基于标签传播的语义重叠社区发现算法

辛 宇 [1] 杨 静 [1] 谢志强 [2]

摘 要 语义社会网络 (Semantic social network, SSN) 是一种由信息节点及链接关系构成的新型复杂网络, 为此以节点邻接关系为挖掘对象的传统社会网络社区发现算法无法有效处理语义社会网络重叠社区发现问题. 由此提出标签传播的语义重叠社区发现算法, 该算法以标签传播算法 (Latent Dirichlet allocation, LDA) 模型为语义信息模型, 利用 Gibbs 取样法建立节点语义信息到语义空间的量化映射; 提出可度量节点间相似性的主成分 (Semantic coherent neighborhood propinquity, SCNP) 模型和语义影响力 (Semantic impact, SI) 模型; 以 SCNP 作为标签传播的权重, 以 SI 作为截断值的参数, 提出一种改进的 Semantic-LPA (Semantic label propagation algorithm) 算法; 提出可度量语义社区发现结果的语义模块度模型, 并通过实验分析, 验证了算法及语义模块度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, 标签传播算法

引用格式 辛宇, 杨静, 谢志强. 基于标签传播的语义重叠社区发现算法. 自动化学报, 2014, **40**(10): 2262−2275

DOI 10.3724/SP.J.1004.2014.02262

An Overlapping Semantic Community Structure Detecting Algorithm by Label Propagation

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

Abstract Since the semantic social network (SSN) is a new kind of complex networks consisting of information nodes and link relationships, the traditional community detection algorithms which depend on the adjacency in social networks are not efficient in the SSN. To solve this problem, an overlapping community structure detecting method in semantic social network is proposed based on label propagation. Firstly, the algorithm utilizes the Gibss sampling method to establish the quantization mapping by which semantic information in nodes can be mapped into the semantic space, with the latent Dirichlet allocation (LDA) as the semantic model. Secondly, a principal component SCNP model is proposed which could measure the propinquity between nodes and the semantic impact model. Thirdly, an improved semantic label propagation algorithm is put forward, with SCNP as the weight of propagation and SI as the parameter of threshold. Finally, a semantic model by which the community structure of SSN can be measured is presented. The efficiency and feasibility of the algorithm and the semantic modularity are verified by experimental analysis.

Key words Semantic social network (SSN), overlapping community, latent Dirichlet allocation (LDA), label propagation algorithm (LPA)

Citation Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping semantic community structure detecting algorithm by label propagation. *Acta Automatica Sinica*, 2014, **40**(10): 2262−2275

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. 社区划分算法的研究内容可分为硬社区划分、重叠社区划分及语义社区划分 3 个阶段. 其中硬社区划分和重叠社区划分研究的出发点是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 (Girvan-Newman)[2]、FN (Fast Newman)[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

收稿日期 2013-08-12 录用日期 2014-02-12
Manuscript received August 12, 2013; accepted February 12, 2014
国家自然科学基金 (61370083, 61073043, 61073041, 61370086, 61402126), 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, 61402126) 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

社区发现的 CPM (Clique percolation method) 算法[4]. 此后, 许多经典算法孕育而生, 如 EAGLE[5]、LFM (Lancichinetti Fortunato method)[6]、COPRA (Community overlap propagation algorithm)[7]、UEOC (Unfold and extract overlapping communities)[8]、蚁群算法[9], 拓扑势算法[10] 等. 近年来, 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.

语义社区划分研究的出发点是根据社会网络中节点语义信息内容 (如微博、社会标签等), 将具有相似信息内容的节点划分为同一社区. 由于划分的社区结构基于信息相似性, 所以划分结果更能体现社区的凝聚性. 又因语义信息需以文本分析为基础, 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 3 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 (Simple social network 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 (Hybrid community discovery framework) 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 (Generic weighted network LDA) 算法[19] 及面向层次划分的 HSN-PAM (Hierarchical social network — Pachinko allocation model)[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. 缺点是利用的语义信息并非文本信息, 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 (Author-topic) 模型[21] 作为 LDA 分析的基本模型, 代表算法有 McCallum 等提出的 ART (Author recipient topic) 模型, 该模型在 AT 模型的基础上加入了链接 (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 (Rule author recipient topic) 模型, 扩展了 ART 模型的在社会计算领域的应用[23]. Zhou 等在 AT 模型中加入了用户节点 (User) 分布取样, 提出了 CUT (Community user topic) 模型[24]. Cha 等根据社交网络中跟帖人的话题 (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 (Hierarchical LDA) 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. 缺点是此类算法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 (Group topic) 模型, 该模型是 ART 模型的扩展, 将组 (Group) 取样替代了 ART 模型的链接取样[26]. 随后 Pathak 等论述了链接取样的必要性, 并在 ART 模型的基础上加入了社区 (Community) 取样, 提出了 CART (Community author recipient topic) 模型[27]. 近年来, 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 (Topic modeling with network structure) 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 (Topic user recipient community model)[29−30] 及 LCTA (Latent community topic analysis) 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, 划分的社区数量需要预先设定, 且不同的预设社区数量产生的社区划分结果差异较大.

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. 为此本文设计了面向语义社会网络的重叠社区发现算法, 创新建立节点语义信

息到语义空间的量化映射, 通过构建可度量节点间相似性的主成分 (Semantic coherent neighborhood propinquity, SCNP) 模型以及语义影响力 (Semantic impact, SI) 模型, 建立一种改进的标签传播社区发现算法, 并提出了评价语义社区划分结果的 SQ (Semantic Q-modularity) 模型, 最后通过实验, 分析本文算法参数取值及算法有效性.

1 语义社会网络的 LDA 关系建模

1.1 LDA 关系表示

语义社会网络的语义信息体现在各节点的文本信息内容上, 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 LDA 建模过程进行描述, 涉及的数学符号如表 1 所示.

LDA 语义数据分别利用 $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 三个向量进行存储, 其中 w_i, d_i, z_i 分别为关键字 i 的编号、所属节点号及所属话题号, 图 1 为 LDA 算法的 $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 数据存储结构, 其中阴影部分表示集合内的相同元素, 如图 1 所示, $w_{i1}=w_{i2}=w_{i4}=w_{i5}$ 说明 w_{i1}, w_{i2}, w_{i4}, w_{i5} 为同一单词, $d_{i1}=d_{i3}=d_{i5}=d_{i6}$ 说明 $w_{i1}, w_{i3}, w_{i5}, w_{i6}$ 是同一节点 d_{i1} 的关键字, 且关键字 w_{i1} 在 d_{i1} 中出现 2 次, $z_{i1}=z_{i2}=z_{i6}$ 说明 w_{i1}, w_{i2}, w_{i6} 隶属同一话题 z_{i1}, 且关键字 w_{i1} 在 z_{i1} 中出现 2 次, z_{i1} 分别隶属于 d_{i1}, d_{i2}.

