计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131246
52(2):499-511,2015

基于随机游走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)
(yangjing@hrbeu.edu.cn)

A Semantic Overlapping Community Detecting Algorithm in Social Networks Based on Random Walk

Xin Yu[1], Yang Jing[1], and Xie Zhiqiang[2]
[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)

Abstract Since the semantic social networks (SSN) is a new kind of complex networks, the community detection is a new investigation relevant to the traditional community detection research. To solve this problem, an overlapping community structure detecting method in semantic social network is proposed based on the random walk strategy. The algorithm establishes the semantic space using latent Dirichlet allocation (LDA) method. Firstly, the quantization mapping is completed by which semantic information in nodes can be changed into the semantic space. Secondly, the semantic influence model and weighed adjacent matrix of SSN are established, with the entropy of nodes in SSN as the semantic information proportion, the distribution ratio of nodes as the weight of adjacent. Thirdly, an improved random walk strategy of community structure detecting in overlapping-SSN is proposed, with the distribution ratio of nodes as parameter, and a semantic modularity model is proposed by which the community structure of SSN can be measured. Finally, the efficiency and feasibility of the proposed algorithm and the semantic modularity are verified by experimental analysis.

Key words random walk; community detection; semantic social network; latent Dirichlet allocation; semantic modularity

摘 要 语义社会网络是由信息节点及社会关系构成的一类新型复杂网络,因此语义社会网络重叠社区发现是传统社区发现研究的新方向.针对这一问题,提出基于随机游走的语义社会网络重叠社区发现算法,该算法首先以LDA(latent Dirichlet allocation)算法为基础建立语义空间,实现节点语义信息到语义空间的量化映射;其次,以语义空间中节点信息熵作为节点语义信息比重,以节点的度分布比率作为节点关系比重,建立节点语义影响力模型及语义社会网络的加权邻接矩阵;再次,以语义影响力模型和加权邻接矩阵为参数,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,并提出可度量语义社区发现结果的语义模块度模型;最后,通过实验分析,验证了所提出的算法及语义模块度模型的有效性和可行性.

关键词 随机游走;社区发现;语义社会网络;LDA算法;语义模块度

中图法分类号 TP18;TP393

收稿日期:2013-09-12;**修回日期**:2014-03-11

基金项目:国家自然科学基金项目(61370083,61370086,61073043,61073041);教育部高等学校博士学科点专项科研基金项目(20112304110011,20122304110012)

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.由此产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的“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,代表算法有 GN[1],FN[2] 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 等人[3] 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,许多经典算法孕育而生,如 EAGLE[4],LFM[5],COPRA[6],UEOC[7] 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.由于语义信息内容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[8] 作为语义处理的核心模型.根据 LDA 模型的应用方式可分为以下 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 等人[9] 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 Henderson 等人在 SSN-LDA 模型的基础上融入了(infinite relational models, IRM)模型[10],提出了 LDA-G 算法[11],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.随后 Henderson 等人[12] 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 等人也在 SSN-LDA 算法的基础上提了面向有权网络的 GWN-LDA 算法[13] 及面向层次划分的 HSN-PAM[14] 算法.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[15] 作为 LDA 分析的基本模型,代表算法有 McCallum 等人[16] 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 等人[17] 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 等人[18] 在 AT 模型中加入了 user 分布取样,提出了 CUT 模型.Cha 等人[19] 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.此类算法的优点在于,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;缺点在于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 等人[20] 提出的 GT 模型,该模型是 ART 模型的扩展,用 group 取样替代了 ART 模型的 recipient 取样.随后 Pathak 等人[21] 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.近年来,话题-社区的关系成为 LDA 模型研究的重点:Mei 等人[22] 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;Sachan 等人和 Yin 等人分别从话题-社区分布和社区-话题分布角度,将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[23-24] 及 LCTA[25] 模型,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.语义社会网络的语义社区发现算法需要以复杂网络中发现直径较小的核心结构和网络簇结构作为研究基础,可对所建立的语义社会网络利用有权网络聚类算法和分类算法实现社区发现[3].对此,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.

本文抽取并量化各节点的语义信息,并以此为基础改造传统关系社区划分中的随机游走社区发现策略,在保障节点的语义信息相似度高的前提下,建立对社区个数无约束的划分方法.在实现语义社区发现的同时创新提出语义社区评价函数,从语义信息量化、语义社区发现和语义社区评价 3 方面进行论述,最后通过实验分析了本文算法的有效性和可行性.

1　节点语义信息的 LDA 模型分析

1.1　构建 LDA 模型

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,所涉及到的数学符号如下:

G 表示全局网络;

$|G|$ 表示语义社会网络中的节点个数;

d_i 表示编号为 i 的节点;

$\boldsymbol{A}$ 表示网络 G 的邻接矩阵,$A_{ij}=1$ 表示节点 d_i 与 d_j 相连;

D 表示节点度数,D_i 表示节点 d_i 的度数;

N 表示关键字表中的文本关键字个数;

N_d 表示表示节点 d 的文本关键字个数;

z 表示某一话题的出现概率,z_{in} 为节点 d_i 中第 n 号文本关键字所属话题的概率;

θ 表示话题分布的先验参数,θ_i 为节点 d_i 中话题 z_{in} 分布的先验参数;

α 表示语义社会网络中 θ 的全局先验参数;

k 表示全局话题个数;

$\boldsymbol{\beta}$ 表示 $k\times N$ 文本关键字-话题概率矩阵,$\beta_{i,j}=p(w=j|z=i)$,$\beta_{i,\cdot}=1$;

w 表示某一特定文本关键字,w_n 的编号为 n 的文本关键字;

w_n^j 表示编号为 n 的节点中编号为 j 文本关键字的出现频数.

LDA 模型是以文本关键字-话题-参数先验关系构成的 3 层贝叶斯模型,根据文献[8]三者之间的关系表达模型如图 1 所示,某一节点 d 的局部语义信息(通过文本关键字表示)模型可表示为

$$p(w|\alpha,\boldsymbol{\beta})=\int p(\theta|\alpha)\Big(\prod_{n=1}^{N}p(w_n|\theta,\boldsymbol{\beta})\Big)\mathrm{d}\theta. \quad (1)$$

全局语义信息的数学模型可表示为

$$p(w|\alpha,\boldsymbol{\beta})=\prod_{i=1}^{|G|}\int p(\theta_i|\alpha)\Big(\prod_{n=1}^{N}p(w_n|\theta,\boldsymbol{\beta})\Big)\mathrm{d}\theta_i. \quad (2)$$

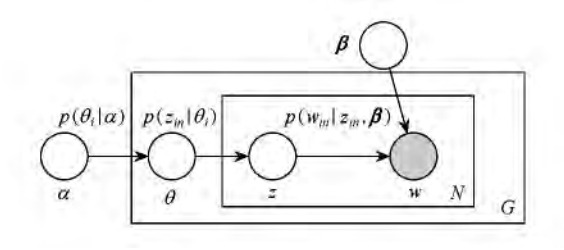

Fig. 1　The LDA plate model.

