中国科学 信息科学
SCIENTIA SINICA Informationis

基于数据场分析的语义重叠社区发现算法

辛宇, 杨静 and 谢志强

Citation: 中国科学: 信息科学 **45**, 918 (2015); doi: 10.1360/N112014-00099-15

View online: http://engine.scichina.com/doi/10.1360/N112014-00099-15

View Table of Contents:http://engine.scichina.com/publisher/scp/journal/SSI/45/7

Published by the 《中国科学》杂志社

中国科学: 信息科学 2015 年 第 45 卷 第 7 期: 918–933
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

基于数据场分析的语义重叠社区发现算法

辛宇[①], 杨静[①*], 谢志强[②]

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
* 通信作者. E-mail: yangjing@hrbeu.edu.cn

收稿日期: 2014–08–05; 接受日期: 2014–10–07; 网络出版日期: 2015–04–16
国家自然科学基金 (批准号: 61370083, 61073043, 61073041, 61370086) 和高等学校博士学科点专项科研基金 (批准号: 20112304110011, 20122304110012) 资助项目

摘要 语义社会网络是一种由信息节点及社会关系构成的新型复杂网络, 而传统社会网络社区发现算法以节点邻接关系为挖掘对象, 因此无法有效处理语义社会网络重叠社区发现问题. 针对这一问题, 提出基于语义数据场的语义重叠社区发现算法, 该算法首先以 LDA (latent dirichlet allocation) 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; 其次, 利用节点间语义坐标及链接关系, 建立节点的语义数据场模型; 再次, 以语义关系强度及语义势能为参数, 提出一种改进的语义社会网络重叠社区发现的随机游走策略; 最后提出可度量语义社区发现结果的语义模块度模型. 通过实验分析, 验证了本文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 语义数据场 随机游走

1 引言

随着网络通讯的发展, 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 3 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属于关系社区划分, 其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 [4] 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. 此后, 重叠社区发现算法成为社区划分研究领域的主流, 许多经典算法应运而生, 如 EAGLE[5], LFM[6], COPRA[7], UEOC[8], 蚁群算法 [9], 拓扑势算法 [10] 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、社会标签等), 将具有相似信息内容的节点划分为同一社区. 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. 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 模型 [11] 作为语义处理的核心模型. 根据 LDA 模型的应用方式可分为 3 类.

引用格式: 辛宇, 杨静, 谢志强. 基于数据场分析的语义重叠社区发现算法. 中国科学: 信息科学, 2015, 45: 918–933, doi: 10.1360/N112014-00099

(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 [12] 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. 由于 Kemp 等 [13] 在 SSN-LDA 模型的基础上融入了 IRM (infinite relational models) 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 [14]. 随后 Henderson 等 [15] 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. Zhang 等 [16] 也在 SSN-LDA 算法的基础上提出了面向有权网络的 GWN-LDA 算法及面向层次划分的 HSN-PAM[17] 算法. Jang 等 [18] 以舆论领袖的影响力及影响范围决定社区大小及紧密性为指导思想, 首先利用 LDA 模型将语义社会网络中的文本信息进行统一挖掘, 并将其挖掘结果作为社会个体的得分, 其次通过对得分及网络相关性的综合评价, 确定舆论领袖及其引领的话题社区. 此类算法的优点在于结构模型简单, 需要的信息量较少, 适合处理大规模数据, 缺点在于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 [19] 作为 LDA 分析的基本模型, 代表算法有 McCallum 等 [20] 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. 随后 McCallum 等 [21] 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 模型在社会计算领域的应用. Zhou 等 [22] 在 AT 模型中加入了 user 分布取样, 提出了 CUT 模型. Cha 等 [23] 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. Natarajan 等 [24] 和 Evans 等 [25] 以 link communit 为切入点, 建立了以 link-content 为语义分析目标的 LCM (link-content model) 模型. 该模型以用户节点间的共享信息及用户节点间的传递信息作为 link-content, 在进行 LDA 建模时将 link 作为社区分类的基本单元. 此类算法的优点在于, 在节点关系基础上结合了本文信息分析, 其划分的社区具有较高的内部相似性, 缺点在于, 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 [26] 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. 随后 Pathak 等 [27] 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. 近些年来, 话题 — 社区的关系成为 LDA 模型研究的重点, Mei 等 [28] 将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. Sachan 等 [29,30] 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 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. 此类算法的优点在于, 语义社区划分准确性高. 缺点在于, 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.

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: (i) 语义社区内部链接关系紧密; (ii) 语义社区内部节点

表 1　数学符号说明

Table 1　The notation description

Notation	Description
G	The global network, G_i is the node i in the network
$\|G\|$	The number of nodes in semantic social network
N	The number of keywords in semantic social network, N_i is the number of keywords in G_i
w	The collection of keywords, w_i is the ID of the ith keyword in w
d	The node ID collection corresponding to the keywords collection w
z	The topic ID collection corresponding to the keywords collection w
$\theta^{(d_i)}$	The topic distribution probability of node d_i
$\lambda^{(j)}$	The keyword distribution probability of topic j, $\lambda_{w_i}^{(j)}$ representing the probability of keyword w_i belonging to topic j, $\lambda_{w_i}^{(j)} = P(w_i \mid z_i = j)$
α	The priori argument of topic distribution probability of each node
β	The priori argument of keyword distribution probability in a special topic

的语义信息相似度高. 为此本文所设计的面向语义社会网络的定义重叠社区发现算法, 建立节点语义信息到语义空间的量化映射, 通过构建节点语义关系强度及语义势能模型, 建立基于语义数据场的随机游走策略, 实现重叠社区发现, 并提出了可评价语义社区划分结果的 SQ 模型, 最后通过实验, 分析本文算法参数并验证有效性.

2　语义社会网络的 LDA 关系建模

2.1　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. 本节内容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, 所涉及的数学符号如表 1 所示.

LDA 语义数据分别利用 w, d, z 3 个集合进行存储, 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编号, 图 1 为 LDA 算法的 w, d, 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}.