从对图 1 的分析可知, $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 三者间存在三层贝叶斯关系, 根据文献 [14] 的文本分析可知, $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 的数学描述如下:

1) $\theta \sim \text{Dirichlet}(\alpha)$, 节点的话题分布 θ 服从参数为 α 的狄利克雷分布;

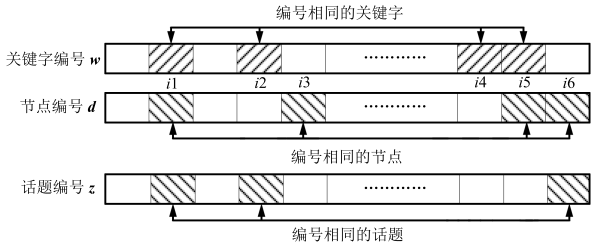


图 1 $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 数据存储结构

Fig. 1 The data storage structure of $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$

2) $z_i|\theta^{(d_i)} \sim \text{Multinomial}(\theta^{(d_i)})$, 节点 d_i 在特定话题分布下, 出现话题 z_i 的概率服从多项式分布;

3) $\lambda \sim \text{Dirichlet}(\beta)$, 关键字服从参数为 β 的狄利克雷分布;

4) $w_i|z_i, \lambda^{(z_i)} \sim \text{Multinomial}(\lambda^{(z_i)})$, 话题 z_i 在特定话题分布下, 出现关键字 w_i 的概率服从多项式分布. 图 2 为关键字 $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 的贝叶斯关系图, 其中箭头指示了 w_i, d_i, z_i 的贝叶斯表达过程, 并以 α 和 β 作为全局参数.

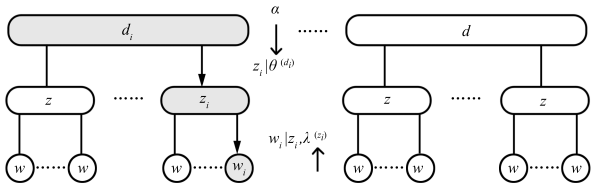


图 2 $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$ 的贝叶斯关系图

Fig. 2 The Bayesian diagram of $\boldsymbol{w}$, $\boldsymbol{d}$, $\boldsymbol{z}$

1.2 Gibbs 迭代过程

$\boldsymbol{w}$, $\boldsymbol{z}$ 的贝叶斯关系表达式为

$$P(z_i=j|w_i)P(w_i)=P(w_i|z_i=j)P(z_i=j) \quad (1)$$

表 1 数学符号说明

Table 1 Mathematical symbols

变量名	变量说明
G	全局网络, G_i 表示网络中的节点 i
$\|G\|$	语义社会网络中的节点个数
N	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
$\boldsymbol{w}$	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
$\boldsymbol{d}$	与关键字的向量 $\boldsymbol{w}$ 对应的节点编号向量, d_i 表示 w_i 所隶属的节点编号
$\boldsymbol{z}$	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, 其最大编号为话题个数 k
$\theta^{(d_i)}$	节点 d_i 的话题分布概率
$\lambda^{(j)}$	话题 j 中关键字的分布, $\lambda_{w_i}^{(j)}$ 表示 w_i 隶属某一话题 j 的概率, $\lambda_{w_i}^{(j)}=P(w_i\|z_i=j)$
α	各节点的话题分布先验参数
β	某一话题内部, 关键字分布的先验参数

其中, $P(w_i) = \sum_{j=1}^{|z|} P(w_i|z_i = j)P(z_i = j)$. Gibbs 取样算法的核心内容在于通过已知样本集合, 建立对某一样本的后验估计, 并对后验估计表达式进行 Gibbs 取样. 实现语义社会网络 LDA 模型的 Gibbs 取样计算, 需要在式 (1) 中加入变量 $\boldsymbol{z}_{-i}$ 和 $\boldsymbol{w}_{-i}$ (表示除去元素 i 的集合 $\boldsymbol{z}$ 和 $\boldsymbol{w}$), 分别作为推断 z_i 和 w_i 的条件, 因此式 (1) 可变形为

$$P(z_i = j|\boldsymbol{z}_{-i}, w_i)P(w_i, \boldsymbol{w}_{-i}) = P(w_i|z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})P(z_i = j|\boldsymbol{z}_{-i}) \quad (2)$$

$$\Rightarrow P(z_i = j|\boldsymbol{z}_{-i}, w_i) \propto P(w_i|z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})P(z_i = j|\boldsymbol{z}_{-i}) \quad (3)$$

根据文献 [16], 式 (3) 的右边分别为

$$P(w_i|z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) = \frac{f_{-i,j}^{(w_i)} + \beta}{f_{-i,j}^{(\cdot)} + |\boldsymbol{w}|\beta} \quad (4)$$

$$P(z_i = j|\boldsymbol{z}_{-i}) = \frac{f_{-i,j}^{(d_i)} + \alpha}{f_{-i,\cdot}^{(d_i)} + |\boldsymbol{z}|\alpha} \quad (5)$$

$$P(z_i = j|\boldsymbol{z}_{-i}, w_i) \propto \frac{f_{-i,j}^{(w_i)} + \beta}{f_{-i,j}^{(\cdot)} + |\boldsymbol{w}|\beta} \frac{f_{-i,j}^{(d_i)} + \alpha}{f_{-i,\cdot}^{(d_i)} + |\boldsymbol{z}|\alpha} \quad (6)$$

其中, $|\boldsymbol{w}|$ 和 $|\boldsymbol{z}|$ 分别表示关键字和话题的个数 (编号的最大值), $f_{-i,j}^{(w_i)}$ 表示关键字 w_i 在话题 j 中的频数, $f_{-i,j}^{(\cdot)}$ 表示话题 j 的关键字总数, $f_{-i,j}^{(d_i)}$ 表示节点 d_i 在话题 j 中的频数, $f_{-i,\cdot}^{(d_i)}$ 表示节点 d_i 的关键字总数. 图 3 为 Gibbs 取样过程, 其中箭头 a 为循环取样过程; 箭头 b 为根据当前样本通过式 (6) 计算 $P(z_i = j|\boldsymbol{z}_{-i}, w_i)$; 箭头 c 为根据 $P(z_i = j|\boldsymbol{z}_{-i}, w_i)$ 对 z_i 进行 Gibbs 取样并修改 z_i 的过程. 当 $P(z_i = j|\boldsymbol{z}_{-i}, w_i)$ 收敛时则结束此过程, 并将 $P(z_i = j|\boldsymbol{z}_{-i}, w_i)$ 按关键字 w_i 归一化, 即可得到关键字 – 话题概率矩阵 $B_{i,j}$, 其中 $B_{i,j} = p(z_i = j|w = i)$, $\boldsymbol{B}_{i,\cdot} = 1$.