图 1　LDA“盘子”模型

根据文献[8],SSN 中节点的 LDA 的生成模型假设如下:

1) $p(\theta|\alpha)\sim Dir(\alpha)$. Dir 为 Dirichlet 分布,其表达式为

$$p(\theta|\alpha)=\frac{\Gamma\Big(\sum_{i=1}^{k}\alpha_i\Big)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\times\cdots\times\theta_k^{\alpha_k-1}. \quad (3)$$

2) $p(z|\theta)\sim Multinomial(\theta)$. $Multinomial$ 为多项式分布.

3) $p(w_n|z_n,\boldsymbol{\beta})=\beta_{z_n,w_n}\times p(w)$的条件概率表达式为

$$p(w|\theta,\boldsymbol{\beta})=\sum_{z=1}^{k}p(w|z,\boldsymbol{\beta})p(z|\theta). \quad (4)$$

根据式(3)(4),式(2)可表示为

$$p(w|\alpha,\boldsymbol{\beta})=\frac{\Gamma\Big(\sum_{i=1}^{k}\alpha_i\Big)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\times\int\Big(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\Big)\Big(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{N_d}(\theta_i\beta_{ij})^{w_n^j}\Big)\mathrm{d}\theta, \quad (5)$$

其中,w_n^j 表示在编号为 n 的节点中编号为 j 文本关键字的出现频数.

1.2 LDA 变分推理过程

根据文献[8],变分推理过程是在 LDA 模型基础上增加如下前提:

1) 加入节点内部估计参数 γ 和 φ,γ 为 θ 的节点语义信息样本估计值,φ 为文档内部话题的后验概率,$\varphi_{i,j}=p(z=j|w=i)$.

2) 假设 γ 和 φ 相互独立.

利用变量 θ 和 z 建立节点内部隐含参数的估计模型如下:

$$q(\theta,z|\gamma,\varphi)=q(\theta|\gamma)\prod_{n=1}^{N}q(z_n|\varphi_{n,\cdot}). \tag{6}$$

变分推理以极大化文本关键字-话题分布的似然函数 $p(w|\alpha,\boldsymbol{\beta})$ 为目标,通过在似然函数中加入样本估计参数 γ 和 φ,实现对全局参数 α 和 $\boldsymbol{\beta}$ 的优化. 为此,式(5)的似然函数表达式可表示如下:

$$\ln p(w|\alpha,\boldsymbol{\beta})=\ln\int\sum_{z=1}^{k}p(\theta,z,w|\alpha,\boldsymbol{\beta})\mathrm{d}\theta=\ln\int\sum_{z=1}^{k}\frac{p(\theta,z,w|\alpha,\boldsymbol{\beta})q(\theta,z)}{q(\theta,z)}\mathrm{d}\theta. \tag{7}$$

根据 Jensen 不等式,

$$\ln p(w|\alpha,\boldsymbol{\beta})\geqslant\int\sum_{z=1}^{k}q(\theta,z)\ln p(\theta,z,w|\alpha,\boldsymbol{\beta})\mathrm{d}\theta-\int\sum_{z=1}^{k}q(\theta,z)\ln q(\theta,z)\mathrm{d}\theta=EP_q[\ln p(\theta,z,w|\alpha,\boldsymbol{\beta})]-EP_q[\ln q(\theta,z)]=L(\gamma,\varphi;\alpha,\boldsymbol{\beta}), \tag{8}$$

其中 EP_q 表示利用估计参数 γ 和 φ 计算的期望. 由于 Dirichlet 分布属于一种指数分布族,根据文献[8]可知:

$$\begin{cases}EP[\ln(\theta_i)|\alpha]=\Psi(\alpha_i)-\Psi\left(\sum_{j=1}^{k}\alpha_j\right);\\ EP[\ln(\theta_i)|\gamma]=\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right).\end{cases} \tag{9}$$

变分推理的优化过程即寻找 $L(\gamma,\varphi|\alpha,\boldsymbol{\beta})$ 的极值过程. 根据式(8),

$$L(\gamma,\varphi|\alpha,\boldsymbol{\beta})=EP_q[\ln p(\theta|\alpha)]+EP_q[\ln p(z|\theta)]+EP_q[\ln p(w|z,\boldsymbol{\beta})]-EP_q[\ln q(\theta)]-EP_q[\ln q(z)]. \tag{10}$$

利用式(9)可得到如下表达式:

$$\begin{aligned}L(\gamma,\varphi|\alpha,\boldsymbol{\beta})=&\left(\ln\Gamma\left(\sum_{j=1}^{k}\alpha_j\right)-\sum_{i=1}^{k}\ln\Gamma(\alpha_i)+\sum_{i=1}^{k}(\alpha_{i-1}-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\right)+\\&\left(\sum_{n=1}^{N}\sum_{i=1}^{k}\varphi_{ni}\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\right)+\\&\left(\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\varphi_{ni}w_n^j\ln\beta_{ij}\right)-\left(\ln\Gamma\left(\sum_{j=1}^{k}\gamma_j\right)-\sum_{i=1}^{k}\ln\Gamma(\gamma_i)+\sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\right)-\left(\sum_{n=1}^{N}\sum_{i=1}^{k}\varphi_{ni}\ln\varphi_{ni}\right).\end{aligned} \tag{11}$$

式(11)包含了(α, $\boldsymbol{\beta}$, γ, φ) 4 个参数,其中 $\sum_{i=1}^{k}\varphi_{ni}=1$,$\sum_{j=1}^{N}\beta_{ij}=1$ 利用拉格朗日乘子法对(α, $\boldsymbol{\beta}$,γ,φ)进行优化求值得到如下结果:

$$\begin{aligned}L_{[\alpha]}=&\sum_{d=1}^{|G|}\left(\ln\Gamma\left(\sum_{j=1}^{k}\alpha_j\right)-\sum_{i=1}^{k}\ln\Gamma(\alpha_i)+\sum_{i=1}^{k}\left((\alpha_i-1)\left(\Psi(\gamma_{di})-\Psi\left(\sum_{j=1}^{k}\gamma_{d_j}\right)\right)\right)\right).\end{aligned} \tag{12}$$

$$\begin{aligned}L_{[\gamma]}=&\sum_{i=1}^{k}(\alpha_{i-1}-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)+\sum_{n=1}^{N}\sum_{i=1}^{k}\varphi_{ni}\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)-\ln\Gamma\left(\sum_{j=1}^{k}\gamma_j\right)+\sum_{i=1}^{k}\ln\Gamma(\gamma_i)-\sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right).\end{aligned} \tag{13}$$

$$L_{[\boldsymbol{\beta}]}=\sum_{d=1}^{|G|}\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{N_d}\varphi_{ni}w_{dn}^j\ln\beta_{ij}+\sum_{i=1}^{k}\lambda_i\left(\sum_{j=1}^{N}\beta_{ij}-1\right). \tag{14}$$