从图 1 的分析可知, w, d, z 3 者之间存在 3 层贝叶斯 (Bayes) 关系, 根据文献 [11] 的文本分析可知, w, d, z 的数学关系式如下:

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

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

(3) $\lambda \sim \text{Dirichlet}(\beta)$, 关键字分布 λ 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 λ 下, 关键字 w_i 出现的概率服从多项式分布.

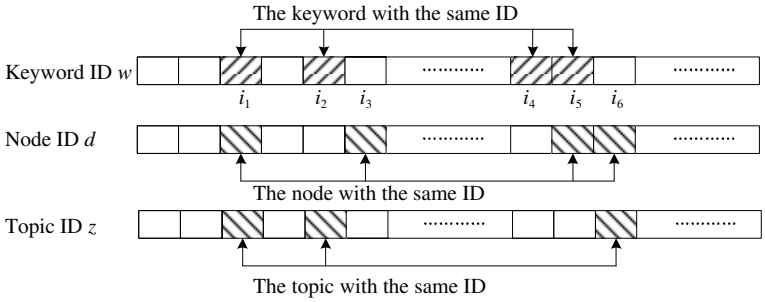

图 1 w, d, z 数据存储结构

Figure 1 The data storage structure of w, d, z

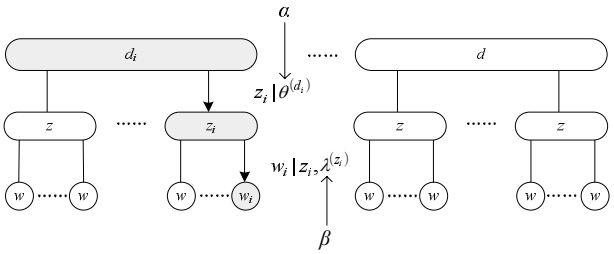

图 2 w, d, z 的贝叶斯关系图

Figure 2 The Bayesian relationship of w, d, z

图 2 为关键字 w, d, 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.

2.2 Gibbs 迭代过程

w, z 的贝叶斯关系表达式为

$$P(z_i = j|w_i)\,P(w_i) = P(w_i|z_i = j)\,P(z_i = j)\,, \tag{1}$$

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为条件, 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 z_{-i} 和 w_{-i} (表示除去元素 i 的集合 z 和 w), 分别作为推断 z_i 和 w_i 的条件, 因此式 (1) 可变形为

$$P(z_i = j|z_{-i}, w_i)\,P(w_i|w_{-i}) = P(w_i|z_i = j, z_{-i}, w_{-i})\,P(z_i = j|z_{-i})\,, \tag{2}$$

$$\Rightarrow P(z_i = j|z_{-i}, w_i) \propto P(w_i|z_i = j, z_{-i}, w_{-i})\,P(z_i = j|z_{-i})\,. \tag{3}$$

根据文献 [13], 式 (3) 的右边分为

$$P(w_i|z_i = j, z_{-i}, w_{-i}) = \frac{f_{-i,j}^{(w_i)} + \beta}{f_{-i,j}^{(\cdot)} + |w|\,\beta}, \tag{4}$$

$$P(z_i = j|z_{-i}) = \int P\left(zi = j|\theta^{(d_i)}\right) P\left(\theta^{(d_i)}|z_{-i}\right) \mathrm{d}\theta^{(d_i)} = \frac{f_{-i,j}^{(d_i)} + \alpha}{f_{-i,\cdot}^{(d_i)} + |z|\,\alpha}, \tag{5}$$

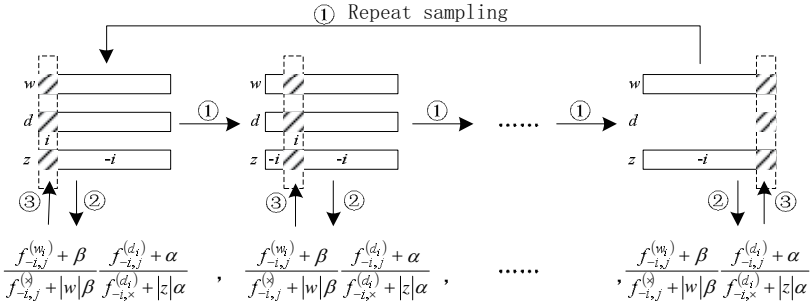

图 3 Gibbs 取样过程

Figure 3 The Gibbs sampling process

$$\Rightarrow P\left(z_i=j|z_{-i},w_i\right)\propto\frac{f_{-i,j}^{(w_i)}+\beta}{f_{-i,j}^{(\cdot)}+|w|\,\beta}\frac{f_{-i,j}^{(d_i)}+\alpha}{f_{-i,\cdot}^{(d_i)}+|z|\,\alpha},\tag{6}$$

其中, $|w|$ 和 $|z|$ 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①为循环取样过程; 箭头②为根据当前样本通过式 (6) 计算 $P(z_i=j|z_{-i},w_i)$; 箭头③根据 $P(z_i=j|z_{-i},w_i)$ 对 z_i 对进行 Gibbs 取样并修改 z_i 的过程. 当 $P(z_i=j|z_{-i},w_i)$ 收敛时则结束此过程, 并将 $P(z_i=j|z_{-i},w_i)$ 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 $B_{i,j}=p(z_i=j|w=i)$.

3 节点的语义量化映射

由于 Gibbs 取样过程所提炼的 k 个话题是以全局语义信息为基础, 因此本文以 LDA 模型提取的 k 个话题作为全局语义空间的 k 维基. $B_{i,\cdot}$ 表达了 i 号关键字对 k 个话题话题的隶属度, 因此向量 $B_{i,\cdot}$ 可表示为 i 号关键字在 k 维语义空间中的坐标, 由于各节点的语义内容由关键字构成, 因此某一节点 G_i 在语义空间中的坐标 (语义坐标) m_i 可通过 G_i 的 N_i 个关键字的加和均值形式表达, 其表达式为

$$m_i=\frac{\sum_{j=1}^{N_i}B_{N_{i,j},\cdot}}{N_i},\tag{7}$$

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.