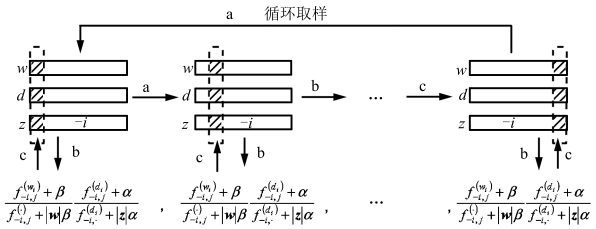


图 3　Gibbs 取样过程

Fig. 3　The process of Gibbs sampling

2　节点的语义量化映射

本文以 LDA 模型提取的 k 个话题作为 k 维语义空间的基, 向量 $\boldsymbol{B}_{i,\cdot}$ 可表示为 i 号关键字在 k 维语义空间中的坐标, 则某一节点 G_i 在语义空间中的坐标 (语义坐标) $\boldsymbol{m}_i$ 可通过 G_i 的 N_i 个关键字加权均值形式表达为

$$m_i = \frac{\sum_{j=1}^{N_i} B_{N_{i,j}}}{N_i} \quad (7)$$

其中, N_i 表示节点 G_i 的关键字个数, $N_{i,j}$ 为 G_i 的第 j 个关键字编号. 通过式 (7) 可将各节点的语义信息量化为 k 维向量 $\boldsymbol{m}$, 因此, 可利用向量内积表达节点间的语义相关性, 从而构造语义相关矩阵 E

$$E_{i,j} = A_{i,j}(m_i \cdot m_j) \quad (8)$$

本文以清华大学 ArnetMiner 系统 Quantifying Link Semantics-publication (QLSP) 数据集的部分数据为例 (其中包含 108 篇论文的 155 条引用关系), 其网络模型如图 4 所示. 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 k 为 5 进行 Gibbs 取样迭代后, 再利用式 (8) 计算语义相关矩阵 E, 其结果如图 5 所示.

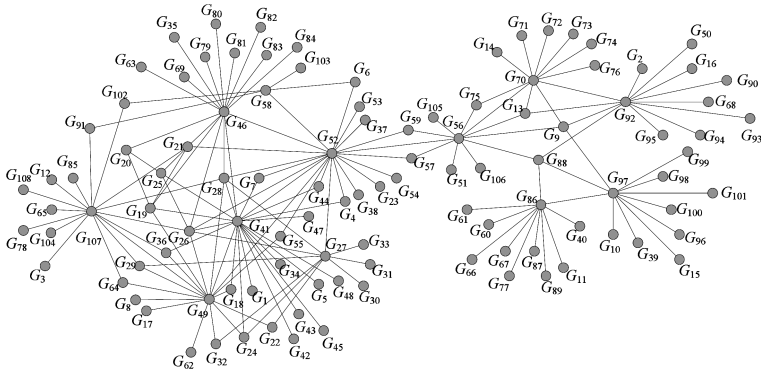


图 4　QLSP 网络拓扑

Fig. 4　The topology of QLSP network

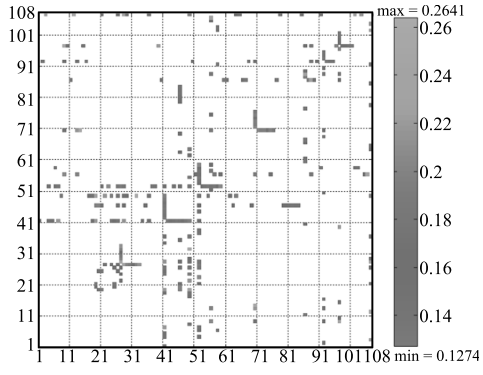


图 5 QLSP 的语义相关矩阵 E

Fig. 5 The semantic correlation matrix E on QLSP

2.1 语义拓扑相关性 SCNP

文献 [32] 定义了网络节点间的拓扑相关性 (Coherent neighborhood propinquity, CNP), 用于度量节点间的关系紧密性, 其表达式为

$$CNP(G_i,G_j)=|A(G_i,G_j)|+|Ne(G_i)\cap Ne(G_j)|+|A(G[Ne(G_i)\cap Ne(G_j)])| \quad (9)$$

其中, $|A(G_i,G_j)|$ 表示相连节点 G_i 和节点 G_j 的边数, $|A(G_i,G_j)|=A_{i,j}$, $Ne(G_i)$ 表示节点 G_i 的相邻节点集, $|Ne(G_i)\cap Ne(G_j)|$ 表示节点 G_i 和节点 G_j 的公共相邻节点个数, $|A(G[Ne(G_i)\cap Ne(G_j)])|$ 表示节点 G_i 和节点 G_j 相邻节点间的链接个数. 文献 [33] 提出了 $weighted\text{-}CNP$, 表达式为

$$weighted\text{-}CNP(G_i,G_j)=|A(G_i,G_j)|+\frac{FE_1}{FE_2}|Ne(G_i)\cap Ne(G_j)|+\frac{FE_1}{FE_3}|A(G[Ne(G_i)\cap Ne(G_j)])| \quad (10)$$

其中, FE_1、FE_2 和 FE_3 分为 $|A(G_i,G_j)|$、$|Ne(G_i)\cap Ne(G_j)|$ 和 $|A(G[Ne(G_i)\cap Ne(G_j)])|$ 的熵值. CNP 和 $weighted\text{-}CNP$ 均以邻接矩阵 A 为参数, 体现的拓扑相关性缺少了语义相关性. 为实现语义社会网络的相关性度量, 本文将语义相关矩阵 E 与拓扑相关性 CNP 相结合, 建立语义拓扑相关性度量 $SCNP$, 表达式

$$SCNP(G_i,G_j)=|E(G_i,G_j)|+\frac{1}{2}|E(Ne(G_i)\cap Ne(G_j),[G_i,G_j])|+|E(G[Ne(G_i)\cap Ne(G_j)])| \quad (11)$$

其中, $|E(G_i,G_j)|$ 表示相连节点 G_i 和节点 G_j 的语义相关性, $|E(G_i,G_j)|=E_{i,j}$, $|E(Ne(G_i)\cap Ne(G_j),[G_i,G_j])|$ 表示节点 G_i 和节点 G_j 的公共相邻节点与 G_i 和 G_j 语义相关性之和, $|E(G[Ne(G_i)\cap Ne(G_j)])|$ 表示节点 G_i 和节点 G_j 相邻节点间的语义相关性之和, 根据式 (10), $weighted\text{-}SCNP$ 为

$$weighted\text{-}SCNP(G_i,G_j)=|E(G_i,G_j)|+\frac{FE_1}{2FE_2}|E(Ne(G_i)\cap Ne(G_j),[G_i,G_j])|+\frac{FE_1}{FE_3}|E(G[Ne(G_i)\cap Ne(G_j)])| \quad (12)$$

本文根据主成分分析 (Principal component analysis, PCA) 思想提出主成分 $SCNP(p\text{-}SCNP)$, 表达式为

$$p\text{-}SCNP(G_i,G_j)=\varepsilon_1|E(G_i,G_j)|+\frac{1}{2}\varepsilon_2|E(Ne(G_i)\cap Ne(G_j),[G_i,G_j])|+\varepsilon_3|E(G[Ne(G_i)\cap Ne(G_j)])| \quad (13)$$

