{8}$$

根据文献 [7], 为实现有效的重叠社区划分, 需要在标签传播过程中降低社区核心节点 (Core) 的标签个数, 增加社区边缘节点 (Border) 的标签个数. 接收准则以节点的语义凝聚力 W 作为划分社区核心节点和社区边缘节点的度量参数, 某一节点的语义凝聚力越大说明该节点作为社区核心节点的合理性越高, 因此, 该节点在接收标签并筛选后的标签应当越少 (越不容易成为重叠节点). 由于截断阈值的取值为 [0, 1], 因此本文将 $W_i/\max(W)$ 作为节点 G_i 的截断阈值 v_i, 其中 $\max(W)$ 表示所有节点的语义凝聚力最大值. 当节点 G_i 在收到其邻居节点 $neighbor(G_i)$ 传来的标签时, 保留集合 $paris_i(c, b)$ 中标签隶属系数大于 v_i 的标签, 作为节点 G_i 的标签, 若节点 G_i 集合 $paris_i(c, b)$ 中标签隶属系数均小于 v_i, 则将标签隶属系数最大的标签作为节点 G_i 的标签.

3.3　终止准则

利用传播准则和接收准则进行循环迭代, 当标签的传播结果收敛 (各节点的标签不再变化) 时终止循环, 将具有相同标签的节点划分为同一社区, 其中具有多个标签的节点为多个社区的重叠节点.

图 5 为本文 BAT-LPA 算法的重叠社区划分结果, σ 取值为 1.2. 图 6 为 QLSP 的初始 4 次标签传播结果. 其中黑色节点为未分配标签的节点, 白色节点为重叠节点.

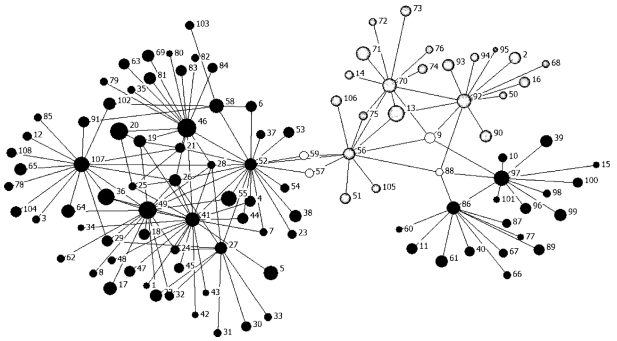
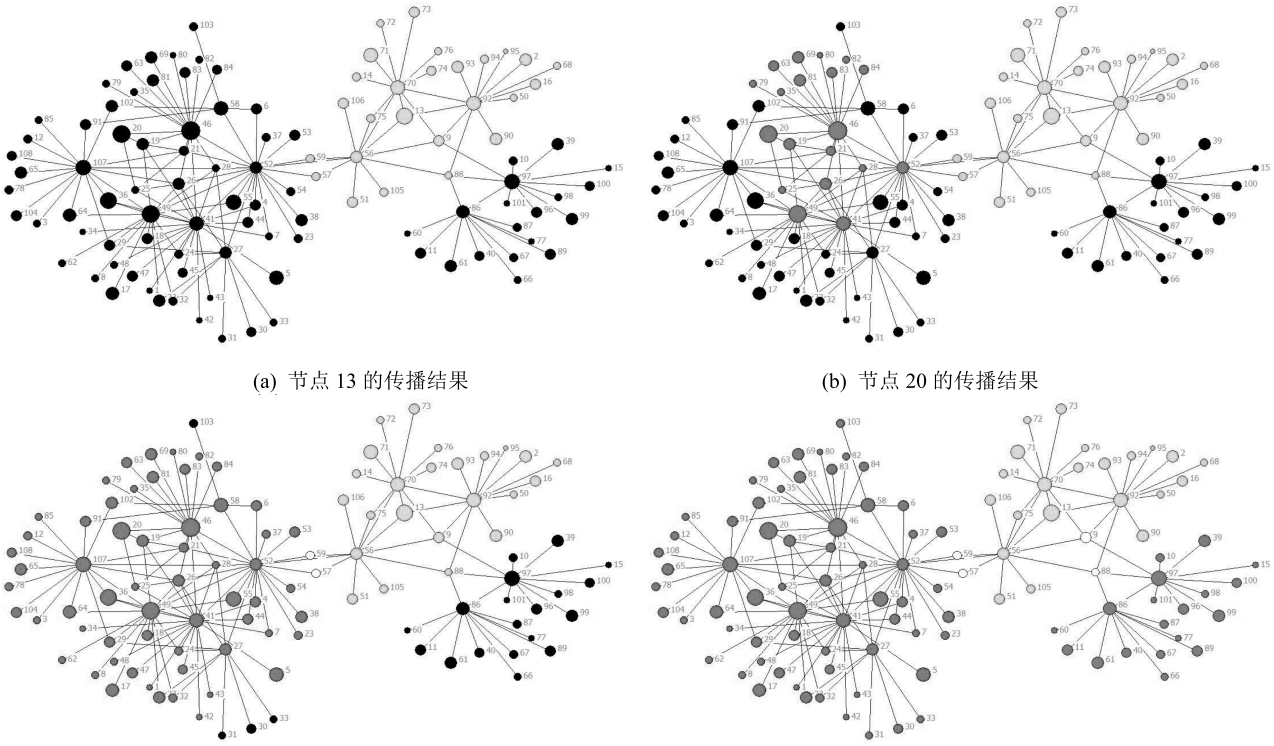