网络 G 的拓扑可看作为 $|G|$ 个节点所构成的关系系统, 其中每个节点受与之相邻节点的影响, 且节点间影响力 (互相作用) 的大小与节点的距离成反比, 因此节点的语义环境受拓扑环境的影响. 根据文献 [10], 各节点间的相互用作构成了网络 G 的社会关系数据场, 为此本文以数据场的模型为基础, 设计了可表达节点语义环境的语义场模型. 根据式 (7) 语义社会网络中各节点的语义信息均可量化为坐标形式, 即语义坐标. 语义坐标可作为节点的质量, 因此根据社会网络数据场的势函数定义 [16], 任意节点 G_i 的语义场 (节点 G_i 的语义环境) 可表示为

$$F_i=\sum_{j=1}^{|G|}m_j\exp\left(-\left(\frac{\mathrm{dis}_{i,j}}{\sigma}\right)^2\right),\tag{8}$$

其中, $\mathrm{dis}_{i,j}$ 为节点 G_i 与节点 G_j 间的距离 (跳数), m_j 为节点 G_j 的语义坐标, σ 为影响因子用于控制每个节点的语义影响范围, 根据高斯函数的数学性质, 每个节点的影响范围近似为 $\lfloor 3\sigma/\sqrt{2}\rfloor$ 跳的局部

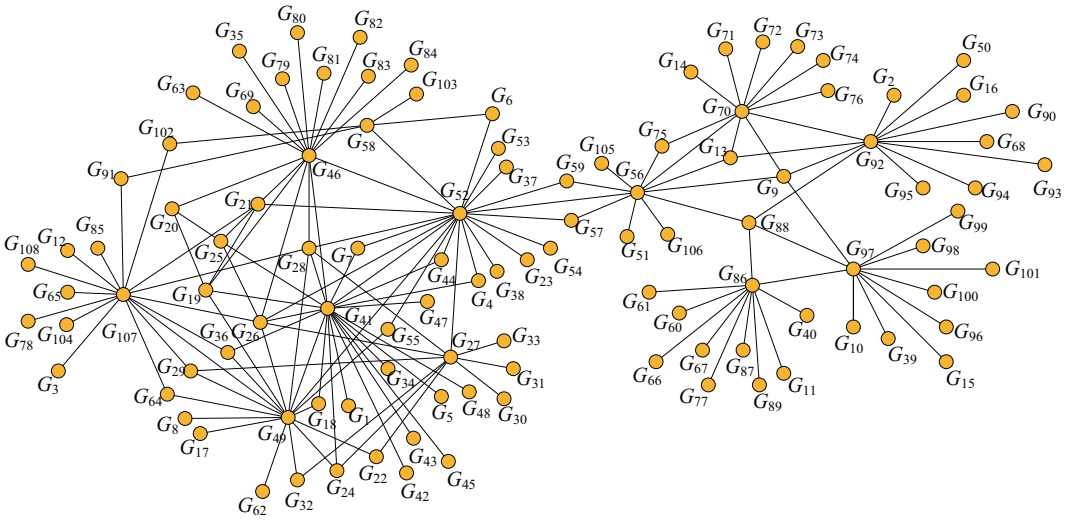

图 4 (网络版彩图) QLSP 网络拓扑

Figure 4 (Color online) The topology of QLSP

区域, 且根据文献 [10] 可知 σ 的最优取值区间为 (1,2). 因此根据式 (8) 中的语义场 F_i 量化表达了节点 G_i 在 3 跳范围内的局部语义环境. 在场势理论中势能表达了物体在当前环境 (场) 中的作用力, 其取值为物体质量与场之积. 为此本文将节点 G_i 的语义坐标 m_i 与其语义场 F_i 的内积作为节点 G_i 的语义势能 E_i. 语义势能 E_i 表达了节点 G_i 对 3 跳范围内的局部语义环境的相互作用力, 其表达式为

$$E_i = m_i \cdot F_i = \sum_{j=1}^{|G|} m_i \cdot m_j \exp\left(-\left(\frac{\mathrm{dis}_{i,j}}{\sigma}\right)^2\right). \tag{9}$$

由于内积可体现向量的相似度, 因此根据式 (9) 节点 G_i 的语义势能 E_i 越大, 则节点 G_i 在网络 G 中的语义影响力越大, 节点 G_i 与其邻近节点的相似度越高且结构越紧密, 即以 G_i 为核心的局部区域内部信息的相似度越高且社区结构越明显 (社区内部紧密度高于边缘紧密度). 因此, 节点 G_i 与其相邻节点 G_j 的语义关系强度 $H_{i,j}$ 可表现为节点 G_i 与其相邻节点 G_j 的语义场 (局部语义环境) F_j 的适应度. 由于语义场 F_j 体现了节点 G_j 的局部语义环境, 则 G_i 与 F_j 的适应度越高则 G_i 与 G_j 的局部语义环境越吻合, 因此可通过节点 G_i 的语义坐标 m_i 与其相邻节点 G_j 的语义场 F_j 的内积进行表达, 其表达为

$$H_{i,j} = m_i \cdot F_j = \sum_{t=1}^{|G|} m_i \cdot m_j \exp\left(-\left(\frac{\mathrm{dis}_{j,t}}{\sigma}\right)^2\right). \tag{10}$$

根据式 (9) 和 (10) 可知, 语义社会网络中各节点的语义势能及节点间的语义关系强度可用语义加权邻接矩阵 $\boldsymbol{H}$ 表示 ($\boldsymbol{H}$ 为非对称矩阵), 若 $H_{i,j} > H_{j,i}(H_{i,j} \neq 0)$ 表示节点 G_i 对语义场 (局部语义环境) F_j 的适应度强于节点 G_j 对语义场 (局部语义环境)F_i 的适应度, 从而使节点 G_j 受到节点 G_j 的吸引, 因此语义加权邻接矩阵 $\boldsymbol{H}$ 表达了网络节点间的吸纳关系.