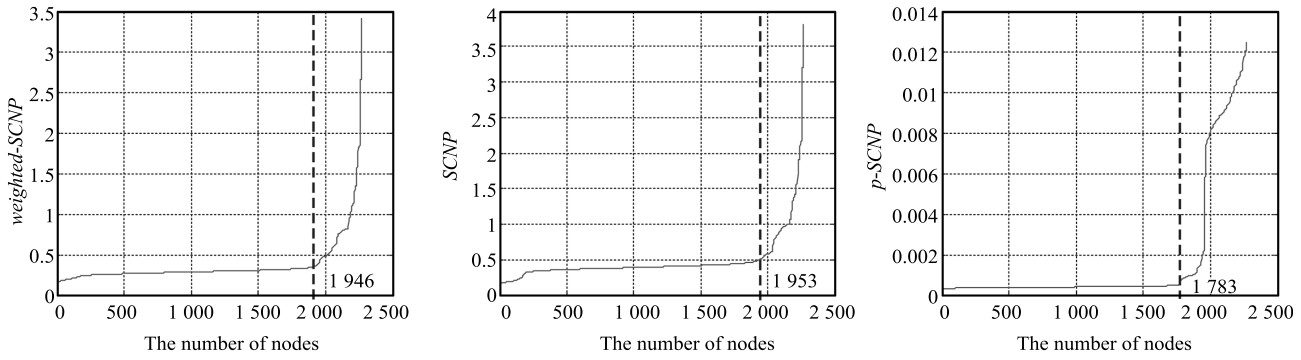
其中, ε 为 $|E(G_i,G_j)|$, $|E(Ne(G_i)\cap Ne(G_j),[G_i,G_j])|$ 和 $|E(G[Ne(G_i)\cap Ne(G_j)])|$ 相关矩阵的最大特征值所对应的特征向量. 由于语义拓扑相关性的意义在于表达节点间的语义差异, 因此语义拓扑相关性排序结果的拐点越小, 则语义社会网络中节点间语义差异的层次越多, 表达的语义差异越充分. 图 6 为 QLSP 数据集 $weighted\text{-}SCNP$, $SCNP$ 和 $p\text{-}SCNP$ 排序结果对比图, 其拐点坐标分别为 1 946, 1 953 和 1 783. 从图 6 可知, 拐点左侧节点的 $weighted\text{-}SCNP$, $SCNP$ 和 $p\text{-}SCNP$ 语义差异不明显, 其中 $p\text{-}SCNP$ 的拐点最小, 因而语义差异较大的节点更多, 更利于充分表达语义拓扑相关性.

2.2 语义影响力分析

语义影响力 (Semantic impact, SI) 是衡量节点在语义社会网络中重要程度的指标. 语义社会网络中, 每一节点的语义影响力取决于该节点的语义坐标及拓扑结构. 网络 G 中每个节点受与之相邻节点的影响, 且节点 G_i 与 G_j 间影响力 $I_{i,j}$ 的大小与节点的距离成反比, 根据文献 [10, 34] 定义的社会网络数据场模型, 节点 G_i 与 G_j 间语义作用力 $I_{i,j}$ 可表达为

$$I_{i,j}=\boldsymbol{m}_i\cdot\boldsymbol{m}_j\exp\left[-\frac{dis_{i,j}^2}{\sigma^2}\right] \quad (14)$$

其中, $dis_{i,j}$ 为节点 G_i 与节点 G_j 间的距离 (跳数), σ 为语义影响力控制因子, 本文在实验章节论述了 σ 的取值. 根据式 (14), 节点 G_i 的语义影响力可表达为 G_i 与网络 G 中所有节点的语义作用力之和, 表达式为



(a) *weighted-SCNP*　(b) *SCNP*　(c) *p-SCNP*

图 6　QLSP 数据集的 *weighted-SCNP*, *SCNP* 和 *p-SCNP* 排序结果对比图

Fig. 6　The comparison figures of *weighted-SCNP*, *SCNP* and *p-SCNP* on QLSP network

$$SI_{i,j} = \sum_{j=1}^{|G|} \boldsymbol{m}_i \cdot \boldsymbol{m}_j \exp\left(-\frac{dis_{i,j}^2}{\sigma^2}\right) \tag{15}$$

其中, $\boldsymbol{m}_i \cdot \boldsymbol{m}_j$ 表达了 G_i 与 G_j 的语义相关性, 若 G_i 与 G_j 语义坐标 $\boldsymbol{m}_i$ 和 $\boldsymbol{m}_j$ 越相似, 则 $\boldsymbol{m}_i \cdot \boldsymbol{m}_j$ 越大; $\exp(-(dis_{i,j}/\sigma)^2)$ 表达了 G_i 与 G_j 的距离相关性, 节点间的距离越大则 $\exp(-(dis_{i,j}/\sigma)^2)$ 越小; 求和表达了节点的度数相关性, 若 G_i 为度数较高的 "富人节点" 则与其距离较小的节点相对较多, 求和后的结果相对较大. 距离相关和度数相关性统称为拓扑相关性, 是传统社会网络中节点重要度的主要参数. 为验证 σ 对 SI 的影响及 SI 对语义相关性和拓扑相关性的平衡作用, 本文从数据集 QLSP 中选取度数不同的 6 个节点 ($G_{90}, G_4, G_{28}, G_{92}, G_{49}, G_{52}$, 度数分别为 1, 2, 6, 12, 17, 21) 进行 σ ($0.5 \sim 15$) 的取值分析, 结果如图 7 所示. 当 σ 大于 5 时 SI 逐渐收敛, 且 SI 的取值不单单依赖于度数, 节点 G_{28} 和 G_{49} 的度数分别为 6 和 17, 节点 G_{49} 的度数相关性较大, 而二者的 SI 取值近似, 验证了 SI 的取值平衡了语义相关性和拓扑相关性.

3　语义重叠社区发现的标签传播算法

标签传播算法 (Label propagation algorithm, LPA) 是社区发现的经典算法, 文献 [35] 提出了 LPA 算法的基本框架 SLPA, 即节点间通过所建立的传播准则 (Speaking rule)、接收准则 (Listening rule) 和终止准则 (Stop criterion) 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 Semantic-LPA (Semantic label propagation algorithm), 需要改进 LPA 算法的传播准则、接收准则和终止准则.

1) Semantic-LPA 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), 函数 $b_t(c, G_i)$ 表示 t 时刻节点 G_i 中标签 c 的标签隶属系数值, 初始时节点 G_i 的集合 $paris_i(c, b) = (i, 1)$. 当节点 G_i 在收到其邻居节点 $neighbor(G_i)$ 传来的标签时, G_i 的集合 $paris_i(c, b)$ 按语义拓扑相关性 $SCNP$ ($weighted\text{-}SCNP$, $p\text{-}SCNP$) 进行加权, 并对标签隶属系数 b 进行归一化. Semantic-LPA 传播准则中标签隶属系数的变化如下:

$$b_t(c, G_i) = \frac{\sum\limits_{G_j \in neighbor(G_i)} b_{t-1}(c, G_j)}{\sum\limits_{G_j \in neighbor(G_i)} SCNP(G_i, G_j)} \tag{16}$$

