计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140308
52(7): 1510-1521, 2015

基于局部语义聚类的语义重叠社区发现算法

辛 宇[1] 杨 静[1] 汤楚蘅[2] 葛斯乔[2]
[1](哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)
[2](哈尔滨工业大学电气工程及自动化学院 哈尔滨 150001)
(xinyu@hrbeu.edu.cn)

An Overlapping Semantic Community Detection Algorithm Based on Local Semantic Cluster

Xin Yu[1], Yang Jing[1], Tang Chuheng[2], and Ge Siqiao[2]
[1](*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)
[2](*School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin* 150001)

Abstract Since the semantic social network (SSN) is a new kind of complex networks, the traditional community detection algorithms depending on the adjacency in social network are not efficient in the SSN. To solve this problem, an overlapping community structure detecting method on semantic social networks is proposed based on the local semantic cluster (LSC). Firstly, the algorithm utilizes the Gibbs sampling method to establish the quantization mapping by which the semantic information in nodes is changed into the semantic space, with the latent Dirichlet allocation (LDA) as the semantic model; Secondly, the algorithm establishes the similarity matrix of SSN, with the relative entropy of semantic coordinate as the measurement of similarity between nodes; Thirdly, according to the character of local small-world in social network, the algorithm proposes the S-fitness model which is the local community structure of SSN, and establishes the LSC method by the S-fitness model; Finally, the algorithm proposes the semantic model by which the community structure of SSN is measured, and the efficiency and feasibility of the algorithm and the semantic modularity are verified by experimental analysis.

Key words semantic social network (SSN); overlapping community structure detection; latent Dirichlet allocation (LDA); relative entropy; Gibbs sampling; local semantic cluster (LSC)

摘 要 语义社会网络是一种包含信息节点及社会关系构成的新型复杂网络,因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.针对这一问题,提出基于局部语义聚类的语义社会网络重叠社区发现算法,该算法:1)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;2)以节点间语义坐标的相对熵作为节点语义相似度的度量,建立节点相似度矩阵;3)根据社会网络的局部小世界特性,提出语义社会网络的局部社区结构 S-fitness 模型,并根据 S-fitness 模型建立了局部语义聚类算法(local semantic clusterm, LSC);4)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;重叠社区发现;LDA 模型;相对熵;Gibbs 取样;局部语义聚类

中图法分类号 TP18; TP393

收稿日期:2014-04-09;**修回日期**:2014-07-21

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)

通信作者:杨 静(yangjing@hrbeu.edu.cn)

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.从研究内容方面,社区划分算法可分为 3 个阶段:硬社区划分、重叠社区划分及语义社区划分.

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 等人[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,许多经典算法孕育而生,如 EAGLE[5]、LFM[6]、COPRA[7]、UEOC[8]、蚁群算法[9]、拓扑势算法[10]等.

语义社会网络是每个节点均具有文本信息内容的社会网络.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.由于语义信息需要以文本分析为基础,因此目前的语义社区划分算法大多以 LDA 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:

1) 关系语义信息的 LDA(latent Dirichlet allocation)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[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.Henderson 等人[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.随后 Henderson 等人[14]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[15-16]也在 SSN-LDA 算法的基础上提了面向有权网络的 GWN-LDA 算法及面向层次化分的 HSN-PAM 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[17]作为 LDA 分析的基本模型,代表算法有 McCallum 等人[18]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[19]在 ART 模型的基础上加入了角色分析过程,提出了 RART 模型,扩展了 ART 模型的在社会计算领域的应用.Zhou 等人[20]在 AT 模型中加入了 user 分布取样,提出了 CUT 模型.Cha 等人[21]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,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[22]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 等人[23]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 等人[24]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 等人[25-26]和 Yin 等人[27]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,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 2 方面条件:1)语义社区内部链接关系紧密;2)语义社区内部节点的语义信息相似度高.为避免社区-话题 LDA 分析中预设社区个数的问题,本文所设计的面向语义社会网络重叠社区发现算法,创新建立节点语义信息到语义空间的量化映射,通过构造语义相似度的度量,提出语义社会网络的局部社区结构 S-fitness 模型,并根据 S-fitness 模型建立了局部语义聚类算法(local semantic cluster, LSC)及评价语义社区划分结果的 SQ,最后通过实验,分析本文算法的有效参数选取及 SQ 评价性能.