图 5　QLSP 数据集的重叠社区划分结果

Fig. 5　The overlapping communities of QLSP dataset



(a) 节点 13 的传播结果 (b) 节点 20 的传播结果

(c) 节点 49 的传播结果 (d) 节点 97 的传播结果

图 6 QLSP 的初始 4 次标签传播结果

Fig. 6 The initial 4 label propagations for QLSP

4 语义重叠社区发现的评价标准

一般的社会网络重叠评价标准以节点拓扑结构为输入, 文献 [5] 所建立的重叠社区模块度 EQ 度量模型为

$$EQ = \frac{1}{X}\sum_{t}\sum_{i\in C_t, j\in C_t}\frac{1}{O_i O_j}\left[A_{i,j} - \frac{R_i R_j}{X}\right] \quad (9)$$

其中, R_i 为节点 G_i 的度数, X 为网络节点的总度数, A 为网络邻接矩阵, O_i 为节点 G_i 所隶属的社区个数. 语义重叠社区需要以节点关系结构和节点语义信息作为基础, 其评价标准不仅要考虑社区内部的关系合理性, 而且需要考虑节点间的语义信息相似性. 为此, 本文引入以语义空间坐标 $\boldsymbol{m}_i$ ($\boldsymbol{m}_i = \theta'_{i,\cdot}$) 为输入的语义信息相似性度量函数 $U(\boldsymbol{m}_i, \boldsymbol{m}_j)$, 即对所划分的社区内部进行了语义相似性加权, 使得 SQ 与社区内部语义相似性正向相关, 建立可评价语义重叠社区的模块度度量模型 SQ, 其表达式为

$$SQ = \frac{1}{X}\sum_{t}\sum_{i\in C_t, j\in C_t}\frac{U(\boldsymbol{m}_i, \boldsymbol{m}_j)}{O_i O_j}\left[A_{i,j} - \frac{R_i R_j}{X}\right] \quad (10)$$

相似性度量函数的选择需要满足以下两方面要求: 1) 由于模块度的取值范围为 (0, 1), 为此语义信息相似性度量函数的取值范围为 (0, 1); 2) 由于节点的语义空间坐标为向量, 为此 $U(\boldsymbol{m}_i, \boldsymbol{m}_j)$ 需要具有计算多维数据相似性的能力. 综合以上两方面要求, 本文选择余弦相似度作为相似性度量函数 $U(\boldsymbol{m}_i, \boldsymbol{m}_j)$, 其表达式为

$$U(\boldsymbol{m}_i, \boldsymbol{m}_j) = \frac{\boldsymbol{m}_i \times \boldsymbol{m}_j}{|\boldsymbol{m}_i|\,|\boldsymbol{m}_j|} = \frac{\sum_{g=1}^{k} m_{i,g} m_{j,g}}{\sqrt{\sum_{g=1}^{k} m_{i,g}^2 \sum_{g=1}^{k} m_{j,g}^2}} \quad (11)$$

从式 (9) 可知, EQ 从链接关系结构 (拓扑结构) 角度对所发现的社区进行度量. 从式 (10) 可知, SQ 在 EQ 的基础上加入了社区内部语义相似度的评价. 若所划分的社区结构越紧密, 社区内部相似度越高, SQ 越大, 使社区结构的评价不单纯依赖于拓扑结构. 因此, SQ 相较 EQ 具有更合理的语义评价特征, 更适合度量语义社区结构. 本文在第 5 节对 SQ 及 EQ 的评价性能进行了分析.

5 实验分析

5.1 话题个数 T 取值分析

话题个数 T 是本文算法 (BAT-LPA 算法) 的

输入参数, 为验证话题个数 T 对语义社区划分结果的影响, 本文选用如下三组数据作为测试数据: 1) 图 3 所示的清华大学 ArnetMiner 系统 QLSP 数据集; 2) 图 7 所示的 Krebs 建立的美国政治之书网络 (Krebs Polbooks Network), 该数据的网络结点代表亚马逊网上书店卖出的有关美国政治的图书, 每本书的政治倾向略有不同, 但总体上分为三类, 且只有 0 或 1 两种选择, 因此为实现语义化模拟, 将与某一 G_i 具有直接相邻关系 (distance = 1) 的节点 G_j 和间接相邻关系 (distance = 2) 节点 G_k 的信息向量之和作为节点 G_i 的信息向量; 3) 图 8 所示的 Newman 建立的海豚家族 (Dolphins) 关系网络, 该网络由两大家族组成, 个数分别为 20 和 42, 共 159 条链接关系, 为模拟语义社会网络的特性, 本文实验借用 Polbooks 网络及 Dolphins 网络的社会关系特性, 并为每个节点生成三维随机数作为节点的语义坐标.