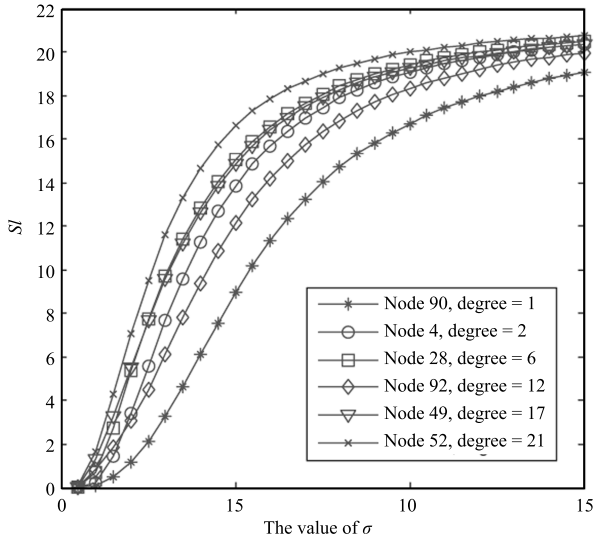


图 7　QLSP 数据集中的不同度数节点的 SI

Fig. 7　The SI of nodes with different degrees on QLSP

2) Semantic-LPA 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. 根据文献 [9], 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) 的标签数量, 增加社区边缘节点 (Border) 的标签数量. 本文 Semantic-LPA 接收准则以语义影响力 SI 作为划分核心节点和社区边缘节点的度量, 节点的语义影响力越大筛选后的标签越少. 由于截断值的取值为 $[0,1]$, 因此本文将 $SI_i/\max(\{SI_1, \cdots, SI_{|G|}\})$ 作为节点 G_i 的截断值 v_i. 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) Semantic-LPA 终止准则, 利用 Semantic-LPA 传播准则和 Semantic-LPA 接收准则进行循环迭代, 当标签的传播结果收敛时 (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.

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 (Extension Q-modularity) 模型为

$$EQ = \frac{1}{X}\sum_{t}\sum_{i\in C_t, j\in C_t}\frac{1}{O_iO_j}\left[A_{i,j} - \frac{R_iR_j}{X}\right] \quad (17)$$

其中, R_i 为节点 d_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 $\boldsymbol{m}_i$ 为输入的语义信息相似性度量函数 $U(\boldsymbol{m}_i, \boldsymbol{m}_j)$, 建立可评价语义重叠社区的模块度模型 SQ, 其表达式为

$$SQ = \frac{1}{X}\sum_{t}\sum_{i\in C_t, j\in C_t}\frac{U(\boldsymbol{m}_i, \boldsymbol{m}_j)}{O_iO_j}\left[A_{i,j} - \frac{R_iR_j}{X}\right] \quad (18)$$

由于模块度的取值范围为 (0, 1), 为此, 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

$$U(\boldsymbol{m}_i, \boldsymbol{m}_j) = \frac{\boldsymbol{m}_i \cdot \boldsymbol{m}_j}{|\boldsymbol{m}_i||\boldsymbol{m}_j|} = \frac{\sum\limits_{g=1}^{k} m_{i,g}m_{j,g}}{\sqrt{\sum\limits_{g=1}^{k} m_{i,g}^2 \sum\limits_{g=1}^{k} m_{j,g}^2}} \quad (19)$$

5 实验分析

5.1 语义影响力控制因子分析

语义影响力控制因子 σ 是本文 Semantic-LPA 算法的输入参数, 为验证参数 σ 对语义社区划分结果的影响, 本文选用如下 3 组数据作为测试数据: 1) 图 4 所示的清华大学 ArnetMiner 系统 QLSP 数据集; 2) 图 8 所示的 Krebs 建立的美国政治之书网络 (Krebs polbooks network), 该数据的网络节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 3 类, 且只有 0 或 1 两种选择, 因此为实现语义化模拟, 将与某一节点 G_i 具有直接相邻关系 (距离为 1) 的节点 G_j 和间接相邻关系 (距离为 2) 的节点 G_k 的信息向量之和作为节点 G_i 的信息向量; 3) 图 9 所示的 Newman 建立的海豚家族 (Dolphins network) 关系网络, 该网络由两大家族组成, 成员个数分别为 20 和 42, 共 159 条链接关系, 为模拟语义社会网络的特性,

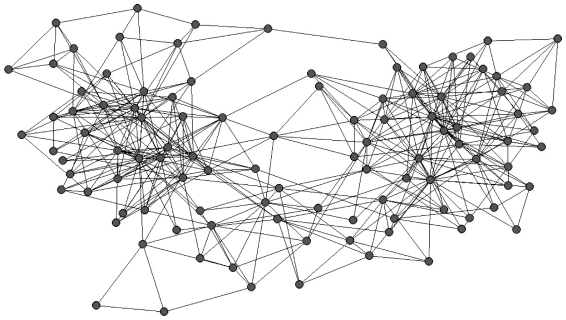


图 8 Polbooks 网络拓扑图

Fig. 8 The topology of polbooks network

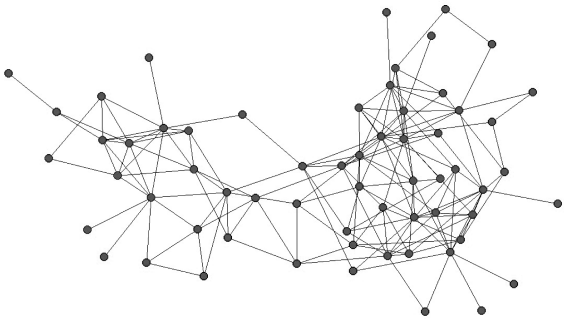


图 9 Dolphins 网络拓扑图

Fig. 9 The topology of dolphins network

本文实验借用 Dolphins 网络的社会关系特性, 并为每个节点生成 3 维随机数作为节点的语义坐标.

本节实验分别对以上 3 组数据集 (QLSP, polbooks, dolphins) 进行参数 σ (0∼5) 的测试, 3 组数据分别以式 (11)∼(13) 所示的语义拓扑相关性 ($SCNP$, $weighted\text{-}SCNP$, $p\text{-}SCNP$) 作为节点度量, 所得社区划分结果的重叠节点个数如图 10 所示. 由图 10 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 σ (0∼5) 条件下, 社区划分结果的 EQ 和 SQ 如图 11 所示, 由式 (17) 和式 (18) 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 11 显示了当参数 $\sigma \in (1, 2)$ 时语义社区划分结果的 SQ 值最高, 且语义社区划分结果 EQ 和 SQ 的极值点不同, 说明以标准模块度 EQ 模型评价语义社区划分结果会导致偏差. 从图 11 的语义拓扑相关性 ($SCNP$, $weighted\text{-}SCNP$, $p\text{-}SCNP$) 比较可知, 利用 $p\text{-}SCNP$ 所得到的结果 SQ 值高于 $SCNP$ 和