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_i 为网络中的节点 i;

$|G|$ 表示语义社会网络中的节点个数;

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

w 表示关键字的向量,w_i 为向量 w 中第 i 个关键字所对应的编号;

d 表示与关键字的向量 w 对应的节点编号向量,d_i 为 w_i 所隶属的节点编号;

z 表示与关键字的向量 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;

$\theta^{(d_i)}$ 表示节点 d_i 的话题分布概率;

$\lambda^{(j)}$ 表示话题 j 中关键字的分布,$\lambda_{w_i}^{(j)}$ 表示关键字 w_i 隶属某一话题 j 的概率,其中 $\lambda_{w_i}^{(j)}=P(w_i|z_i=j)$;

α 表示各节点的话题分布先验参数;

β 表示话题内部关键字分布的先验参数.

LDA 语义数据分别利用 w,d,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 w,d,z 数据存储结构,其中阴影部分表示向量内的相同元素,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 三者之间存在 3 层贝叶斯关系,根据文献[11]可知,w,d,z 的关系如下:

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

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

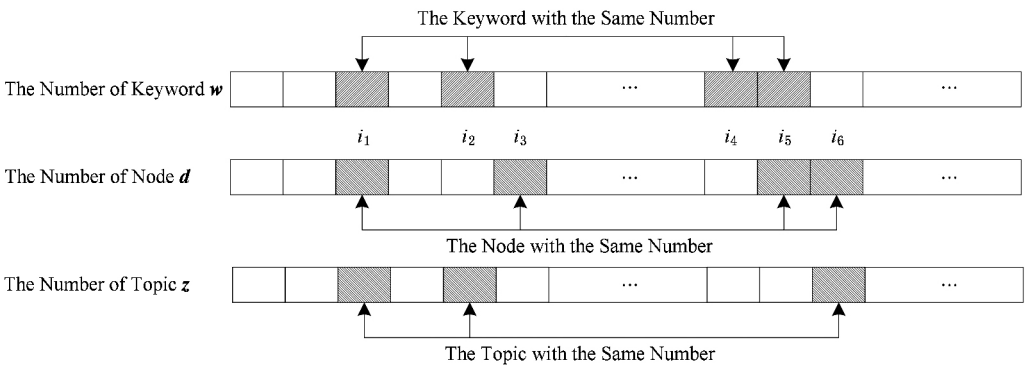


Fig. 1 The data storage structure of w,d,z.

图 1 w,d,z 数据存储结构

③ $\lambda\sim$Dirichlet(β),关键字分布 λ 服从参数为 β 的狄利克雷分布;

④ $w_i|z_i,\lambda^{(z_i)}\sim$Multinomial($\lambda^{(z_i)}$),话题 z_i 在特定话题分布 λ 下,出现关键字 w_i 的概率服从多项式分布.

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.

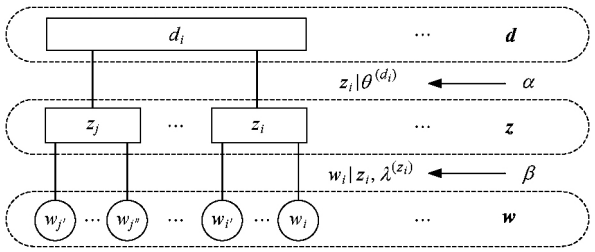


Fig. 2　The bayesian diagram of w,d,z.

图 2　w,d,z 的贝叶斯关系图

1.2　Gibbs 迭代过程

w,z 的贝叶斯关系表达式为

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

其中,$P(w_i)=\sum_{j=1}^{|z|}P(w_i\mid z_i=j)P(z_i=j)$.