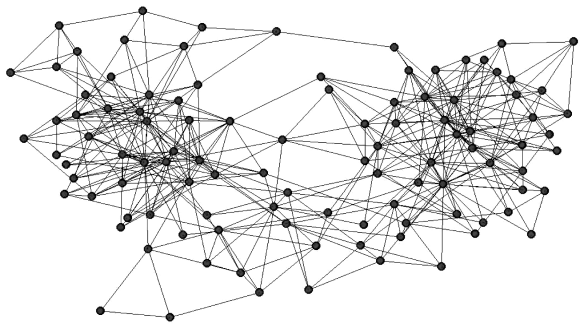


图 7　Polbooks 网络拓扑图

Fig. 7　The topology of Polbooks network

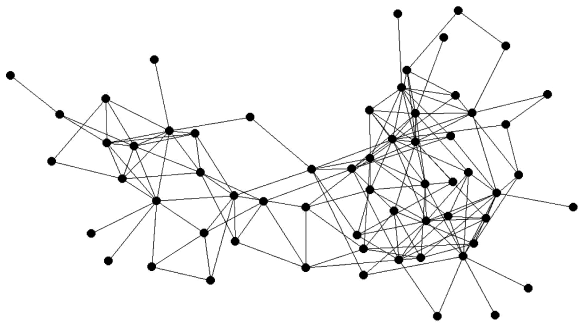


图 8　Dolphins 网络拓扑图

Fig. 8　The topology of Dolphins network

本节实验分别对以上三组数据集 (QLSP, Polbooks, Dolphins) 进行话题个数 T 的取值实验. 图 9 为三组数据集的在话题个数为 T 取值为 (1 ∼ 20) 下的社区个数 (CS) 及 EQ 和 SQ 的对比结果 (σ 取 1.5). 当话题个数增加时各节点的语义坐标较近似, 导致 T 个特征值所构成的向量 $\mathbf{\Lambda} = \{\lambda_1, \cdots, \lambda_T\}$ 的模减小, 从而使各节点的语义凝聚力的取值减小. 根据 BAT-LPA 算法的接收准则可知, 当语义凝聚力较小时, 各节点所保留的标签较多, 从而导致社区个数增加. 在图 9 的社区个数对比中, $CS \in (0, 10)$ 说明了这一过程. 通过 EQ 和 SQ 的对比可知, 当话题个数达到某一最优值后, EQ 和 SQ 的取值随话题个数的增加而下降.

为对比话题个数 T 不同取值的语义社区划分结果, 图 10 截选了三组数据在 $T = \{6, 12, 18\}$ 下的社区划分结果, 其中黑色节点为重叠节点.

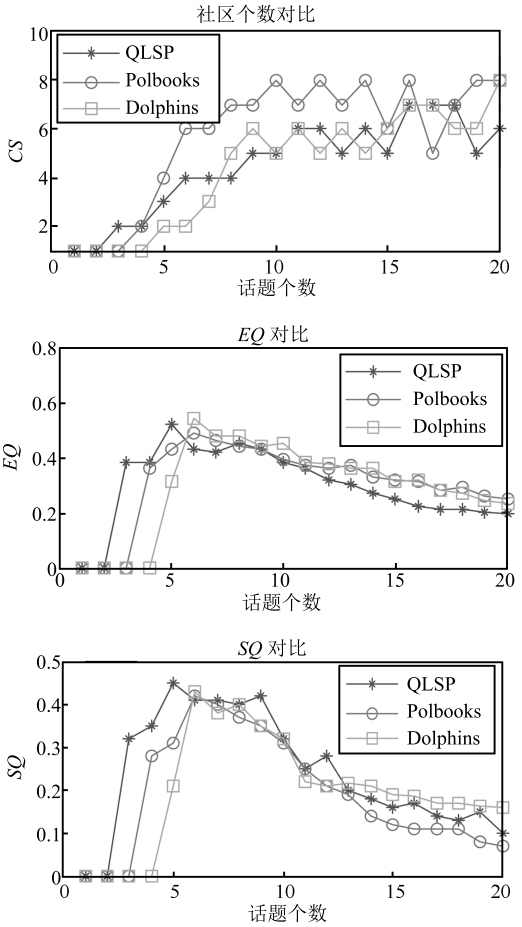


图 9　三组数据集在话题个数为 (1 ∼ 20) 下的对比

Fig. 9　The comparison of the three datasets on the topic (1 ∼ 20)

5.2　距离影响因子 σ 分析

本节实验分别对以上三组数据集 (QLSP, Polbooks, Dolphins) 进行参数 $\sigma \in (0, 5)$ 的测试 (T 取值为 5), 所得社区划分结果的重叠节点个数如图 11 所示. 根据式 (2)、式 (7) 及图 4 的分析可知, 当参数 σ 增加时, 节点的取样范围减小且扩散系数所影响的距离减小, 导致语义凝聚力所代表的区域范围变小, 且在进行社区发现时减小了标签传播范围. 由于语义凝聚力所代表的区域出现集中化且标签的

传播不够充分, 从而影响了社区间的重叠性. 从图 11 所示的对比结果可分析出, 参数 σ 的取值越大, 重叠节点的个数越小, 当 $\sigma > 3$ 时重叠节点个数趋于 0.

3 组数据在参数 $\sigma \in (0, 5)$ 条件下, 社区划分结果的 EQ 和 SQ 如图 12 所示, 从式 (9) 和式 (10) 的对比可知, SQ 加入了语义信息相似性度量函数 U, 且 $U(\boldsymbol{m}_i, \boldsymbol{m}_j) < 1$, 使得 SQ 的总体趋势小于 EQ. 图 12 直观显示了当参数 $\sigma \in (1, 2)$ 时语义社区划分结果的 SQ 值最高, 且 EQ 和 SQ 的极值点