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 篇论文和 155 条引用关系), 其网络模型如图 4 所示. 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 k 为 5 进行 Gibbs 取样迭代后, 再利

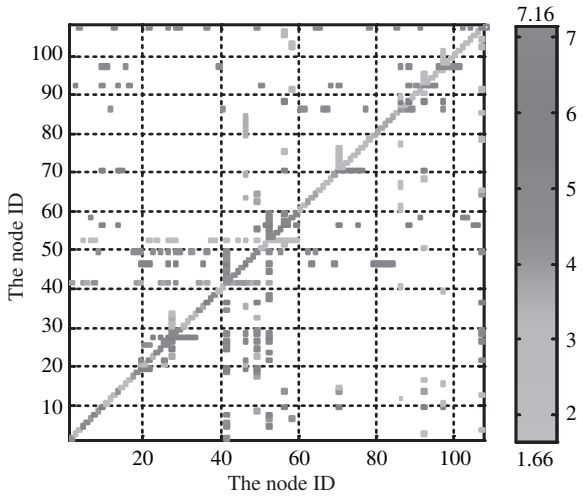

图 5　QLSP 的语义加权邻接矩阵 $\boldsymbol{H}$ 及语义势能 $\boldsymbol{E}$

Figure 5　The semantic weighted adjacency matrix of $\boldsymbol{H}$ and $\boldsymbol{E}$

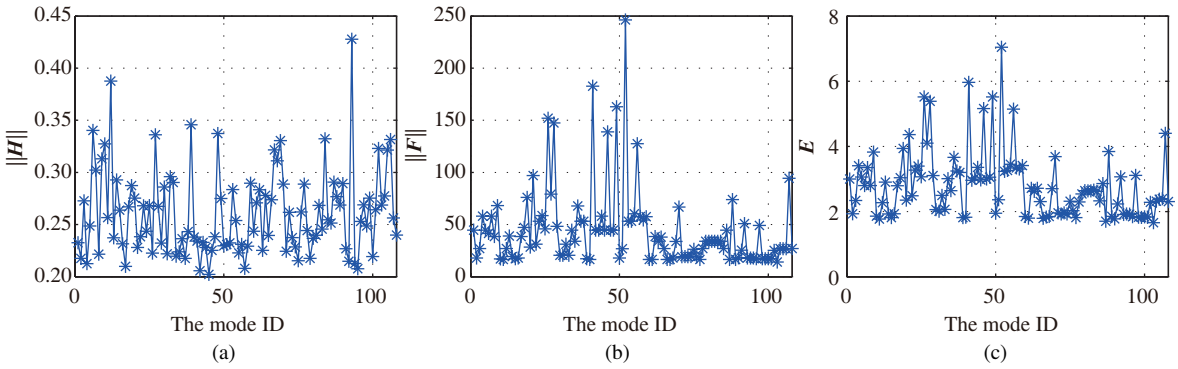

图 6　(网络版彩图) $||\boldsymbol{m}||$, $||\boldsymbol{F}||$ 和语义势能 $\boldsymbol{E}$ 的对比图

Figure 6　(Color online) The comparison of $||\boldsymbol{m}||$, $||\boldsymbol{F}||$ and $\boldsymbol{E}$. (a) The semantic coordinates mold $||\boldsymbol{H}||$; (b) the semantic field mold $||\boldsymbol{F}||$; (c) the semantic potential $\boldsymbol{E}$

用式 (9,10) 语义加权邻接矩阵 $\boldsymbol{H}$ 结果如图 5 所示, 其中对角线为语义势能 $\boldsymbol{E}$. 图 6 为各点节点语义坐标 $\boldsymbol{m}$ 模值 $||\boldsymbol{m}||$、语义场 $\boldsymbol{F}$ 模值 $||\boldsymbol{H}||$ 和语义势能 $\boldsymbol{E}$ 的对比图.

4　语义重叠社区发现的随机游走策略

随机游走策略是语义社区发现的经典策略, 该策略以一步转移概率矩阵为基础, 计算以 s 为出发点的 l 步抵达概率分布, 并将概率分布较高的节点作为 s 社区 (s 所在社区) 的内部节点 [8]. 由上一节的分析可知, 语义社会网络拓扑结构可通过加权邻接矩阵 $\boldsymbol{H}$ 和语义势能 $\boldsymbol{E}$ 进行表示, 因此, 为实现语义社会网络的社区发现, 可对随机游走策略进行 3 方面的改进:

(1) 设定一步转移概率矩阵, 由于加权邻接矩阵 $\boldsymbol{H}$ 表达了语义社会网络节点间的吸纳关系, $H_{i,j}$ 越大则 G_j 相对于节点 G_i 的吸引力越强, 因此节点 G_i 加入 G_j 所在社区的概率越大. 又由于随机游走策略的一步转移概率矩阵 $\boldsymbol{S}$ 中, $S_{i,j}$ 的值表示从节点 G_i 游走到 G_j 的概率, 因此 $S_{i,j}$ 的值越大节

点 G_i 加入 G_j 所在社区的概率越大. 综合加权邻接矩阵 $\boldsymbol{H}$ 和一步转移概率矩阵 $\boldsymbol{S}$ 的特性, 可将行向量归一化后的加权邻接矩阵 $\boldsymbol{H}$ 作为随机游走策略的一步转移概率矩阵 $\boldsymbol{S}$, 其表达式为

$$S_{ij} = \frac{H_{ij}}{\sum_{r=1}^{|G|} H_{ir}}. \tag{11}$$

假设 l 步抵达概率分布为 ε_s^l, 其中 $\varepsilon_s^l(i)$ 表示一个 agent 从节点 s 出发, 经过 l 次转移后最终到达节点 G_i 的概率, 则 $\varepsilon_s^l(i)$ 可通过迭代式 (12) 进行表达

$$\varepsilon_s^l(i) = \sum_{r=1}^{|G|} \varepsilon_s^{l-1}(r) S_{ri}. \tag{12}$$

(2) 设定局部扩展节点, 由于语义势能 E_i 表达了节点 G_i 在网络 G 中的语义影响力, 且在随机游走策略中局部扩展节点是网络中影响力最大的结点, 因此可选择语义势能 $\boldsymbol{E}$ 最高的节点作为局部扩展节点.