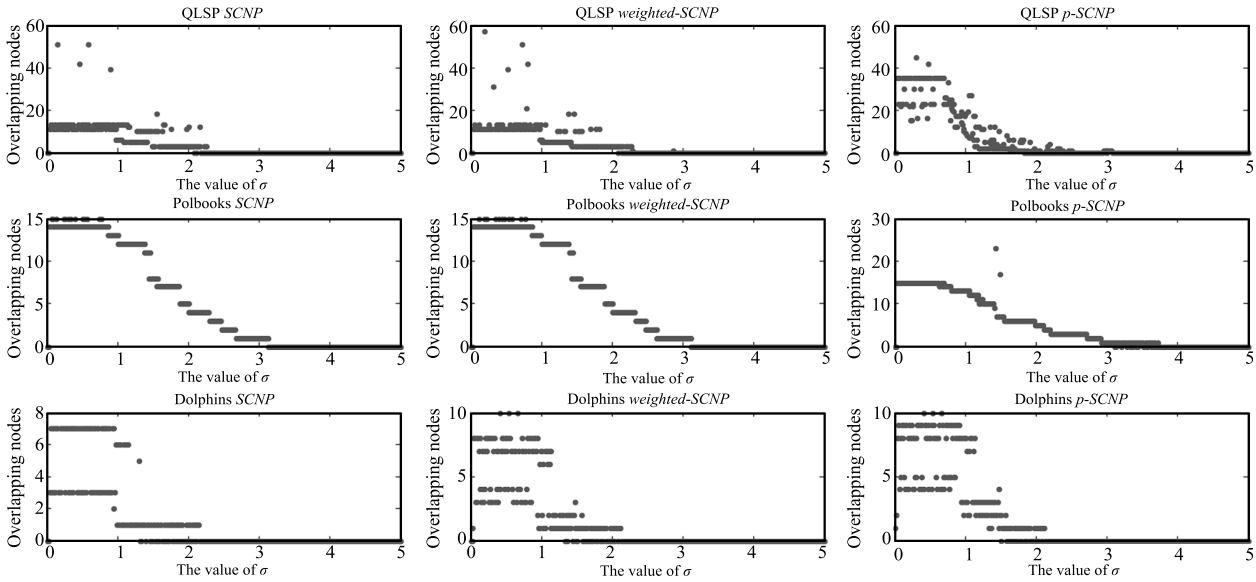


图 10　3 组数据的重叠节点数

Fig. 10　The number of overlapping nodes on the three datasets

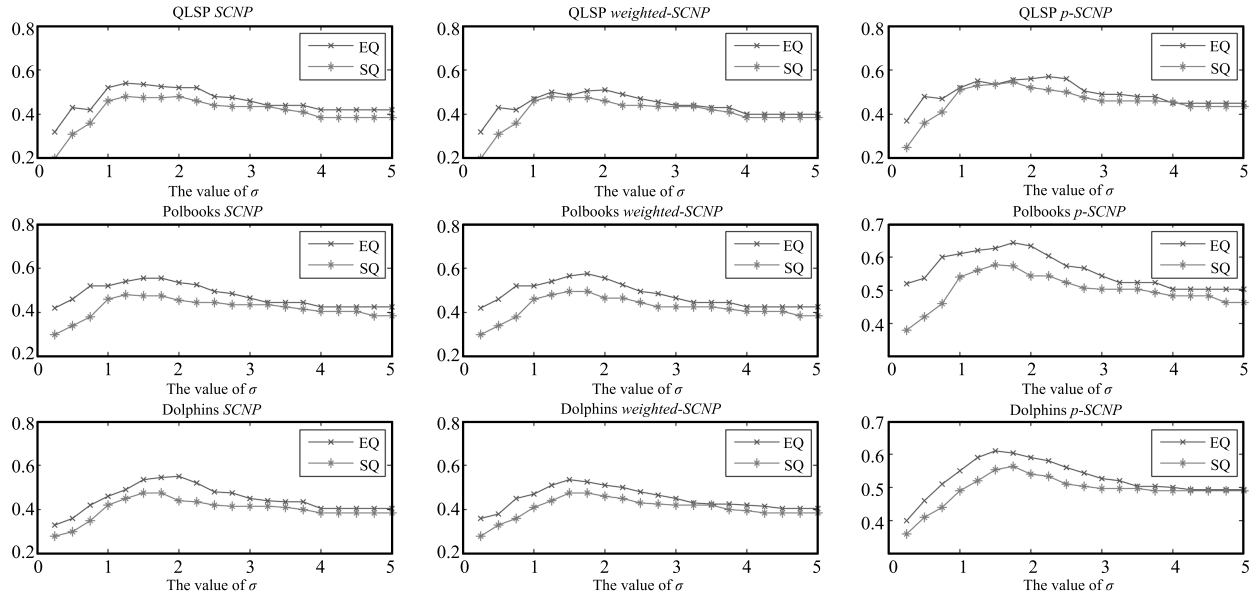


图 11　3 组数据的 EQ 和 SQ

Fig. 11　The EQ and SQ on three datasets

weighted-SCNP.

为对比 σ 不同取值的语义社区划分结果, 实验在图 12 中列举了 3 组数据在参数 σ 为 0.5, 1.5, 3.5, 且语义拓扑相关性为 *p-SCNP* 时的社区划分结果 (深色节点为重叠节点). 当 σ 取值越大时, 根据式 (15), 节点间的影响力 SI 越小, 划分的语义社区重叠节点数越少.