其中,w_n^j 是节点 d_n 中第 j 号文本关键字的出现次数,$\sum_{j=1}^{N_d}\varphi_{ni}w_n^j\ln\beta_{ij}=\varphi_{ni}\sum_{j=1}^{N_d}w_n^j\ln\beta_{ij}=\varphi_{ni}\ln\beta_{id_n}$,$L_{[\varphi]}$ 表示如下:

$$\begin{aligned}L_{[\varphi_{ni}]}=&\varphi_{ni}\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)+\\&\varphi_{ni}\ln\beta_{id_n}-\varphi_{ni}\ln\varphi_{ni}+\lambda_n\left(\sum_{j=1}^{k}\varphi_{ni}-1\right).\end{aligned} \tag{15}$$

式(12)~(15)分别对(α,$\boldsymbol{\beta}$,γ,φ)求导数可得到(α,$\boldsymbol{\beta}$,γ,φ)的极值关系式如下:

$$\Psi\left(\sum_{j=1}^{k}\alpha_j\right)-\Psi(\alpha_i)=\sum_{d=1}^{|G|}\Psi\left(\sum_{j=1}^{k}\gamma_j\right)-\Psi(\gamma_i). \tag{16}$$

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\varphi_{ni}. \tag{17}$$

$$\beta_{ij}\propto\sum_{d=1}^{|G|}\sum_{n=1}^{N_d}\varphi_{d_ni}w_{dn}^j. \tag{18}$$

$$\varphi_{ni}\propto\beta_{id_n}\exp\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right). \tag{19}$$

根据式(16)~(19)变分推理的参数训练过程分

为各节点局部语义信息参数循环训练过程(训练 γ,φ)和总体语义信息参数训练过程(训练 α,β),局部语义信息参数循环训练过程是总体语义信息参数过程的子过程. 图 2 为训练过程的盘子模型图:

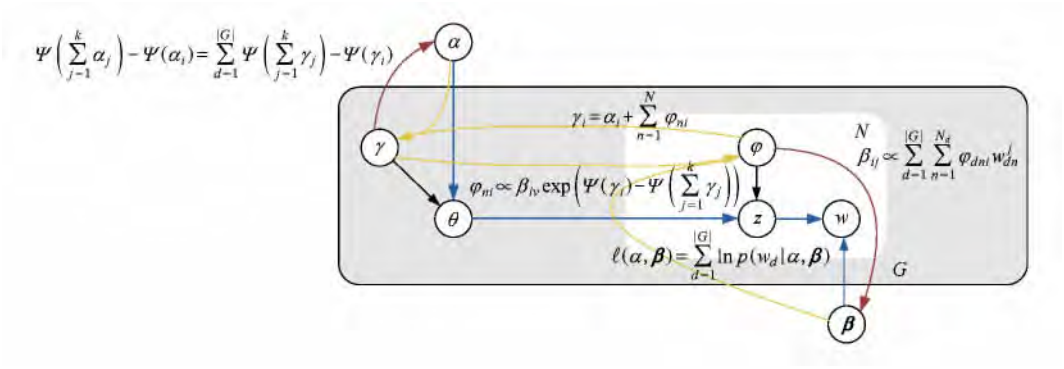

Fig. 2　The proceeding of LDA variational inference.

图 2　LDA 变分推理过程

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,$\boldsymbol{\varphi}_{i,\cdot}$ 表示为 i 号文本关键字在 k 维语义空间中的坐标,则某一节点 d_i 在语义空间中的坐标(语义坐标)$\boldsymbol{m}_i$ 可通过 d_i 的 N_i 个关键字的加权均值形式表达,其表达式为

$$\boldsymbol{m}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\varphi}_{j,\cdot}, \tag{20}$$

其中,N_i 为节点 d_i 的关键字个数.

在语义社会网络节点的关系度量方面,由于每个节点均有语义空间中的量化表示,因此可将向量内积 $E_{ij}=\boldsymbol{m}_i\cdot\boldsymbol{m}_j$ 作为节点 d_i 和节点 d_j 的语义相似性度量,其中 $E_{ij}>0$,且 E_{ij} 越大节点 d_i 和节点 d_j 的相似性越高. 由此可根据节点的相似性构造语义社会网络的加权邻接矩阵 $\boldsymbol{E}$:

$$E_{ij}=\begin{cases}\boldsymbol{m}_i\cdot\boldsymbol{m}_j, & a_{i,j}=1;\\ 0, & a_{i,j}=0.\end{cases} \tag{21}$$

在语义社会网络节点的影响力度量方面,节点 d_i 在语义社会网络中的语义影响力 M_i,可通过以下 2 方面进行建模:

1) 语义信息比重. 可通过节点 d_i 信息含量进行度量,节点 d_i 的信息含量越大则其语义信息比重越大. 由于节点 d_i 的语义信息可量化为 k 维向量 $\boldsymbol{m}_i$,其信息含量大小可用向量 $\boldsymbol{m}_i$ 归一化后的信息熵值(entropy)表达.

2) 节点关系比重. 可利用节点 d_i 的度分布比率表示. 由于社会网络具有"小世界"特性,节点的度分布服从幂律分布,其密度函数 $P(i)\propto\exp(-D_i/\sigma)$,