(3) 设定截断策略, 为减少算法的复杂度本文采用所有节点的平均概率值 $\kappa = \sum_{i=1}^{|G|} \varepsilon_s^l(i)/G_n$ 作为截断阈值, 将 $\varepsilon_s^l > \kappa$ 的节点 G_i 与 s 划分为同一社区即为 s 初始社区.

根据文献 [8] 可知, 若利用归一化后的邻接矩阵作为一步转移概率矩阵, 在进行随机游走策略进行社区发现时, l 取值的不同会导致各社区结构进入局部混合状态 (出现社区结构) 时间的不同, 所发现的社区结果波动较大. 随着 l 取值的增加, 网络中节点度数较大的节点会逐渐被划归为同一社区, 导致社区划分结果的效果逐渐劣化. 其原因在于上述策略中, 同一节点的游出概率均为该节点度数的倒数, 由于缺少对局部适应度的考虑, 导致度数高的节点在概率分布 ε_s^l 中具有相对优势. 本文算法利用了行向量归一化后的加权邻接矩阵 $\boldsymbol{H}$ 作为一步转移概率矩阵, 使节点的游出概率取决于该节点对相邻节点的局部语义环境的适应度, 提高了与出发节点 s 相适应的节点在概率分布 ε_s^l 中比重, 当 l 取值增加时社区结构稳定且同一社区内部节点的适应度较高. 本文在实验章节分析了算法对 l 取值的稳定性. 为得出稳定的社区结构, 本文首先建立社区集合 C, 并在截断策略结束后, 首先在 s 初始社区找出 s 的最大连通子图作为 s 社区, 再判断 s 社区与 C 中各社区的相似度 (节点重复度), 若 s 社区与 C 中某一元素 (s' 社区) 的重复度超过 η, 则合并 s 社区和 s' 社区, 否则将 s 加入 C. 本文在实验章节中分析了 l 和 η 的取值. 由于分别建立了随机游走策略所需的一步转移概率矩阵和局部扩展节点, 可根据随机游走策略算法框架设计语义重叠社区发现策略, 其流程描述如下:

(1) 建立社区集合 C 以保存划分出的社区;

(2) 选择未被划分社区的节点中语义势能 $\boldsymbol{E}$ 最大的节点 s 作为局部扩展节点;

(3) 利用马尔可夫 (Markov) 动力学方法计算节点 s 的 l 步抵达概率分布 ε_s^l;

(4) 采用截断策略, 抽取局部扩展节点 s 的初始社区结构;

(5) 在 s 初始社区中找出 s 的最大连通子图作为 s 社区;

(6) 判断 s 社区与 C 中各社区的相似度后更新社区集合 C;

(7) 如果仍存在未划分社区的节点则转 (2), 否则结束.

本文算法的整体流程如图 7 所示. 利用本文算法对 QLSP 数据集进行社区划分的结果如图 8 所示 (其中参数为 $l = 10, \eta = 40\%$).

5 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关系结构为输入, 文献 [5] 所建立的重叠社区模块度 EQ 模

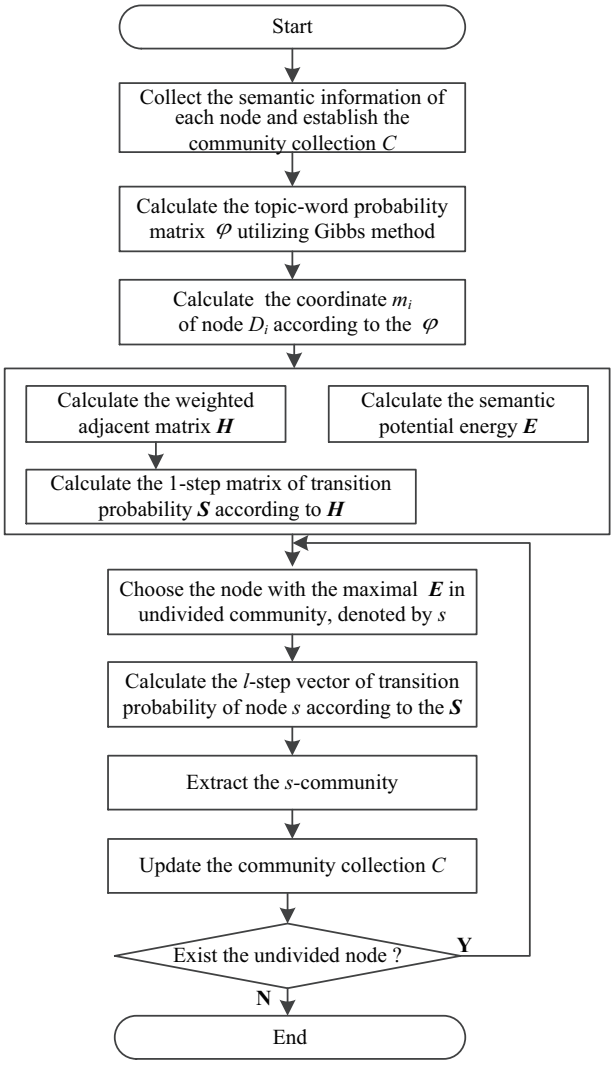

图 7　算法流程图

Figure 7　Algorithm flow chart

型为

$$\mathrm{EQ}=\frac{1}{X}\sum_{t}\sum_{i\in C_t,j\in C_t}\frac{1}{O_iO_j}\left[A_{i,j}-\frac{R_iR_j}{X}\right], \tag{13}$$

其中, R_i 为节点 d_i 的度数, X 为网络节点的总度数, $\boldsymbol{A}$ 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 m_i 为输入的语义信息相似性度量函数 $U(m_i,m_j)$, 建立可评价标语义重叠社区的模块度模型 SQ, 其表达式为

$$\mathrm{SQ}=\frac{1}{X}\sum_{t}\sum_{i\in C_t,j\in C_t}\frac{U\left(m_i,m_j\right)}{O_iO_j}\left[A_{i,j}-\frac{R_iR_j}{X}\right]. \tag{14}$$