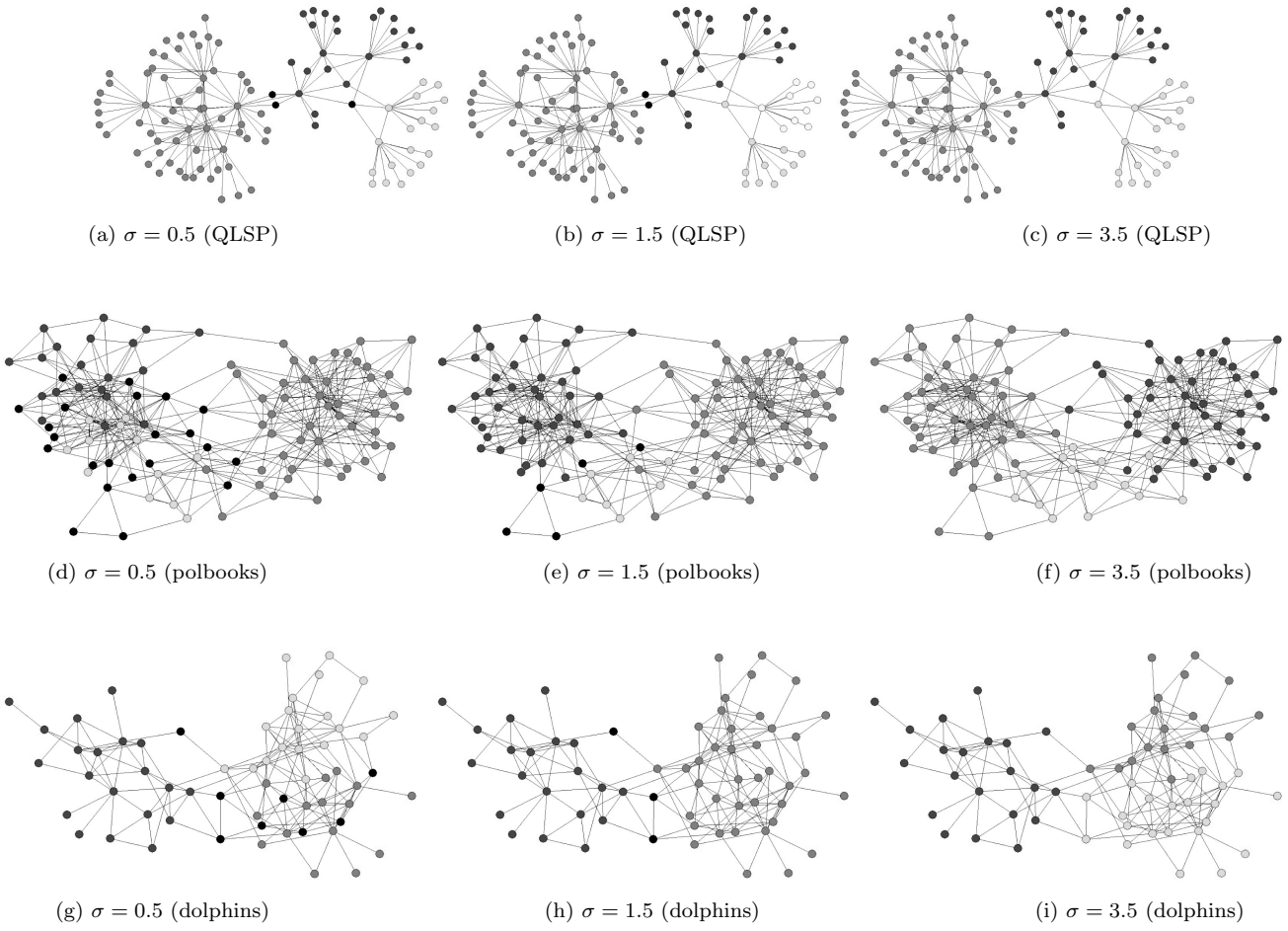
5.2 重叠社区发现算法比较分析

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, 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+ 等. 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 和 CPM 算法, 因此本文仅对 GN、FN、LPAm、LPAm+、LFM、COPRA、UEOC 等算法进行求解, 图 13 为以上各算法的社区划分结果, 其中深色节点为重叠节点, 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 (0.5831) 高

表 2 经典算法的 SQ 和 EQ 值

Table 2 Values of SQ and EQ by classical algorithms

算法	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
LPA	0.4003	0.5410
UEOC	0.4071	0.4410
Semantic-LPA	0.5430	0.5720



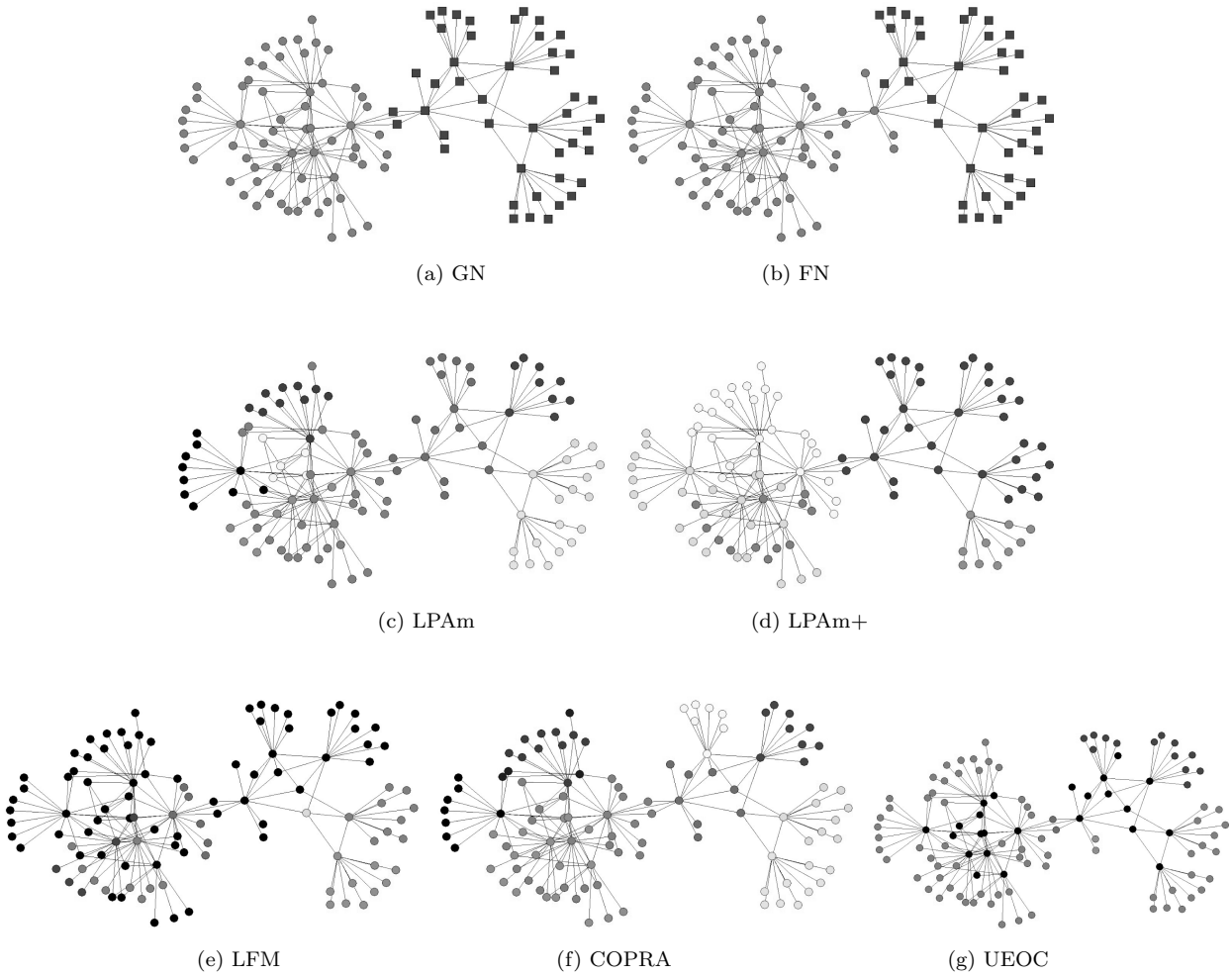
(a) $\sigma = 0.5$ (QLSP) (b) $\sigma = 1.5$ (QLSP) (c) $\sigma = 3.5$ (QLSP)

(d) $\sigma = 0.5$ (polbooks) (e) $\sigma = 1.5$ (polbooks) (f) $\sigma = 3.5$ (polbooks)

(g) $\sigma = 0.5$ (dolphins) (h) $\sigma = 1.5$ (dolphins) (i) $\sigma = 3.5$ (dolphins)

图 12 σ 在不同取值下的社区结构

Fig. 12 Community structures with different σ



(a) GN (b) FN

(c) LPAm (d) LPAm+

(e) LFM (f) COPRA (g) UEOC

图 13 各算法的社区划分结果

Fig. 13 Community results with classical algorithms

于本文算法 (0.5720), 但 SQ 值均低于本文算法 (0.5430), 由此验证了传统社区划分算法的 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.