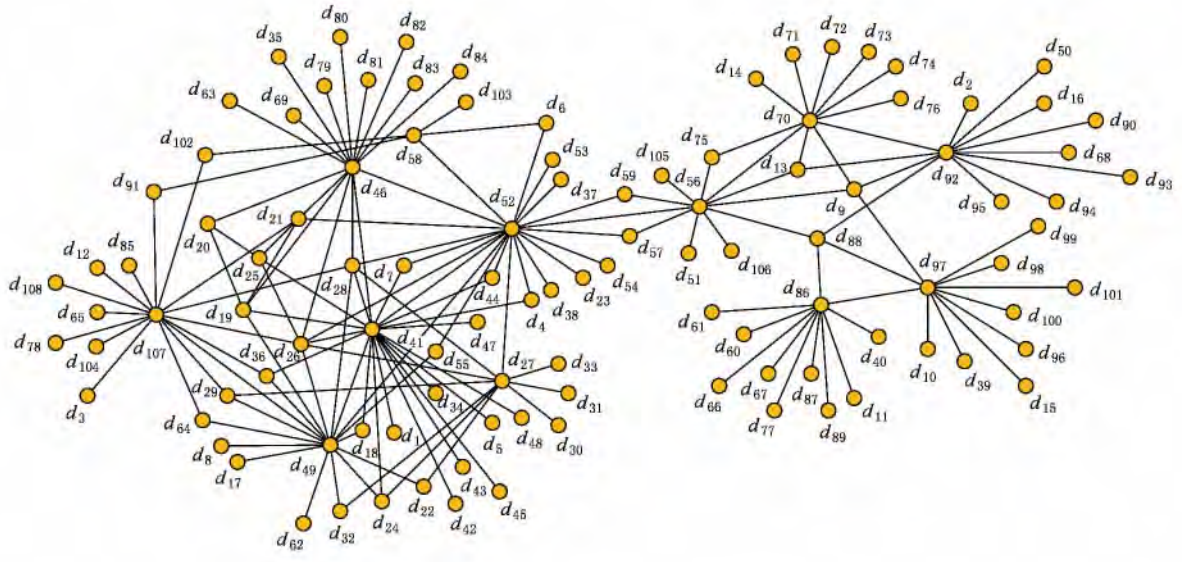

Fig. 3　The network topology of QLSP.

图 3　QLSP 网络拓扑

σ 为幂律分布的参数. 在社会网络中度数越大的节点出现概率越低,对社会网络的关系影响越大,因此可采用出现概率倒数 $\exp(D_i/\sigma)$ 的形式作为节点关系比重的表达式,为平衡语义信息比重与节点关系比重的比例,σ 度数取最大值 $\max(D)$. 通过以上分析,节点 d_i 在语义社会网络中的影响力 M_i 可表示为

$$M_i = entropy(normalize(\boldsymbol{m}_i))\exp[D_i/\sigma] = \left(-\sum_{j=1}^{k}\ln\left(m_{i,j}\Big/\sum_{j=1}^{k}m_{i,j}\right)\right)\exp[D_i/\max(D)]. \tag{22}$$

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、155 条引用关系),其网络模型如图 3 所示. 本文算法分别在每篇论文的摘要中抽取 10 个关键字作为论文节点的语义信息,并设置话题个数 k 为 3,利用式(21)和式(22)计算各节点的语义信息比重、节点关系比重及语义影响力 M_i,其结果如图 4 所示. 图 5 为该网络的加权邻接矩阵 $\boldsymbol{E}$.

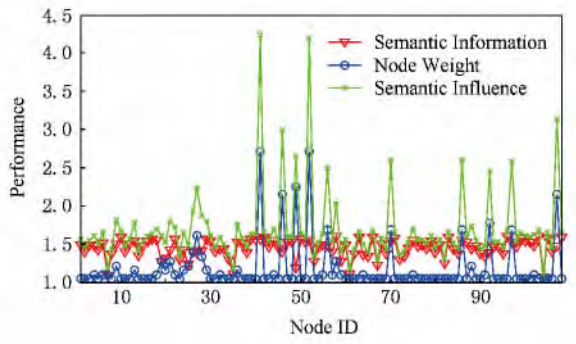

Fig. 4 The comparison chart of semantic influence on QLSP.

图 4 QLSP 网络的语义影响力对比图

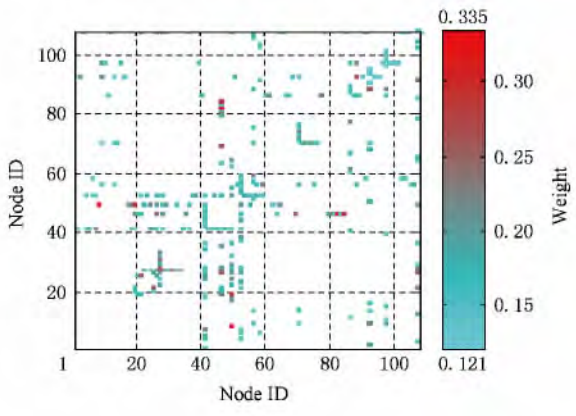

Fig. 5 The weighted adjacent matrix of QLSP.

图 5 QLSP 网络的加权邻接矩阵

3 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 的社区内部节点. 由第 2 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 $\boldsymbol{E}$ 和语义影响力 M 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以下 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 $\boldsymbol{E}$ 表达了语义链接关系的强弱,E_{ij} 越大节点 d_j 相对于节点 d_i 的其他相邻节点的语义信息相似度越高,因此节点 d_j 与节点 d_i 同属一个社区的概率较大. 为使随机游走策略的社区划分结果体现节点语义信息相似度,可将行向量归一化后的加权邻接矩阵 $\boldsymbol{E}$ 作为随机游走策略的一步转移概率矩阵 $\boldsymbol{H}$,其表达式为

$$H_{ij} = E_{ij}\Big/\sum_{r=1}^{|G|}E_{ir}. \tag{23}$$

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 d_i 的概率,则 $\varepsilon_s^l(i)$ 可通过迭代式(24)进行表达:

$$\varepsilon_s^l(i) = \sum_{r=1}^{|G|}\varepsilon_s^l(r)H_{ri}. \tag{24}$$

2) 设定局部扩展节点. 由于语义影响力 M 表达了节点的语义信息比重和节点关系比重,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节点,因此可选择语义影响力 M 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 $\varepsilon_s^l(i) > \kappa$ 的节点 d_i 与 s 划分为同一社区即为 s 初始社区.

根据文献[7]可知,利用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)的时间不同,所发现的社区结果波动较大,且会导致 s 初始社区出现不连通的情况. 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. 本文在第 5 节分析 l 和 η 的取值.

由此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,根据文献[7]的随机游走算法框架,设计语义重叠社区发现策略,整体流程如图 6 所示. 具体描述如下:

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影响力最大的节点 s 作为局部扩展节点;

3) 利用马尔可夫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;

5) 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.