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 $U(m_i,m_j)$, 其表

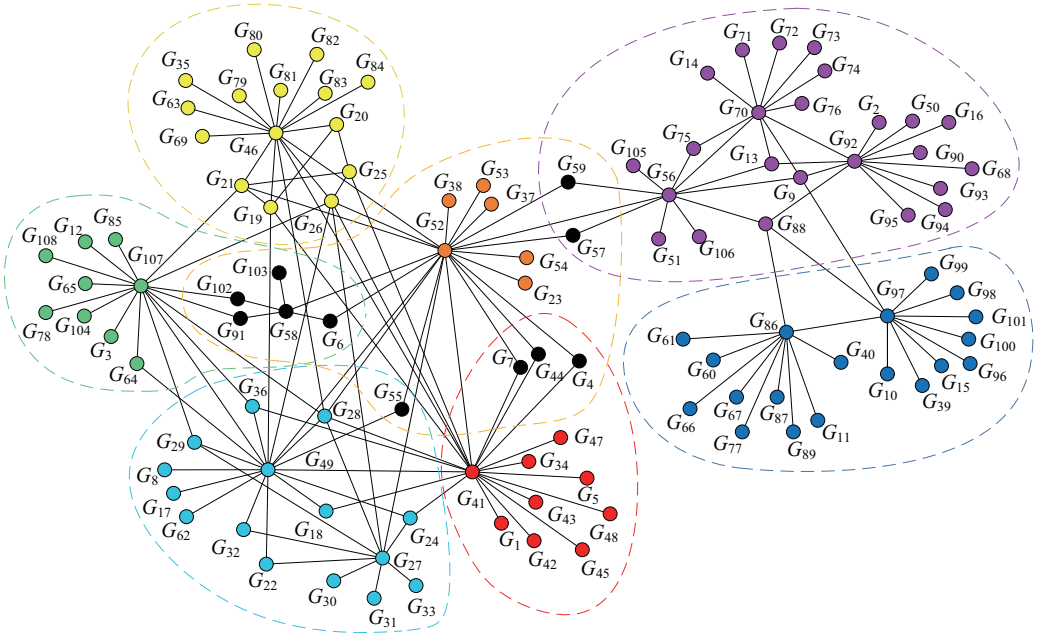

图 8 (网络版彩图) QLSP 网络语义社区划分结果

Figure 8 (Color online) The detected communities of QLSP network

达式为

$$U(m_i, m_j) = \frac{m_i \cdot m_j}{|m_i|\,|m_j|} = \frac{\sum_{g=1}^{k} m_{i,g} m_{j,g}}{\sqrt{\sum_{g=1}^{k} {m_{i,g}}^2 \sum_{g=1}^{k} {m_{j,g}}^2}}. \tag{15}$$

本文在实验章节对 SQ 进行了实验分析.

6 实验分析

6.1 随机游走步长 l 和社区的重复度 η 的取值分析

随机游走步长 l 和社区重复度 η 是本文算法 SFR(semantic field randwalk) 的输入参数, 为验证参数 l 和 η 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) 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 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) Newman 建立的海豚家族 (dolphins network) 关系网络, 该网络由两大家族组成员个数分别为 20 和 42, 共 159 条链接关系, 为模拟语义社会网络的特性, 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) 进行参数分别为 l (2 $\sim$ 12) 和 η (20% $\sim$ 90%) 的测试, 3 组数据的 SQ 函数取值结果如图 9 所示, 3 组数据的最大 SQ 值分别为 0.4173, 0.4165, 0.4474, EQ 分别为 0.4359, 0.4351, 0.4738. 对应的参数 l 和 η 取值分别为 ($l = 8, \eta =$

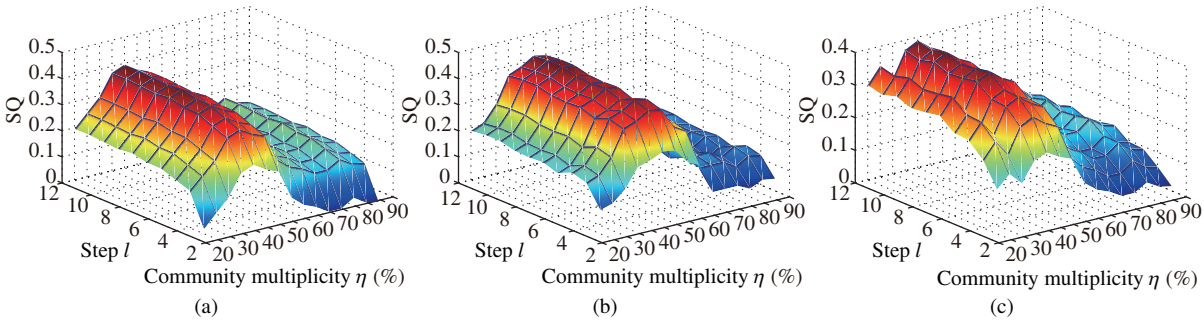

图 9 (网络版彩图) 3 组数据的 SQ 值对比图

Figure 9 (Color online) The comparison of SQ on the three datasets. (a) QLSP dataset; (b) polbooks dataset; (c) dolphins dataset

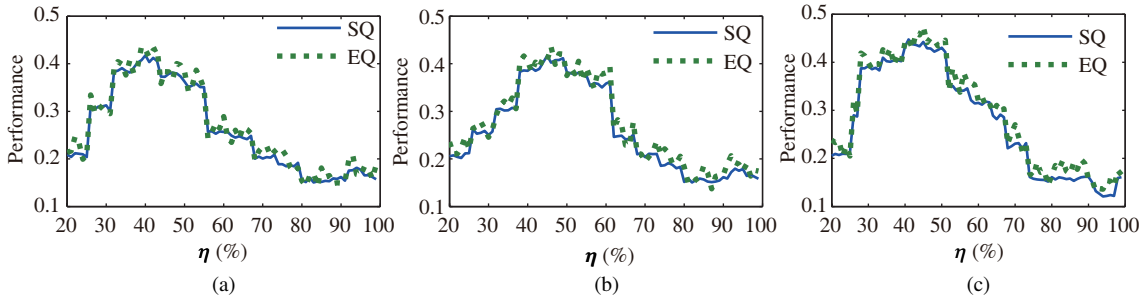

图 10 (网络版彩图) EQ 和 SQ 对比结果

Figure 10 (Color online) The comparison of EQ and SQ. (a) QLSP dataset; (b) polbooks dataset; (c) dolphins dataset

41%), $(l = 10, \eta = 46\%)$, $(l = 12, \eta = 42\%)$. 从图 9 所示的对比结果可分析出: (1) 当 η 取值固定且 l 大于 4 时各组数据的 SQ 值趋于稳定且结果趋于最优, 其原因在于当 l 大于 4 时已进行了充分游走, 此时 l 的取值对社区划分结果的影响较小, 说明了本文算法克服了随机游走的波动性问题; (2) 社区重复度 η 对本文 SFR 算法结果的影响较大, 且 η 在 35% ~ 45% 内为最佳.