Gibbs 取样算法的核心内容在于通过建立已知样本分布为条件,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 $\boldsymbol{z}_{-i}$ 和 $\boldsymbol{w}_{-i}$(表示除去元素 i 的向量 $\boldsymbol{z}$ 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

$$P(z_i=j\mid \boldsymbol{z}_{-i},w_i)P(w_i\mid \boldsymbol{w}_{-i})=P(w_i\mid z_i=j,\boldsymbol{z}_{-i},\boldsymbol{w}_{-i})P(z_i=j\mid \boldsymbol{z}_{-i})\Rightarrow\tag{2}$$

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

根据文献[18],式(3)的右边分别为

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

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

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

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 $P(z_i=j\mid \boldsymbol{z}_{-i},w_i)$;箭头③为根据 $P(z_i=j\mid \boldsymbol{z}_{-i},w_i)$ 对 z_i 进行 Gibbs 取样并修改 z_i 的过程.当 $P(z_i=j\mid \boldsymbol{z}_{-i},w_i)$ 收敛时结束此过程,并将 $P(z_i=j\mid \boldsymbol{z}_{-i},w_i)$ 按关键字 w_i 归一化,即可得到关键字-话题概率矩阵 $\boldsymbol{\varphi}$,其中 $\varphi_{i,j}=p(z_i=j\mid w=i)$,$\boldsymbol{\varphi}_{i,\cdot}=1$.

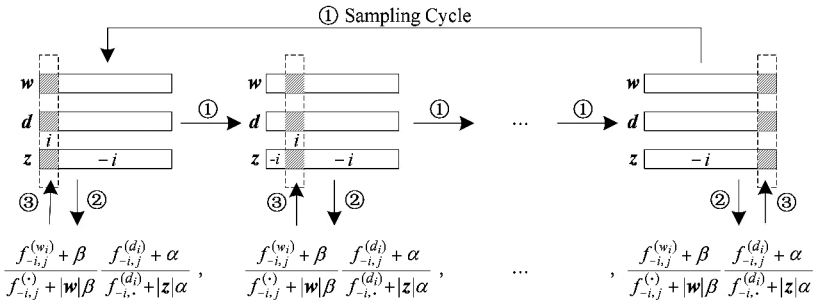


Fig. 3　The process of Gibbs sampling.

图 3　Gibbs 取样过程

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,向量 $\boldsymbol{\varphi}_{i,\cdot}$ 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 个关键字的加权均值形式表达,其表达式为

$$\boldsymbol{m}_i = \sum_{j=1}^{N_i} \boldsymbol{\varphi}_{N_{i,j},\cdot} \Big/ N_i, \tag{7}$$

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.

利用式(7)可量化表示各个节点的语义信息，因此，各节点间的语义相似性可通语义坐标 $\boldsymbol{m}_i$ 进行度量. 根据式(7)可知语义坐标 $\boldsymbol{m}_i$ 各元素之和为 1，因此语义坐标可看作各节点信息在语义空间的分布，为此本文选择度量分布相似性的相对熵(KL 散度)作为节点间语义相似度的度量，即相对熵 $KL(\boldsymbol{m}_i, \boldsymbol{m}_j)$ 越小，$\boldsymbol{m}_i, \boldsymbol{m}_j$ 的相似性越高，由于相对熵具有非对称性 $KL(\boldsymbol{m}_i, \boldsymbol{m}_j) \neq KL(\boldsymbol{m}_j, \boldsymbol{m}_i)$，对此选择 $(KL(\boldsymbol{m}_i, \boldsymbol{m}_j) + KL(\boldsymbol{m}_j, \boldsymbol{m}_i))/2$ 作为节点 $\boldsymbol{m}_i$ 和节点 $\boldsymbol{m}_j$ 的相似度 h_{ij}，其表达式如下：

$$h_{ij} = \frac{KL(\boldsymbol{m}_i, \boldsymbol{m}_j) + KL(\boldsymbol{m}_j, \boldsymbol{m}_i)}{2} = \frac{1}{2}\sum_{t=1}^{k}\left(m_{i,k}\ln\frac{m_{i,k}}{m_{j,k}} + m_{j,k}\ln\frac{m_{j,k}}{m_{i,k}}\right). \tag{8}$$