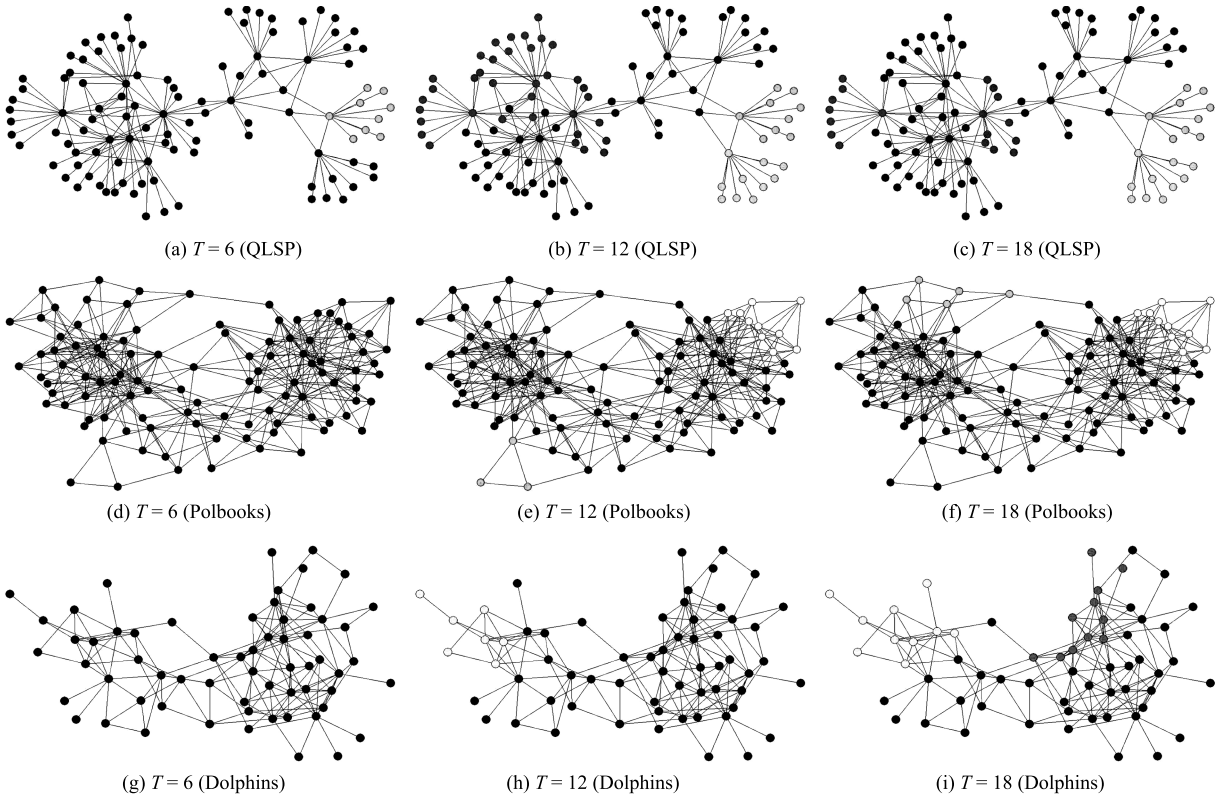


图 10 T 在不同取值下的网络划分结果

Fig. 10 The community structures with different T

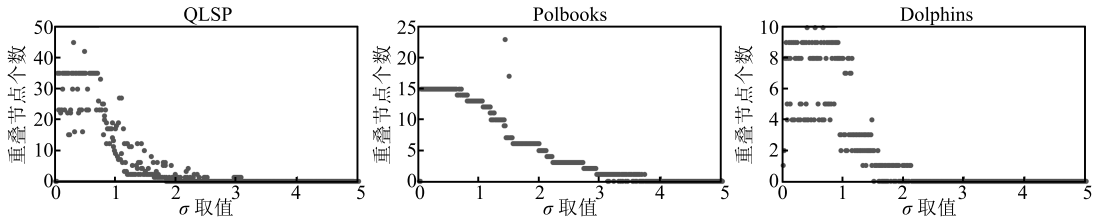


图 11 3 组数据的重叠节点数

Fig. 11 The number of overlapping nodes of three datasets

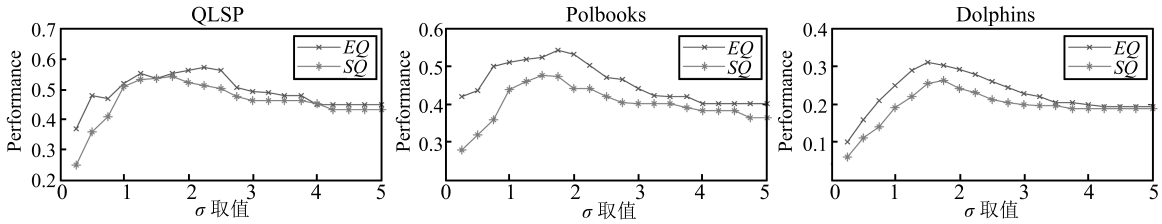


图 12 3 组数据的 EQ 和 SQ 对比

Fig. 12 The comparison of the three datasets on EQ and SQ

不同. 由于 SQ 在 EQ 的基础上加入了社区内部 EQ 的极值点仅代表此处所划分社区的关系结构最合理, 而 SQ 的极值点代表了此处社区内部语义相似性及关系结构最合理. 该实验验证了由于缺少社区内部语义相似性度量, 导致 EQ 在评价语义社区结构时会产生偏差. 为对比 σ 对语义社区划分结果的影响, 本节实验在图 13 中列举了 3 组数据在参数 σ 为 0.5, 1.5, 3.5 (T 取值为 5) 时的社区划分结果. 当 σ 取值越大时, 所划分的语义社区重叠节点数越少.

5.3　重叠社区发现算法比较分析

本节实验目的在于分析经典社区发现算法在面向语义社会网络时划分结果的偏差性, 因此本节实验仅以 QLSP 数据集进行举例说明. 社区发现中经典的社区发现算法包括 GN, FN, LFM, COPRA, UEOC, EAGLE, CPM, LPAm, LPAm+. 其中 LFM, COPRA, UEOC, EAGLE, CPM 为重叠社区发现算法, 由于 QLSP 数据集仅含一个 clique 社区 (26, 28, 41, 46, 49, 52), 不适用于 EAGLE, CPM 算法, 因此本文仅对 GN, FN, LPAm, LPAm+, LFM ($\alpha = 0.85$), COPRA (threshold = 0.5), UEOC (step = 20) 算法进行求解. 图 14 为以上各算法的社区划分结果, 其中黑色节点为重叠节点, 各算法的 SQ 和 EQ 值如表 2 所示. 以上经典算法以链接关系优化划分为导向, 从表 2 中的结果可分析出, 经典算法的 EQ 值最高为 0.5831 (LPAm+), 高于本文算法 (0.5608), 但 SQ 值均低于本文算法 (0.5118), 由此验证了传统面向链接关系的社区划分算法的 EQ 较高, 但在处理语义社区划分问题时 SQ 较低, 所划分的社区结果与语义社区的理想结果偏差较大.