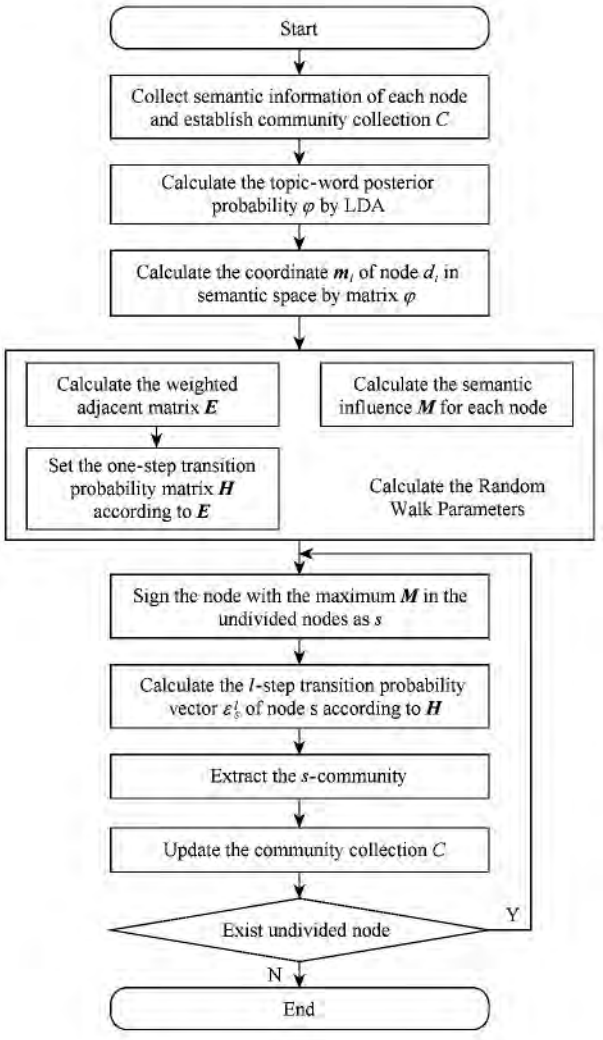

Fig. 6　The algorithm flow chart.

图 6　算法流程图

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关系结构为输入,文献[4]所建立的重叠社区模块度 EQ 模型为

$$EQ=\frac{1}{R}\sum_{t=1}^{|C|}\sum_{i\in C_t,j\in C_t}\frac{1}{O_iO_j}\left(A_{i,j}-\frac{D_iD_j}{R}\right),\quad(25)$$

其中,D_i 为节点 d_i 的度数,R 为网络节点的总度数,$\boldsymbol{A}$ 为网络邻接矩阵,O_i 为节点 d_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 $U(\boldsymbol{m}_i,\boldsymbol{m}_j)$,建立评价语义重叠社区的模块度模型 SQ,其表达式为

$$SQ=\frac{1}{R}\sum_{t=1}^{|C|}\sum_{i\in C_t,j\in C_t}\frac{U(\boldsymbol{m}_i,\boldsymbol{m}_j)}{O_iO_j}\left(A_{i,j}-\frac{D_iD_j}{R}\right),\quad(26)$$

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度量函数 $U(\boldsymbol{m}_i,\boldsymbol{m}_j)$,表达式为

$$U(\boldsymbol{m}_i,\boldsymbol{m}_j)=\frac{\boldsymbol{m}_i\cdot\boldsymbol{m}_j}{|\boldsymbol{m}_i||\boldsymbol{m}_j|}=\sum_{g=1}^{k}m_{i,g}m_{j,g}\Big/\sqrt{\sum_{g=1}^{k}m_{i,g}\sum_{g=1}^{k}m_{j,g}},\quad(27)$$

本文将在第 5 节对 SQ 进行实验分析.

5　实验分析

5.1　参数 l 和 η 的取值分析

本节实验利用图 3 所示的 QLSP 数据集分析随机游走步长 l 和重复度 η 对结果的影响. 其中 l 控制抵达概率分布 ε_s^l,从而直接影响社区划分结果. l 取值过小会导致游走过程不充分,过大会导致社区划分

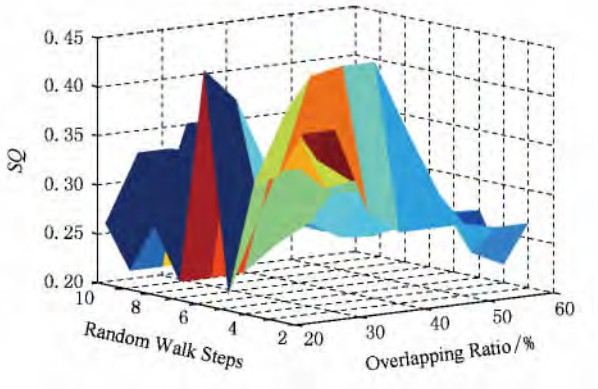

Fig. 7　The SQ of QLSP network.

图 7　QLSP 网络的 SQ 值

结果因陷入局部混合状态使社区结构紊乱;社区重复度 η 控制初始社区整合,若 η 取值过小则导致重叠社区过多,取值过大则导致离散边界节点和重复节点过多. 图 7 为 l 和 η 分别取(2～10)和(20%～90%)时 QLSP 数据集的 SQ 函数值,从图 7 可以直观分析出当 $l=3$, $\eta=40\%$ 时,社区划分结果最佳, SQ 为 0.443, EQ 为 0.432,结果如图 8 所示. 图 9 为 l 和 η 在不同取值下的结果,其中黑色节点表示重叠节点,从中可直观验证 l 和 η 对社区划分结果的影响.