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(包含 108 篇论文中 155 条引用关系). 本文算法分别在每篇论文的摘要中抽取 6 个关键字作为论文节点的语义信息，以话题个数 $k=5$ 进行 Gibbs 取样迭代后，再利用式(7)～(8)计算 KL 相似度矩阵 $\boldsymbol{H}$，其结果如图 4 所示，网络模型如图 5 所示.

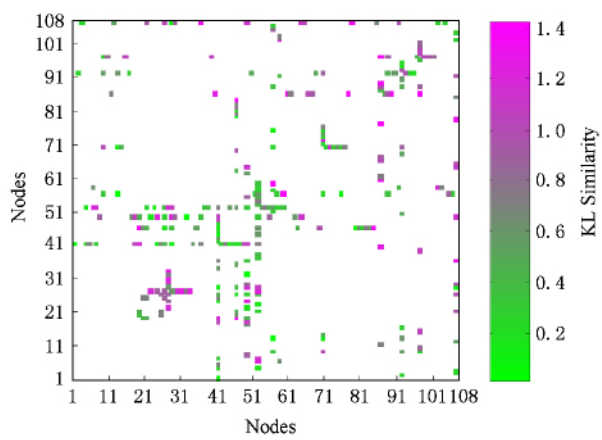


Fig. 4 The similarity matrix $\boldsymbol{H}$ on QLSP network.

图 4 QLSP 的相似度矩阵 $\boldsymbol{H}$

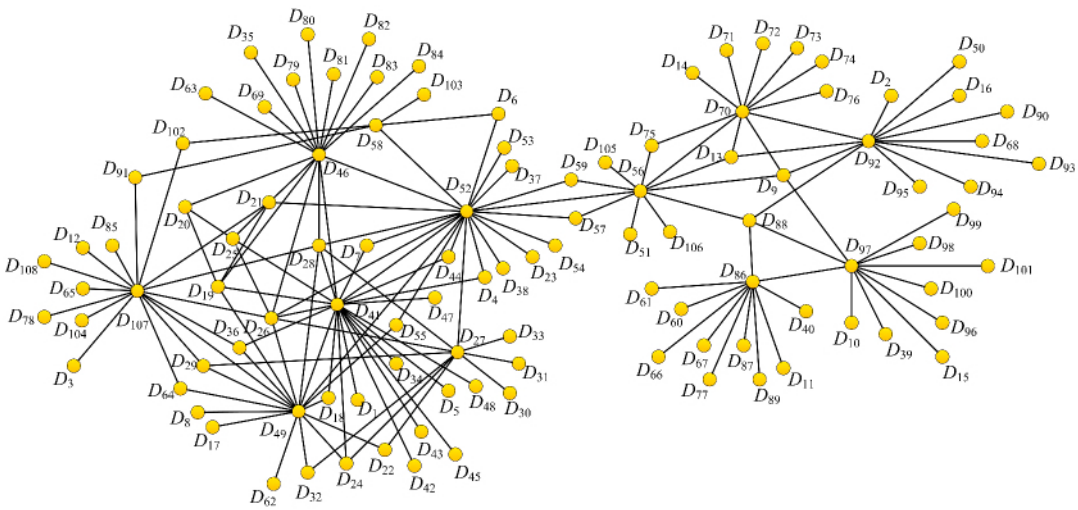


Fig. 5 The topology of QLSP network.