表 2　经典算法的 SQ 和 EQ 值

Table 2　Values of SQ and EQ by classical algorithms

Methods	SQ	EQ
GN	0.3584	0.5417
FN	0.3157	0.4061
LPAm	0.3235	0.4329
LPAm+	0.4311	0.5831
LFM	0.2329	0.4254
COPRA	0.4003	0.5410
UEOC	0.4071	0.4410
BAT-LPA	0.5118	0.5608

5.4　真实数据集比较

本实验从清华大学 ArnetMiner 系统的 QLSP

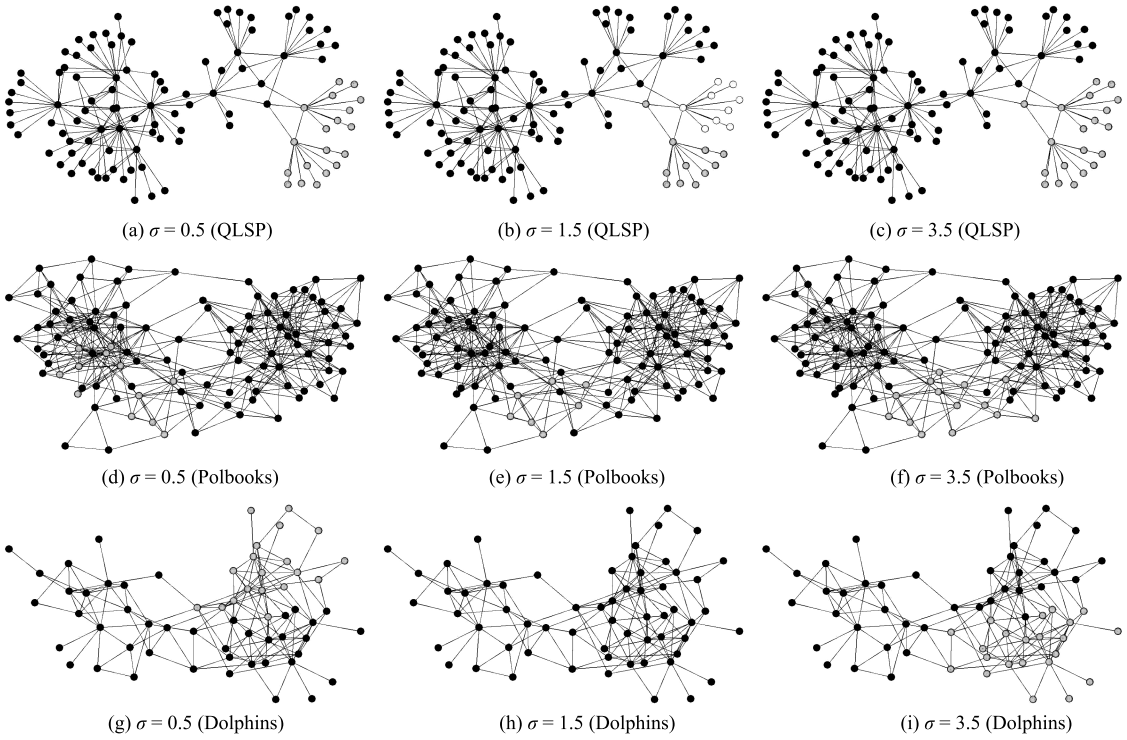


图 13　σ 在不同取值下的社区结构

Fig. 13　The communities structures of different σ

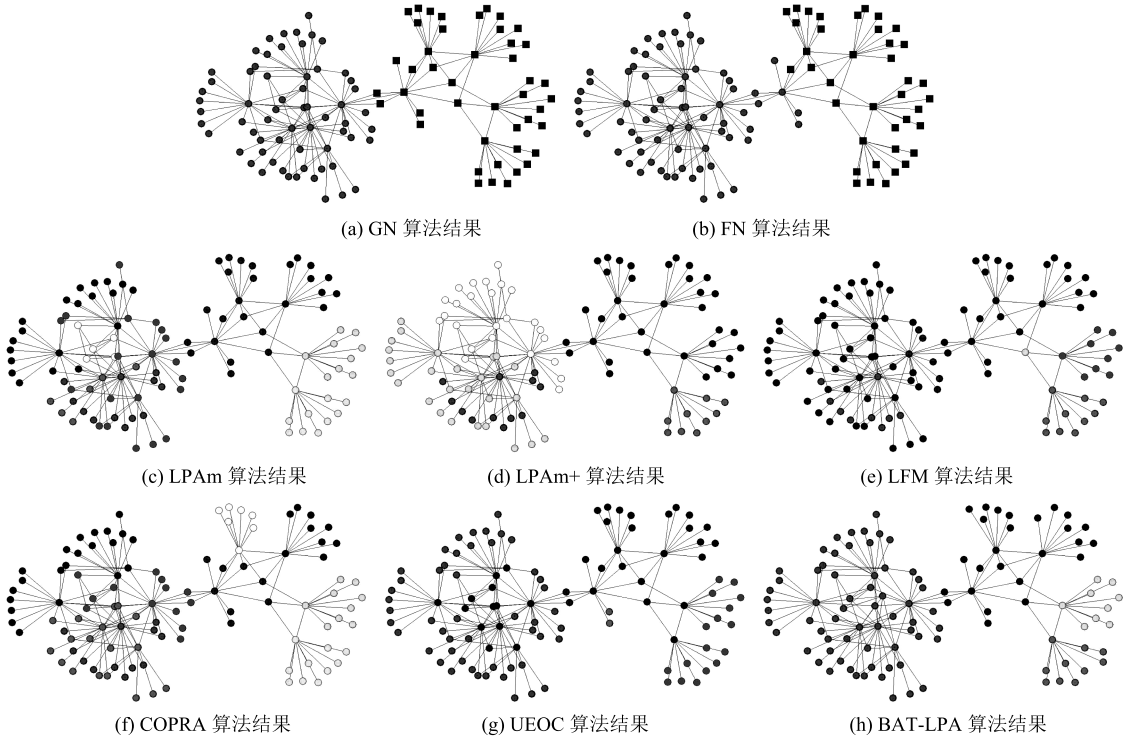


图 14 各算法的社区划分结果

Fig. 14 Community results with classical algorithms

完整数据集 (共 805 个节点)、Aminer-FOAF-Data-Set (AFD) 数据集 (截取 2 000 个节点)、Citation Network Dataset (CND) 数据集 (共 2 555 个节点)、DBLP (April 12, 2006) 数据集 (1 200 000 个节点) 中分别截取 1 500 个节点作为 DBLP (A) 数据集和 2 000 个节点作为 DBLP (B) 数据集, 共 5 组真实实验数据, 用于分析本文算法与经典算法的比较结果.

表 3 为各算法对上述数据集的执行结果, 其中本文 BAT-LPA 运行参数为 $\sigma = 1.5$, $T = 5$, 表 3 包括 EQ, SQ 及社区个数 CS, 图 15 和图 16 分别为各算法的 EQ 和 SQ 直方图, 其中图 15 的结果表示本文 BAT-LPA 算法在 EQ 标准下的结果较差, 其原因在于: BAT-LPA 算法在进行标签传播时, 其接收准则以节点的语义凝聚力 W 作为划分社区核心节点和社区边缘节点的度量参数, 且语义凝聚力受节点间语义相似性的影响, 所划分社区结构考虑了链接关系的紧密性和语义的相似性. 