6.2 SQ 与 EQ 的比较分析

本节实验计算了实验 1 中, 3 组数据 η 不同取值条件下的 EQ 值, 3 组数据的 EQ 和 SQ 值的对比如图 10 所示. 从式 (13) 和式 (14) 的对比可知, SQ 加入了语义信息相似性度量函数 U 且 $U(m_i, m_j) < 1$, 使得 SQ 的总体趋势小于 EQ. QLSP 数据集的 EQ 最大值为 0.4359 所对应的 $\eta = 40\%$, 与 SQ 的最大值为 0.4173 所对应的 $\eta = 41\%$ 存在偏差, 而 $\eta = 41\%$ 所对应的社区结构的 EQ 值为 0.4236. 对于 polbooks 和 dolphins 数据集的 EQ 最大值分为 0.4351($\eta = 46\%$) 和 0.4738($\eta = 42\%$), 与 SQ 的最大值 0.4165($\eta = 47\%$) 和 0.4474($\eta = 46\%$) 同样存在偏差.

6.3 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, LPA, UEOC, EAGLE, CPM. 其中 LFM, LPA, 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,

表 2 经典算法的 SQ 和 EQ 值

Table 2 The SQ and EQ on classical algorithms

Algorithm	SQ	EQ
GN	0.3584	0.4617
FN	0.3157	0.4061
LFM	0.2329	0.3254
LPA	0.4003	0.541
UEOC	0.4071	0.441
SFR	0.4173	0.4359

FN, LFM, LPA, UEOC 算法进行求解, 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 (0.5236) 高于本文算法 (0.4359) 但 SQ 值均低于本文算法 (0.4173), 由此验证了, 传统面向链接关系的社区划分算法的 EQ 较高, 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.

6.4 多种数据集比较分析

本实验以清华大学 ArnetMiner 系统的 QLSP 完整数据集 (共 805 个节点), Aminer-FOAF-DataSet (AFD) 数据集 (截取 2000 个节点), Citation Network Dataset (CND) 数据集 (共同 2555 个节点), DBLP (April 12, 2006) 数据集 (1200000 个节点) 中分别截取 (A) 1500 个节点数据集和 (B) 2000 个节点数据集作为实验数据. 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 SFR 算法的运行参数为 $\eta = 40\%$, $l = 10$, 表 3 包括 EQ, SQ 及社区个数 CS.

6.5 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 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 和表 5 分别为 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 (CS) 表示各算法执行前的社区预设数. 从表 4 与表 5 的分析可知, Enron 数据集的最佳个数为 10. 本文算法的社区个数为 11, EQ 和 SQ 取值分别为 0.322 和 0.308. 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.

6.6 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、多数据集分析 4 个方面进行分析, 所得出的结论如下:

(1) SFR 算法的最优参数取值为 $\eta = [35\%, 45\%]$ 且克服了 l 取值对随机游走结果的波动性;

(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, SFR 相对于经典重叠社区发现算法更有效;

(4) SFR 对于各类语义社会网络具有普遍适用性;

(5) 相较于各类语义社区发现算法, SFR 算法无需预设社区个数且结果较好.

表 3　各数据集的 SQ, EQ 及 CS

Table 3　The SQ, EQ and CS on each dataset

Algorithm		QLSP	AFD	CND	DBLP(A)	DBLP(B)
GN	EQ	0.3108	0.1325	0.1928	0.2823	0.3192
	SQ	0.231	0.1597	0.1891	0.2139	0.2865
	CS	10	25	39	17	16
FN	EQ	0.4216	0.1525	0.2235	0.3191	0.2618
	SQ	0.3216	0.1392	0.1721	0.2916	0.2561
	CS	10	27	37	19	16
LFM	EQ	0.3668	0.1473	0.2406	0.4052	0.3641
	SQ	0.3172	0.1321	0.2172	0.3317	0.3133
	CS	12	24	33	22	12
LPA	EQ	0.4198	0.3186	0.1119	0.383	0.4113
	SQ	0.2891	0.2177	0.1202	0.2971	0.3217
	CS	13	21	35	21	13
UEOC	EQ	0.3849	0.2312	0.2648	0.3658	0.3183
	SQ	0.3177	0.2218	0.2271	0.2964	0.2011
	CS	12	24	30	22	14
SFR	EQ	0.3236	0.2415	0.2125	0.3443	0.3114
	SQ	0.3532	0.2713	0.2754	0.3592	0.3647
	CS	13	26	33	25	16

表 4　各类语义社区发现算法的 EQ 值

Table 4　The EQ on each semantic community detection algorithm

Algorithm	CS=6	CS=8	CS=10	CS=12	CS=14
TURCM	0.198	0.271	0.339	0.331	0.283
CART	0.152	0.249	0.302	0.294	0.255
CUT	0.133	0.231	0.266	0.278	0.227
LCTA	0.164	0.239	0.278	0.311	0.249

表 5　各类语义社区发现算法的 SQ 值

Table 5　The SQ on each semantic community detection algorithm

Algorithm	CS=6	CS=8	CS=10	CS=12	CS=14
TURCM	0.173	0.231	0.281	0.31	0.261
CART	0.122	0.226	0.256	0.268	0.226
CUT	0.126	0.215	0.233	0.235	0.202
LCTA	0.161	0.208	0.243	0.279	0.215

7　结论

本文针对语义社会网络社区划分的问题, 提出了 SFR 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, 可有效解决语义社会网络中的重叠社区发现问题.