图 5 QLSP 网络拓扑

3 局部语义聚类算法

由于社会网络具有局部小世界特性，社区网络中的社区结构的规模远小于整体网络规模，从而使社区划分的结果对网络整体拓扑的依赖较小，因此通过局部聚类方法所挖掘出的社区结构具有较强的社区独立性，且算法复杂度较低，可面向大规模网络. 文献[6]提出了度量局部社区结构 fitness 模型，其表达式为

$$F_C = \frac{R_{\text{in}}^C}{(R_{\text{in}}^C + R_{\text{out}}^C)^{\eta}}, \tag{9}$$

其中，F_c 为社区 C 的 fitness，R_{in}^C 为社区 C 的内部度数，R_{out}^C 为社区 C 的外部度数(社区 C 与其它社区的链接数)，η 为社区规模控制参数 $\eta(0.5<\eta<2)$.

为使局部社区结构满足局部语义特性，可根据节点间的语义相似性构造语义社会网络节点度分布模型 SR，由于语义相似度矩阵 $\boldsymbol{H}$ 中 h_{ij} 越小，节点

D_i 和 D_j 的相似性越高,可通过利用 $\exp(-h_{i,j})$ 作为节点 D_i 和其邻接节点 D_j 紧密程度的度量,因此节点 D_i 的度分布可表示为如下形式:

$$SR_i = \{\exp(-h_{i,j}) \mid h_{i,j} \neq 0\}. \tag{10}$$

语义社会网络的局部社区结构 S-fitness 模型可表示如下:

$$SF_C = \frac{SR_{\text{in}}^C}{(SR_{\text{in}}^C + SR_{\text{out}}^C)^{\omega}}, \tag{11}$$

其中, $SR_{\text{in}}^C = \sum_{i=1}^{|G|}\sum_{j=1}^{|G|} h_{i,j}, D_i, D_j \in C, SR_{\text{out}}^C = \sum_{i=1}^{|G|}\sum_{j=1}^{|G|} h_{i,j}, D_i \in C, D_j \in G$, ω 为语义社区规模控制参数,本文在实验章节对 ω 的取值进行了实验分析.

设 C' 为与社区 C 直接相邻的节点集合,若某一节点 $D_i \in C'$,则节点 D_i 相对于社区 C 的聚合度可表达如下:

$$SF_C^i = SF_{C+\{D_i\}} - SF_C, \tag{12}$$

其中 $C+\{D_i\}$ 表示节点 D_i 加入社区 C 后所形成的新社区;若某一节点 $D_i \in C$,则节点 D_i 相对于社区 C 的聚合度可表达如下:

$$SF_C^i = SF_C - SF_{C-\{D_i\}}, \tag{13}$$

其中 $C-\{D_i\}$ 表示节点 D_i 离开社区 C 后,所形成的新社区.

通过以上分析,局部语义聚类算法 LSC 的简要执行步骤如下:

1) 随机选择某一未被划分的节点 D_i,以该节点作为初始社区 $C=\{D_i\}$,与 C 直接相邻的节点加入 C';

2) 按式(12)计算 C' 中与 C 聚合度最大的节点 $D_j(SF_C^j=\max(SF_C^{C'}))$,若 $SF_C^j<0$ 则转 4),否则将 D_j 加入 C 并更新 C' 转 3);

3) 按式(13)计算 C 内聚合度最小的节点 $D_t(SF_C^t=\min(SF_C^C))$,若 $SF_C^t<0$ 则将 D_t 从 C 中移除并更新 C' 转 3),否则转 2);

4) C 中的节点被划分为同一社区,若此时存在未被划分的节点则转 1),否则结束.

以上步骤中 1)为语义社区的初始选择;2)为语义社区的局部扩散过程;3)为语义社区扩散后的内部调整过程;4)为语义社区经过局部扩散和内部调整后的稳定过程. 经过以上的局部语义聚类过程,可实现语义社会网络的重叠社区发现. 图 6 为本文 LSC 算法的社区划分结果(参数 $\omega=0.92$).