因此, BAT-LPA 算法的 EQ 值与仅强调链接关系紧密性的经典算法相比, 效果较差. 图 16 的结果显示了本文 BAT-LPA 算法的 SQ 值较高, 说明了 BAT-LPA 算法在链接关系紧密性和语义相似性两方面评价条件下, 效果较好. 从图 15 和图 16 的比对可知, 相较于传统经典算法, 本文 BAT-LPA 算法更适合处理语义社会网络的社区发现问题.

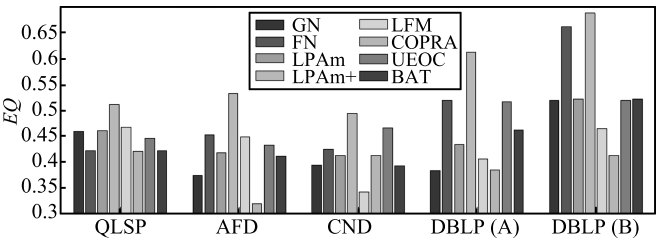


图 15 各算法的 EQ 直方图

Fig. 15 The histogram of EQ with different classical algorithms

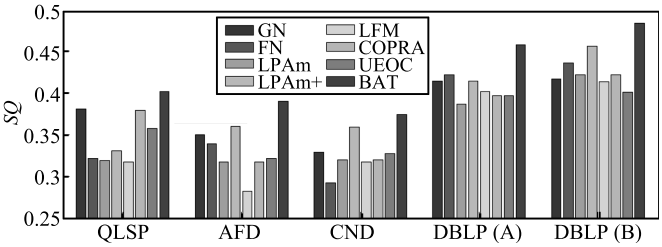


图 16 各算法的 SQ 直方图

Fig. 16 The histogram of SQ with different classical algorithms

5.5 语义社会网络社区发现算对比分析

本节实验对比各类需要预先设定社区个数的语义社区发现算法, 以语义社区发现算法中通用的 Enron[22−23] 数据集作为实验数据集. Enron 数据

集是 Enron 公司 150 个用户的交互数据, 共包含 500 000 条数据. 表 4 为经 LDA 分析后从 Enron 数据集中抽取的 4 组话题. 表 5 和表 6 为 Enron 数据集分别在 TURCM, CART, CUT, LCTA 算法下的 EQ 值及 SQ 值, 表中社区个数表示各算法执行前的预设社区个数.

从对表 5 和表 6 的分析可知, 当社区个数为 10 时, 各算法的 SQ 取值达到最高; 当社区个数为 12 时, 大多数算法的 EQ 取值达到最高, 其原因在于各类面向语义分析的社区发现算法的 EQ 最优值与 SQ 最优值会产生偏差. 根据 SQ 在评价语义社区划分结果时相较于 EQ 更合理, 因此对于 TURCM, CART, CUT, LCTA 算法, Enron 的最佳社区个数为 10. 本文算法的 EQ 和 SQ 取值分别为 0.317 和 0.305, 社区个数为 11. 通过对比可知, 本文算法的结果近于同类算法的 SQ 最优值 (TURCM, $SQ = 0.310$, $CS = 10$), 且本文算法无需预先设定社区个数, 由此验证了本文算法相对于同类算法的优越性.