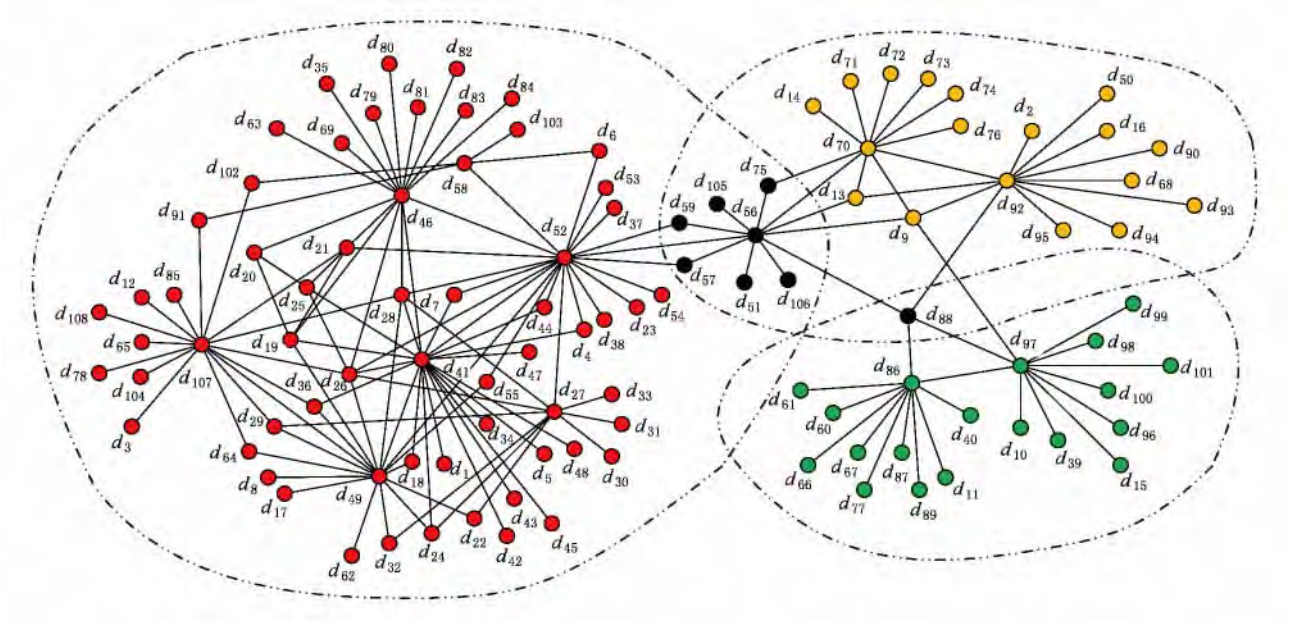

Fig. 8 The semantic community of QLSP network.

图 8 QLSP 网络语义社区划分结果

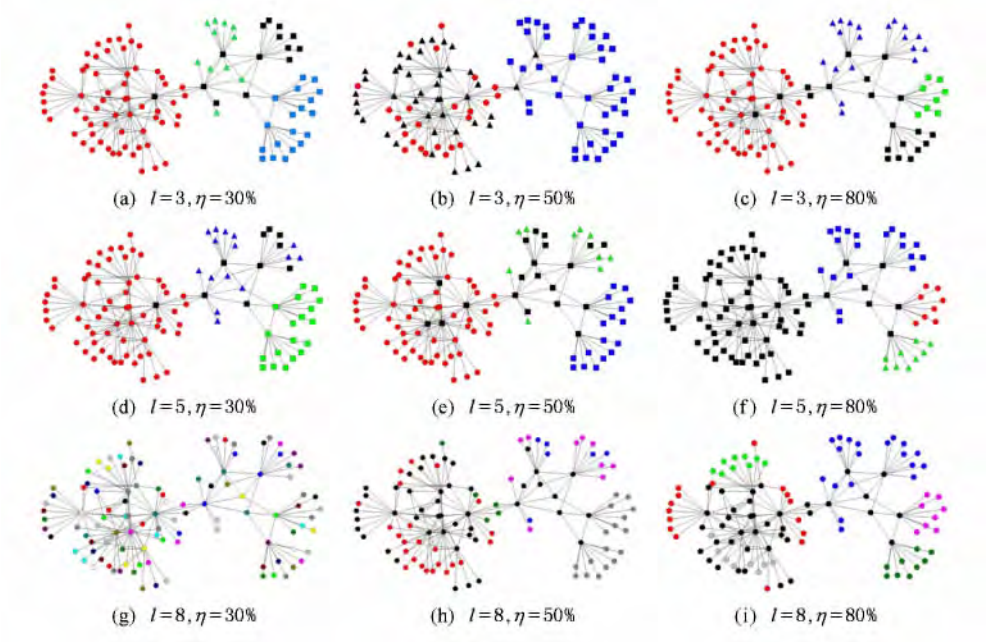

Fig. 9 Community structures with different l and η.

图 9 l 和 η 在不同取值下的网络划分结果

5.2 SQ 与 EQ 的比较分析

本节实验计算了 QLSP 数据集不同划分结果的 EQ 和 SQ,并按步长 l(2～10)将结果分为 9 组,每组中 η 的取值为 20%～90%, EQ 和 SQ 的对比如图 10 所示. 从式(25)和式(26)的对比可知, SQ 加入了语义信息相似性度量函数 $U(\boldsymbol{m}_i,\boldsymbol{m}_j)<1$,使得 SQ 的总体趋势小于 EQ, QLSP 中 SQ 与 EQ 的最优社区结构分别为 $l=3$, $\eta=40\%$ 和 $l=4$, $\eta=30\%$,

验证了仅以节点关系为输入的 EQ,在评价面向节点具有语义信息属性的社区划分结果时会产生偏差.

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应对语义社会网络时划分结果存在的偏差,因此本节实验仅以 QLSP 数据集为例进行说明. 社区发现中经典算法包括 GN[1], FN[2], LFM[5], COPRA[6], UEOC[7], EAGLE[4], CPM[3],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, LFM, COPRA, UEOC 算法进行求解,图 11 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 1 所示. 从表 1 的结果可知,经典算法的 EQ 值高于本文算法(0.432),但 SQ 值均低于本文算法(0.443),由此验证了面向链接关系的算法不适用于语义社区发现,同时验证了本文算法相较于传统算法的优越性.

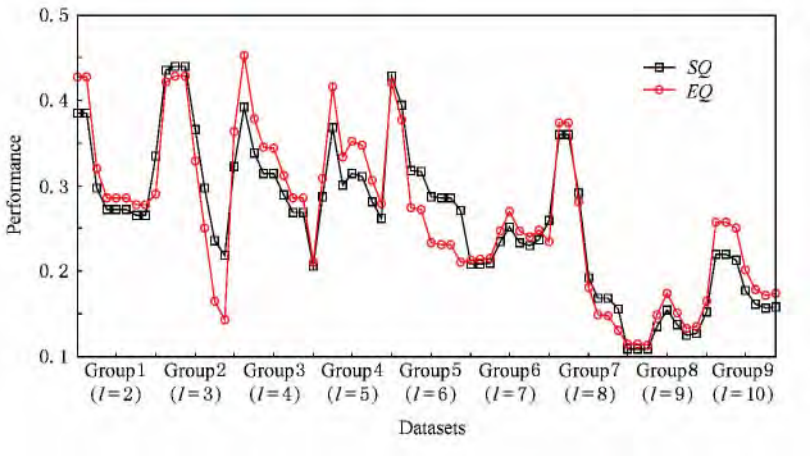

Fig. 10　The comparison chart of EQ and SQ.