5.3 真实数据集比较

实验从清华大学 ArnetMiner 系统的 QLSP 完整数据集 (共 805 个节点), Aminer-FOAF-DataSet (AFD) 数据集 (截取 2 000 个节点), Citation network dataset (CND) 数据集 (共 2 555 个节点) 和 DBLP (April 12, 2006) 数据集 (1 200 000 个节点) 中分别截取 1 500 个节点数据集 (DBLP(A)) 和 2 000 个节点数据集 (DBLP(B)) 作为实验数据. 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 Semantic-LPA 运行参数为 $\sigma = 1.5$, 表 3 包括 EQ, SQ 及社区个数 (Community size, CS), 图 14 和图 15 分别为各算法的 EQ 和 SQ 直方图, 其中图 14 的结果表示本文 Semantic-LPA 算法结果在 EQ 标准下的结果较差, 图 15 的结果充分验证了本文 Semantic-LPA 算法的 SQ 较高, 语义社区划分结果更精确, 从图 14 和图 15 的比对可知, 相较于传统经典算法, 本文 Semantic-LPA 算法更适合处理语义社会网络的社区发现问题.

5.4 语义社区网络社区发现算法比较分析

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] 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 0.5 M 条数据, 423 M 数据量. 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 分别为 Enron 数据集在 TURCM、CART、CUT、LCTA 算法下的 EQ 值及 SQ 值, 表中社区数量表示各算法执行前的社区预设数量. 从表 5 和表 6 的分析可知, Enron 数据集的最佳个数为 10. 本文算法的 EQ

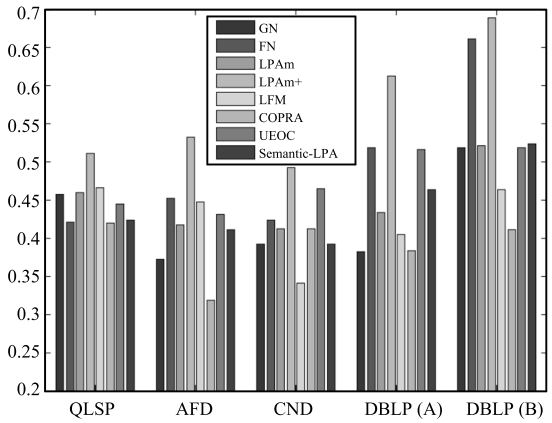


图 14 各算法的 EQ 直方图

Fig. 14 The histogram of EQ with different classical algorithms

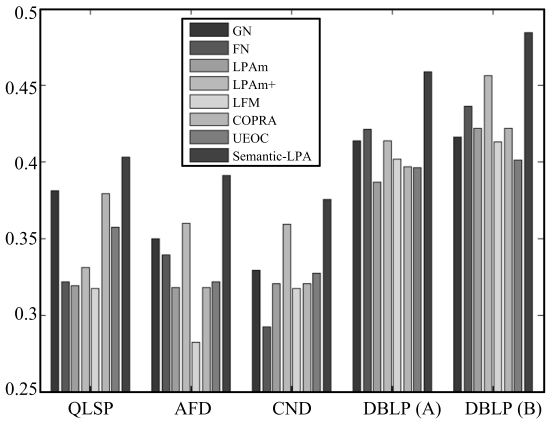


图 15 各算法的 SQ 直方图

Fig. 15 The histogram of SQ with different classical algorithms

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
Semantic-LPA	EQ	0.4236	0.4115	0.3925	0.4643	0.5238
	SQ	0.4032	0.3913	0.3754	0.4592	0.4847
	CS	13	26	33	25	16

和 SQ 取值分别为 0.318 和 0.304. 通过对比可知, 本文算法的结果近于同类算法的最优值, 且无需预先设定社区个数, 由此验证了本文算法相对同类算法的优越性.

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
Semantic-LPA	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
Semantic-LPA	0.164	0.221	0.304	0.256	0.213

5.5 实验总结

本文实验部分分别从参数取值、SQ 有效性、经典算法比较和多数据集分析 4 个方面进行分析, 结论如下: 1) Semantic-LPA 算法的最优参数取值为 $\sigma \in (1,2)$, 且所提出的 $p\text{-}SCNP$ 的实验效果优于 $SCNP$ 和 $weighted\text{-}SCNP$; 2) SQ 相对于 EQ 更适合评价语义社区划分结果; 3) 在面向具有语义关系的社区划分问题时, Semantic-LPA 相对于经典重叠社区发现算法更有效; 4) Semantic-LPA 相对于 TURCM、CART、CUT、LCTA 等语义社区划分算法有效性更高.

6 结论

本文针对语义社会网络社区划分的问题, 提出 Semantic-LPA 算法, 该算法将语义社会网络的语义特性和社会关系特性相融合, 可有效解决语义社会网络中的重叠社区发现问题. 本文的工作在于: 1) 利用 Gibbs 取样法构建语义空间, 并将节点的语义信息映射为语义空间内的坐标, 实现了节点的语义信息可度量化; 2) 利用节点的语义坐标、链接关系及 PCA 思想提出了 $SCNP$ ($weighted\text{-}SCNP$, $p\text{-}SCNP$), 以度量节点间的相关性; 3) 利用数据场模型构建了语义影响力模型 SI, 并以 SI 作为截断值, 设计了面向语义重叠社区划分的标签传播算法; 4) 提出了评价语义社区划分结果的 SQ 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, Semantic-LPA 算法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. 所提出的 SQ 相对于 EQ 更适合评价语义社区划分结果. 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一种研究方法, 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. Complex network clustering algorithms. *Journal of Software*, 2009, **20**(1): 54−66
(杨博, 刘大有. 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, **9**(12): 7921−7826

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 structures of complex networks in nature and societ. *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*, 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

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. Markov random walk under constraint for discovering overlapping communities in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011, P05031

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. 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. 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 Intelligence and Security Informatics. New Brunswick, NJ: IEEE, 2007. 200−207

16 Kemp C, Tenenbaum J B, Griffiths T L, Yamada Y, Ueda N. Learning systems of concepts with an infinite relational model. In: Proceedings of the 21st National Conference on Artificial Intelligence. Boston, MA: 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. New York: ACM, 2009. 1456−1461

18 Henderson K, Eliassi-Rad T, Papadimitriou S, Faloutsos C. HCDF: A hybrid community discovery framework. In: Proceedings of the 2010 SIAM International Conference on Data Mining. Columbus, OH: SIAM, 2010. 754−765

19 Zhang H, 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. Boston, MA: AAAI Press, 2007, **7**: 663−668

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

21 Steyvers M, Smyth P, Rosen-Zvi M, Groffiths 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: ACM, 2004. 306−315

22 McCallum A, Corrada-Emmanuel A, Wang X R. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, 2005. 3

23 McCallum A, Wang X, 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, Lee C L, Zha H Y. Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web. New York: 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: 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: 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. 8

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

29 Sachan M, Contractor D, Faruquie T, 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: ACM, 2011. 2349−2352

30 Sachan M, Contractor D, Faruquie T, Subramaniam V. Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web. New York: 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): Article No. 63, DOI: 10.1145/2337542.2337548

32 Zhang Y Z, Wang J Y, Wang Y, Zhou L Z. Parallel community detection on large networks with propinquity dynamics. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 997−1006

33 Lou H, Li S H, Zhao Y X. Detecting community structure using label propagation with weighted coherent neighborhood propinquity. *Physica A: Statistical Mechanics and Its Applications*, 2013, **392**(14): 3095−3105

34 Zhu X J Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning (ICML-2003). Piscataway, N J: IEEE, 2003. 912−919

35 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. Washington, CD: 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.)