表 3　各数据集的执行结果

Table 3　Results from classical algorithms on different datasets

算法		QLSP	AFD	CND	DBLP (A)	DBLP (B)
GN	EQ	0.4581	0.3725	0.3928	0.3823	0.5192
	SQ	0.381	0.3497	0.3291	0.4139	0.4165
	CS	10	25	39	17	16
FN	EQ	0.4216	0.4525	0.4235	0.5191	0.6618
	SQ	0.3216	0.3392	0.2921	0.4216	0.4361
	CS	10	27	37	19	16
LPAm	EQ	0.4598	0.4176	0.4119	0.4331	0.5215
	SQ	0.3191	0.3177	0.3202	0.3871	0.4217
	CS	16	30	35	31	23
LPAm+	EQ	0.5108	0.5325	0.4928	0.6123	0.6892
	SQ	0.331	0.3597	0.3591	0.4139	0.4565
	CS	10	21	24	12	13
LFM	EQ	0.4668	0.4473	0.3406	0.4052	0.4641
	SQ	0.3172	0.2821	0.3172	0.4017	0.4133
	CS	12	24	33	22	12
COPRA	EQ	0.4198	0.3186	0.4119	0.383	0.4113
	SQ	0.3791	0.3177	0.3202	0.3971	0.4217
	CS	13	21	35	21	13
UEOC	EQ	0.4449	0.4312	0.4648	0.5158	0.5183
	SQ	0.3577	0.3218	0.3271	0.3964	0.4011
	CS	12	24	30	22	14
BAT-LPA	EQ	0.4214	0.4103	0.3913	0.4611	0.5216
	SQ	0.4021	0.3902	0.3743	0.4581	0.4836
	CS	14	25	34	21	15

表 4　Enron 数据集的话题分组

Table 4　The topics extracted from Enron

Topic	California power	Gas trans	Trading	Deals
word	power	gas	price	meeting
	transmission	energy	market	contract
	energy	Enron	dollar	report
	calpx	transco	nymex	Enron
	California	chris	trade	deal

表 5 各类语义社区发现算法的 EQ 值

Table 5 The EQ of various semantic community detection algorithms

The number of communities	6	8	10	12	14
TURCM[29−30]	0.198	0.271	0.339	0.331	0.283
CART[27]	0.152	0.249	0.302	0.304	0.255
CUT[24]	0.133	0.231	0.266	0.278	0.227
LCTA[31]	0.164	0.239	0.278	0.311	0.249
BAT	0.182	0.294	0.318	0.301	0.248

表 6 各类语义社区发现算法的 SQ 值

Table 6 The SQ of various semantic community detection algorithms

The number of communities	6	8	10	12	14
TURCM[29−30]	0.173	0.231	0.31	0.281	0.261
CART[27]	0.122	0.226	0.268	0.256	0.226
CUT[24]	0.126	0.215	0.235	0.231	0.202
LCTA[31]	0.161	0.208	0.279	0.243	0.215
BAT	0.164	0.221	0.304	0.256	0.213

6 结论

本文针对一般语义社会网络社区划分需要预先设定社区个数的问题且社区结果仅为硬划分, 提出了 BAT-LPA 算法, 该方法将语义社会网络的语义特性和社会关系特性相融合, 在进行社区发现时利用 BAT 取样分析的结果作为节点语义凝聚力的度量, 有效避免了预设社区个数的问题. 本文算法设计的创新思想在于: 1) 提出 BAT 模型, 并建立了以节点为核心的 block 场取样方法, 以增加局部社区的规模特性; 2) 建立了可度量语义凝聚力的方法, 并以语义凝聚力为参数改进了标签传播算法 LPA, 以实现语义重叠社区划分; 3) 提出了评价语义社区划分结果的 SQ 度量模型.

本文算法的实验分析验证了在面向具有语义关系的社区划分问题时, BAT-LPA 相较于经典重叠社区发现算法更有效, 且对于各类语义社会网络具有普遍适用性. 另外, 本文算法可为动态语义社会网络、大规模数据语义社会网络及语义社区推荐等研究领域提供基础, 对深入研究语义社会网络具有一定的理论和实际意义.

References

1 Yang Bo, Liu Da-You, Liu Jinming, Jin Di, Ma Hai-Bin. Complex network clustering algorithms. *Journal of Software*, 2009, **20**(1): 54−66
(杨博, 刘大有, Liu Jinming, 金弟, 马海宾. 复杂网络聚类方法. 软件学报, 2009, **20**(1): 54−66)

2 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of National Academy of Science of the United States of America*, 2002, **99**(12): 7821−7826

3 Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, **69**(6): 066133

4 Palla G, Derenyi I, Farkas I, Vicsde T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, **435**(7043): 814−818

5 Shen H W, Cheng X Q, Cai K, Hu M B. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and Its Applications*, 2009, **388**(8): 1706−1712

6 Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks.*New Journal of Physics*, 2009, **11**(3): 033015-27

7 Gregory S. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 2010, **12**(10): 103018

8 Jin D, Yang B, Baquero C, Liu D Y, He D X, Liu J. A Markov random walk under constraint for discovering overlapping communities in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011, **2011**(5): P05031-21

9 Jin Di, Yang Bo, Liu Jie, Liu Da-You, He Dong-Xiao. Ant colony optimization based on random walk for community detection in complex networks. *Journal of Software*, 2012, **23**(3): 451−464
(金弟, 杨博, 刘杰, 刘大有, 何东晓. 复杂网络簇结构探测 — 基于随机游走的蚁群算法. 软件学报, 2012, **23**(3): 451−464)