图 10　EQ 和 SQ 对比结果

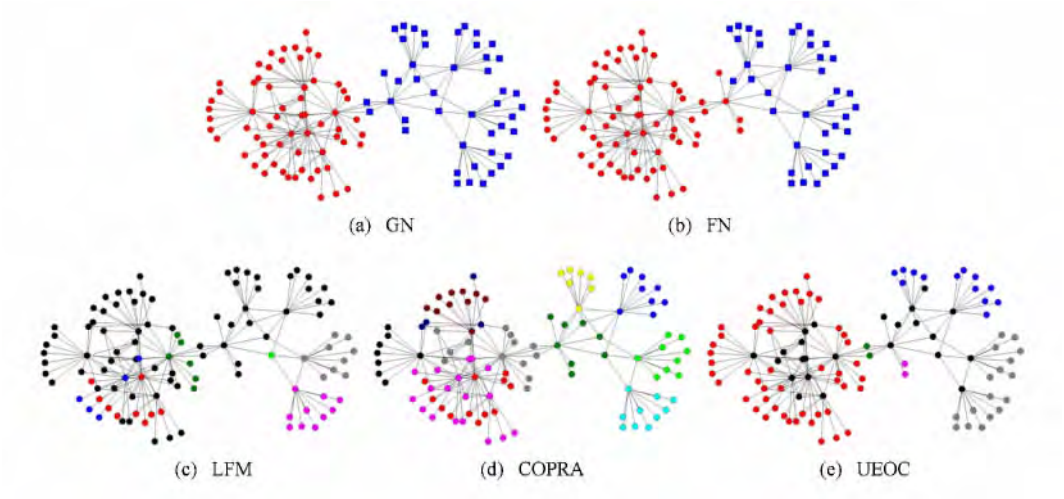

Fig. 11　The community results from classical algorithms.

图 11　各算法的社区划分结果

Table 1　The Value of SQ and EQ by Classical Algorithms

表 1　经典算法的 SQ 和 EQ 值

Algorithms	SQ	EQ
GN	0.358 4	0.461 7
FN	0.315 7	0.406 1
LFM	0.232 9	0.325 4
COPRA	0.420 3	0.541 0
UEOC	0.407 1	0.441 0

5.4 真实数据集比较分析

本实验以清华大学 ArnetMiner 系统的 QLSP 完整数据集(共 805 个节点)、AFD(aminer foaf dataset)数据集(截取 2 000 个节点)、CND(citation network dataset)数据集(共 2 555 个节点)、DBLP(database systems and logic programming) April 12,2006 数据集(1 200 000 个节点)中分别截 1 500 个节点作为 DBLP(A)数据集和 2 000 个节点作为 DBLP(B)数据集作为实验数据,分析本文算法与经典算法的比较结果. 表 2 为各算法对上述数据集的执行结果(本文 SR 算法的运行参数为 $l=3$, $\eta=40\%$),包括 EQ, SQ 及社区个数 CS,图 12 和图 13 分别为各算法的 EQ 和 SQ 直方图,其中图 12 的结果表示本文 SR 算法结果在 EQ 标准下的结果较差,图 13 的结果充分验证了本文 SR 算法的语义社区划分结果更精确,从图 12 和图 13 的比对可知,相较于传统算法本文 SR 算法更适合处理语义社会网络社区发现问题.

Table 2 Results from Classical Algorithms with Different Datasets

表 2 各数据集的执行结果

Algorithms	Measurement	QLSP	AFD	CND	DBLP(A)	DBLP(B)
GN	*EQ*	0.31	0.13	0.19	0.28	0.31
	SQ	0.23	0.15	0.18	0.21	0.28
	CS	10	25	39	17	16
FN	*EQ*	0.42	0.15	0.22	0.31	0.26
	SQ	0.32	0.13	0.17	0.29	0.25
	CS	10	27	37	19	16
LFM	*EQ*	0.36	0.14	0.24	0.4	0.36
	SQ	0.31	0.13	0.21	0.33	0.31
	CS	12	24	33	22	12
COPRA	*EQ*	0.41	0.31	0.11	0.38	0.41
	SQ	0.28	0.21	0.12	0.29	0.32
	CS	13	21	35	21	13
UEOC	*EQ*	0.38	0.23	0.26	0.36	0.31
	SQ	0.31	0.22	0.22	0.29	0.2
	CS	12	24	30	22	14
SR	*EQ*	0.31	0.25	0.28	0.37	0.3
	SQ	0.34	0.24	0.25	0.35	0.33
	CS	11	20	33	20	13

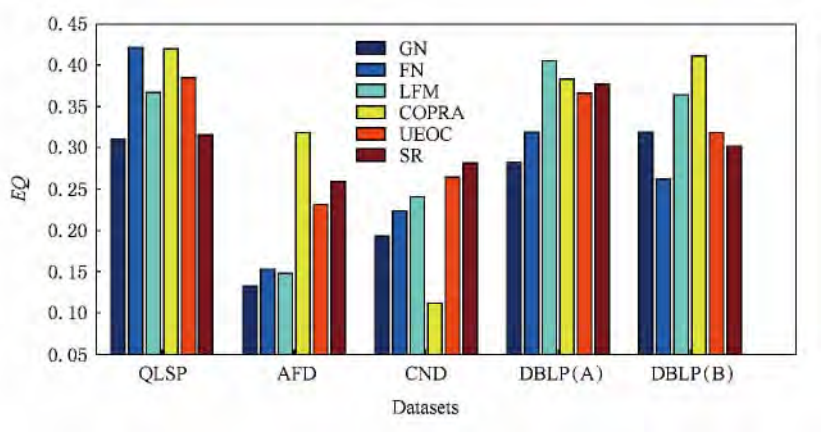

Fig. 12 The histogram of *EQ* for different classical algorithms.

图 12 各算法的 *EQ* 直方图

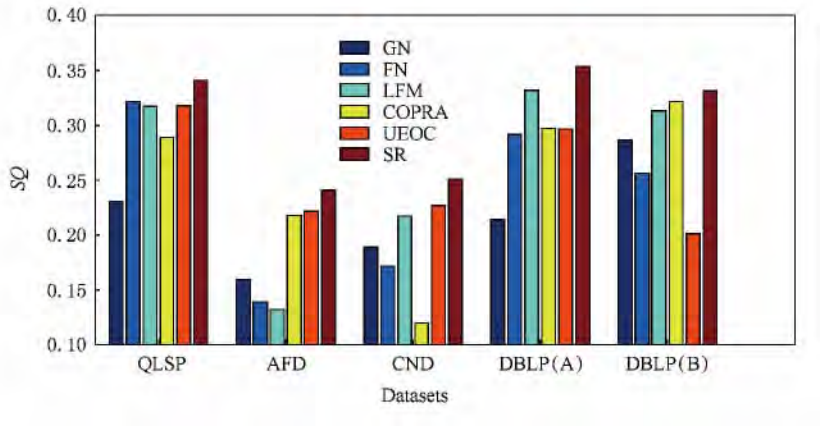

Fig. 13　The histogram of *SQ* for different classical algorithms.

图 13　各算法的 *SQ* 直方图

5.5　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 0.5×10^6 条数据,423×10^6 B. 表 3 为经 LDA 分析后从 Enron 数据集中抽取的 4 组话题. 表 4 与表 5 分别为 Enron 数据集分别在 TURCM,CART,CUT,LCTA 算法下的 *EQ* 值及 *SQ* 值,其中社区个数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 0.325 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.