本文算法设计的创新思想在于 (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, 实现节点的语义信息可度量化; (2) 利用节点的语义坐标及链接关系, 建立了节点的语义数据场模型, 并据此构建了节点语义关系强度及语义势能模型; (3) 利用语义关系强度及语义势能模型构建了面向语义重叠社区发现的随机游走策略; (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, SFR 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.

参考文献

1 Yang B, Liu D Y, Jin D, et al. Complex network clustering algorithms. J Softw, 2009, 20: 54–66 [杨博, 刘大有, 金弟, 等. 复杂网络聚类方法. 软件学报, 2009, 20: 54–66]

2 Girvan M, Newman M E J. Community structure in social and biological networks. P Nat Acad Sci, 2002, 99: 7821–7826

3 Newman M E J. Fast algorithm for detecting community structure in networks. Phys Rev E, 2004, 69: 066133

4 Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structures of complex networks in nature and societ. Nature, 2005, 435: 814–818

5 Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks. Phys A, 2009, 388: 1706–1712

6 Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks. New J Phys, 2009, 11: 033015

7 Gregory S. Finding overlapping communities in networks by label propagation. New J Phys, 2010, 12: 103018

8 Jin D, Yang B, Baquero C, et al. A Markov random walk under constraint for discovering overlapping communities in complex networks. J Stat Mech Theory Exp, 2011, 2001: P05031

9 Jin D, Yang B, Liu J, et al. Ant colony optimization based on random walk for community detection in complex networks. J Softw, 2012, 23: 451–464 [金弟, 杨博, 刘杰, 等. 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: 451–464]

10 Gan W Y, He N, Li D Y. Community discovery method in networks based on topological potential. J Softw, 2009, 20: 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: 2241–2254]

11 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022

12 Zhang H, Qiu B, Giles C L, et al. An LDA-based community structure discovery approach for large-scale social networks. Intell Secur Inform, 2007, 31: 200–207

13 Kemp C, Tenenbaum J B, Griffiths T L, et al. Learning systems of concepts with an infinite relational model. In: Proceedings of the AAAI Conference on Artificial Intelligence, Boston, 2006, 5

14 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, 2009. 1456–1461

15 Henderson K, Eliassi R T, Papadimitriou S, et al. HCDF: a hybrid community discovery framework. In: Proceedings of SIAM International Conference on Data Mining, Columbus, 2010. 754–765

16 Zhang H, Giles C L, Foley H C, et al. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of the AAAI Conference on Artificial Intelligence, Arlington, 2007. 7: 663–668

17 Zhang H, Li W, Wang X, et al. HSN-PAM: finding hierarchical probabilistic groups from large-scale networks. In: Proceedings of 7th IEEE International Conference on Data Mining Workshops, Omaha, 2007. 27–32

18 Jang J, Myaeng S H. Discovering dedicators with topic-based semantic social networks. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, Boston, 2013. 1–12

19 Steyvers M, Smyth P, Rosen Z M, et al. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, 2004. 306–315

20 McCallum A, Corrada E A, Wang X. Topic and role discovery in social networks. J Artif Intell Res, 2006, 29: 139–152

21 McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email. J Artif Intell Res, 2007, 30: 249–272
22 Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, 2006. 173–182
23 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, Portland, 2012. 565–574
24 Nagarajan N, Sen P, Chaoji V. Community detection in content-sharing social networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, 2013. 82–89
25 Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities. Phys Rev E, 2009, 80: 016105
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, Chicago, 2005. 28–35
27 Pathak N, DeLong C, Banerjee A, et al. Social topic models for community extraction. In: Proceedings of the 2nd SNA-KDD Workshop, Las Vegas, 2008. 1–8
28 Mei Q, Cai D, Zhang D, et al. Topic modeling with network regularization. In: Proceedings of the 17th International Conference on World Wide Web, Beijing, 2008. 101–110
29 Sachan M, Contractor D, Faruquie T, et al. Probabilistic model for discovering topic based communities in social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, 2011. 2349–2352
30 Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web, Lyon, 2012. 331–340
31 Yin Z, Cao L, Gu Q, et al. Latent community topic analysis: integration of community discovery with topic modeling. ACM Trans Intell Syst Technol, 2012, 3: 63–70

A semantic overlapping community detection algorithm based on semantic data fields

XIN Yu[1], YANG Jing[1*] & XIE ZhiQiang[2]

1 *College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001*, China*;
2 *College of Computer Science and Technology, Harbin University of Science and Technology, Harbin* 150080*, China*
*E-mail: yangjing@hrbeu.edu.cn

Abstract Semantic social networks (SSNs) represent a new type of complex network; consequently, they cannot be analyzed efficiently by traditional community detection algorithms that depend on social network adjacency. To solve this problem, an overlapping community structure-detecting method for SSNs is proposed, and is based on semantic data fields. First, the paper proposes an algorithm that utilizes Gibbs sampling to create the quantization mapping that enables semantic information in nodes to be moved into semantic space, using LDA (latent dirichlet allocation) as the semantic model. Second, it establishes a semantic data field model, using the semantic coordinates and link relationships of nodes. Third, it proposes an improved random walk strategy that employs an overlapping community structure-detecting algorithm for SSNs, using the semantic relationship strength and the semantic potential of nodes as parameters. Finally, it proposes the semantic model by which an SSN community structure can be measured. The efficiency, feasibility, and semantic modularity of the algorithm is verified by experimental analysis.

Keywords semantic social networks, overlapping community, LDA, semantic data field, random walk

XIN Yu was born in 1987. He received his Master's degree from the Harbin University of Science and Technology, Harbin. Currently, he is a Senior Researcher at SNA. His research interests include data and knowledge engineering, and enterprise intelligence computing.

XIE ZhiQiang was born in 1962. He holds a PhD, and is currently a professor and PhD student supervisor. His research interests include data and knowledge engineering, and enterprise intelligence computing.

YANG Jing was born in 1962. She holds a PhD, and is currently a professor and PhD student supervisor. Her research interests include data and knowledge engineering, and enterprise intelligence computing.