10 Gan Wen-Yan, He Nan, Li De-Yi, Wang Jian-Min. Community discovery method in networks based on topological potential. *Journal of Software*, 2009, **20**(8): 2241−2254
(淦文燕, 赫南, 李德毅, 王建民. 一种基于拓扑势的网络社区发现方法, 软件学报, 2009, **20**(8): 2241−2254)

11 Jin Di, Liu Jie, Yang Bo, He Dong-Xiao, Liu Da-You. Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, 2011, **37**(7): 873−882
(金弟, 刘杰, 杨博, 何东晓, 刘大有. 局部搜索与遗传算法结合的大规模复杂网络社区探测. 自动化学报, 2011, **37**(7): 873−882)

12 He Dong-Xiao, Zhou Xu, Wang Zuo, Zhou Chun-Guang, Wang Zhe, Jin Di. Community mining in complex Networks-Clustering combination based genetic algorithm. *Acta Automatica Sinica*, 2010, **36**(8): 1160−1170
(何东晓, 周栩, 王佐, 周春光, 王喆, 金弟. 复杂网络社区挖掘－基于聚类融合的遗传算法. 自动化学报, 2010, **36**(8): 1160−1170)

13 Yang Bo, Liu Jie, Liu Da-You. A random network ensemble model based generalized network community mining algorithm. *Acta Automatica Sinica*, 2012, **38**(5): 812−822
(杨博, 刘杰, 刘大有. 基于随机网络集成模型的广义网络社区挖掘算法. 自动化学报, 2012, **38**(5): 812−822)

14 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

15 Zhang H Z, Qiu B J, Giles C L, Foley H C, Yen J. An LDA-based community structure discovery approach for large-scale social networks. In: Proceedings of the 2007 IEEE on Intelligence and Security Informatics. New Brunswick, NJ: IEEE, 2007. 200−207

16 Kemp C, Tenenbaum J B, Griffiths T L, Yamada T, Ueda N. Learning systems of concepts with an infinite relational model. In: Proceedings of the 21st National Conference on Artificial Intelligence. California: AAAI Press 2006. 381−388

17 Henderson K, Eliassi R T. Applying latent dirichlet allocation to group discovery in large graphs. In: Proceedings of the 2009 ACM Symposium on Applied Computing. Honolulu, Hawaii, USA: ACM, 2009. 1456−1461

18 Henderson K, Eliassi R T, Papadimitriou S, Faloutsos C. HCDF: A hybrid community discovery framework. In: Proceedings of the 2010 SIAM Conference. SDM. 2010. 754−765

19 Zhang H Z, Giles C L, Foley H C, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2007. 663−668

20 Zhang H Z, Li W, Wang X R, Giles C L, Foley H C, Yen J. HSN-PAM: finding hierarchical probabilistic groups from large-scale networks. In: Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Omaha, Nebraska, USA: IEEE, 2007. 27−32

21 Steyvers M, Smyth P, Rosen Z M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2004. 306−315

22 McCallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, 2005, **3**(1): 1−7

23 McCallum A, Wang X R, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 2007, **30**(1): 249−272

24 Zhou D, Manavoglu E, Li J, Giles C L, Zha H Y. Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web. Edinburgh, Scotland, UK: ACM, 2006. 173−182

25 Cha Y, Cho J. Social-network analysis using topic models. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2012. 565−574

26 Wang X, Mohanty N, McCallum A. Group and topic discovery from relations and text. In: Proceedings of the 3rd International Workshop on Link Discovery. New York, USA: ACM, 2005. 28−35

27 Pathak N, DeLong C, Banerjee A, Erickson K. Social topic models for community extraction. In: Proceedings of the 2nd SNA-KDD Workshop. Las Vegas, Nevada, USA: ACM, 2008. 1−10

28 Mei Q Z, Cai D, Zhang D, Zhai C X. Topic modeling with network regularization. In: Proceedings of the 17th International Conference on World Wide Web. Beijing, China: ACM, 2008. 101−110

29 Sachan M, Contractor D, Faruquie T A, Subramaniam V. Probabilistic model for discovering topic based communities in social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2011. 2349−2352

30 Sachan M, Contractor D, Faruquie T A, Subramaniam V L. Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web. New York, USA: ACM, 2012. 331−340

31 Yin Z J, Cao L L, Gu Q Q, Han J W. Latent community topic analysis: integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 2012, **3**(4): 63

32 Xie J R, Szymanski B K, Liu X M. SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Proceedings of the 11th IEEE International Conference of Data Mining Workshops. Vancouver, BC: IEEE, 2011. 344−349

辛 宇 哈尔滨工程大学计算机科学与技术学院博士研究生. 2011 年获哈尔滨理工大学计算机科学与技术学院硕士学位. 主要研究方向为数据库与知识工程.
E-mail: xinyu@hrbeu.edu.cn
(**XIN Yu** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. He received his master degree from Harbin University of Science and Technology in 2011. His research interest covers database and knowledge engineering.)

杨 静 哈尔滨工程大学计算机科学与技术学院教授. 主要研究方向为数据库与知识工程. 本文通信作者.
E-mail: yangjing@hrbeu.edu.cn
(**YANG Jing** Professor at the College of Computer Science and Technology, Harbin Engineering University. Her research interest covers database and knowledge engineering. Corresponding author of this paper.)

谢志强 哈尔滨理工大学计算机科学与技术学院教授. 主要研究方向为数据库与知识工程.
E-mail: xiezhiqiang@hrbust.edu.cn
(**XIE Zhi-Qiang** Professor at the College of Computer Science and Technology, Harbin University of Science and Technology. His research interest covers database and knowledge engineering.)