Table 3　Topics Extracted from Enron

表 3　Enron 数据集的话题分组

Topic	California Power	Gas Transportation	Trading	Deals
Word	Power	Gas	Price	Meeting
	Transmission	Energy	Market	Contract
	Energy	Enron	Dollar	Report
	Calpx	Transco	Nymex	Enron
	California	Chris	Trade	Deal

Table 4　The *EQ* of Various Semantic Community Detection Algorithms

表 4　各类语义社区发现算法的 *EQ* 值

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

Table 5　The *SQ* of Various Semantic Community Detection Algorithms

表 5　各类语义社区发现算法的 *SQ* 值

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

5.6　实验总结

本文实验部分分别从 SR 参数取值、评价函数有效性、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) SR 算法的最优参数取值为 $l=3$,$\eta=40\%$;

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,SR 相对于经典重叠社区发现算法更有效;

4) SR 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,SR 算法无需预设社区个数且结果较好.

6　结　　论

本文针对语义社会网络社区划分的问题提出了 SR 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,结合了随机游走算法框架实现语义社区划分,其创新思想在于:1)利用 LDA 算法构建语义空间,并将节点的语义信息映射为语义空间内的坐标,使节点的语义信息可度量化;2)以节点语义

空间坐标的信息熵值作为节点语义信息比重,以节点的度分布比率作为节点关系比重,并根据节点的语义信息比重和关系比重建立节点语义影响力模型及加权邻接矩阵;3)根据加权邻接矩阵和节点语义影响力建立语义社区重叠社区发现的随机游走策略;4)建立了评价语义社区划分结果的 SQ 模型.

本文实验分析验证在面向具有语义关系的社区划分问题时,SR 算法相对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、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] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of National Academy of Science, 2002, 9(12): 7921-7826

[2] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 1-8

[3] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[4] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A, 2009, 388(8): 1706-1712

[5] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 1-8

[6] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2010, 12(10): 1-9

[7] Jin D, Yang B, Baquero C, et al. A Markov random walk under constraint for discovering overlapping communities in complex networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2011, 8(5): 1-11

[8] Blei D M, Ng A Y, Jordan MI. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 21(3): 993-1022

[9] Zhang H, Qiu B, Giles C L, et al. An LDA-based community structure discovery approach for large-scale social networks [C] // Proc of the 7th Conf on Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2007: 200-207

[10] Kemp C, Tenenbaum J B, Griffiths T L, et al. Learning systems of concepts with an infinite relational model [C] // Proc of the 21st Association for the Advancement of Artificial Intelligence. Palo Alto: AAAI, 2006: 5-13

[11] Henderson K, Eliassi R T. Applying latent Dirichlet allocation to group discovery in large graphs [C] // Proc of the 2009 ACM Symp on Applied Computing. New York: ACM, 2009: 1456-1461

[12] Henderson K, Eliassi-Rad T, Papadimitriou S, et al. HCDF: A Hybrid community discovery framework [C] // Proc of the 10th SIAM Int Conf on Data Mining. Philadelphia: SDM, 2010: 754-765

[13] Zhang H, Giles C L, Foley H C, et al. Probabilistic community discovery using hierarchical latent gaussian mixture model [C] // Proc of the 22nd Association for the Advancement of Artificial Intelligence. Palo Alto: AAAI, 2007: 663-668

[14] Zhang H, Li W, Wang X, et al. HSN-PAM: Finding hierarchical probabilistic groups from large-scale networks [C] // Proc of the 7th Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2007: 27-32

[15] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery [C] // Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 306-315

[16] McCallum A, Corrada EA, Wang X. Topic and role discovery in social networks [J]. Computer Science Department Faculty Publication Series, 2005, 16(3): 1-7

[17] McCallum A, Wang X, Corrada-Emmanuel. A. Topic and role discovery in social networks with experiments on enron and academic email [J]. Journal of Artificial Intelligence Research, 2007, 30(8): 249-272

[18] Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discovering e-communities [C] // Proc of the 15th Int Conf on World Wide Web. New York: ACM, 2006: 173-182

[19] Cha Y, Cho J. Social-network analysis using topic models [C] // Proc of the 35th ACM SIGIR Int Conf on Research and Development in Information Retrieval. New York: ACM, 2012: 565-574

[20] Wang X, Mohanty N, McCallum A. Group and topic discovery from relations and text [C] // Proc of the 3rd Int Conf of Workshop on Link Discovery. New York: ACM, 2005: 28-35

[21] Pathak N, DeLong C, Banerjee A, et al. Social topic models for community extraction [C] // /Proc of the 2nd SNA-KDD Conf on Workshop. New York: ACM, 2008: 1-8

[22] Mei Q, Cai D, Zhang D, et al. Topic modeling with network regularization [C] // Proc of the 17th Int Conf on World Wide Web. New York: ACM, 2008: 101-110

[23] Sachan M, Contractor D, Faruquie T, et al. Probabilistic model for discovering topic based communities in social networks [C] // Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 2349-2352

[24] Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks [C] // Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 331-340

[25] Yin Z, Cao L, Gu Q, et al. Latent community topic analysis: Integration of community discovery with topic modeling [J]. ACM Trans on Intelligent Systems and Technology, 2012, 3(4): 1-20

Xin Yu, born in 1987. PhD candidate at Harbin Engineering University. Student member of China Computer Federation. His main research interest covers database and knowledge engineering, enterprise intelligence computing (xinyu@hrbeu.edu.cn).

Yang Jing, born in 1962. Professor and PhD supervisor at Harbin Engineering University. Senior member of China Computer Federation. Her research interest covers database and knowledge engineering, enterprise intelligence computing (yangjing@hrbeu.edu.cn).

Xie Zhiqiang, born in 1962. Postdoctoral at Harbin Engineering University. Professor at Harbin University of Science and Technology. Senior member of China Computer Federation. His research interest covers enterprise intelligence computing, database and knowledge engineering (xiezhiqiang@hrbust.edu.cn).

勘误启事

本刊 2014 年第 11 期发表的“DTN 中基于生命游戏的拥塞控制策略”(第 2393-2407 页)一文中,因作者疏忽,将该文 2.3.3 节图 9 和 2.4 节处的文献标引遗漏,此两处参考了文献“王恩,杨永健,杜占玮. 基于马尔可夫相遇时间间隔预测的拥塞控制策略[J]. 吉林大学学报(工学版), 2014,44(1):149-157”,特此补充;另 3.1 节“投递成功率=成功投递到目的节点的报文数量/网络中产生的报文总数”应改为“投递成功率=成功投递的报文种类数/网络中产生的报文种类数”. 谨向广大读者致歉.

《计算机研究与发展》编辑部