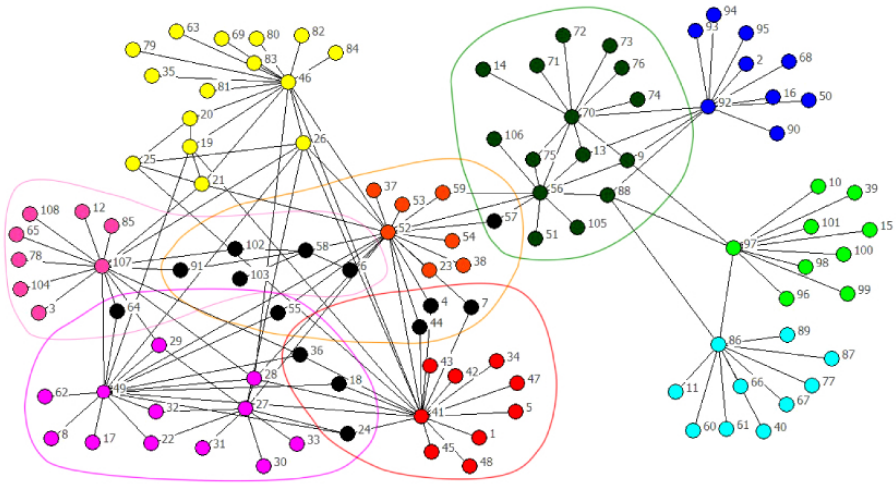


Fig. 6　The semantic community of QLSP network.

图 6　QLSP 网络语义社区划分结果

4　语义重叠社区评价模型

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 为

$$EQ = \frac{1}{X}\sum_t \sum_{i\in C_t, j\in C_t} \frac{1}{O_i O_j}\left[A_{i,j} - \frac{R_i R_j}{X}\right], \tag{14}$$

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 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 $U(\boldsymbol{m}_i, \boldsymbol{m}_j)$,建立可评价标语义重叠社区的模块度模型 SQ,其表达式为

$$SQ = \frac{1}{X}\sum_t \sum_{i\in C_t, j\in C_t} \frac{U(\boldsymbol{m}_i, \boldsymbol{m}_j)}{O_i O_j}\left[A_{i,j} - \frac{R_i R_j}{X}\right]. \tag{15}$$

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:

$$U(\boldsymbol{m}_i,\boldsymbol{m}_j)=\frac{\boldsymbol{m}_i\cdot\boldsymbol{m}_j}{|\boldsymbol{m}_i||\boldsymbol{m}_j|}=\sum_{g=1}^{k}m_{i,g}m_{j,g}\Big/\sqrt{\sum_{g=1}^{k}m_{i,g}\sum_{g=1}^{k}m_{j,g}}. \tag{16}$$

本文在第 5 节对 SQ 进行实验分析.

5 实验分析

5.1 ω 取值分析

规模控制参数 ω 是语义聚类算法的输入参数,为验证规模控制参数 ω 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 7 所示的 Krebs 建立的美国政治之书网络(Krebs polbooks network),该数据的网络结点代表亚马逊网上书店卖出的有关美国政治的图书,每本书的政治倾向略有不同,但总体上分为 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 D_i 具有直接相邻关系(距离为 1)的节点 D_j 和间接相邻关系(距离为 2)节点 D_k 的信息向量之和作为节点 D_i 的信息向量;

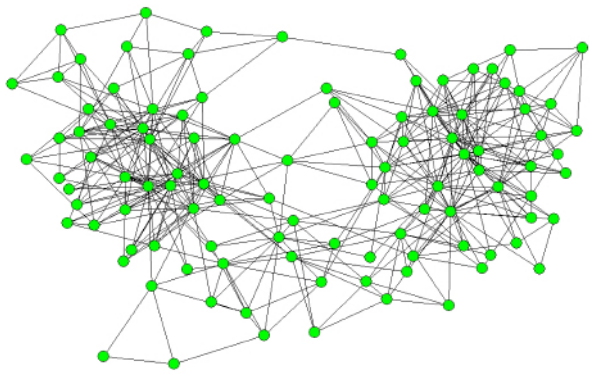


Fig. 7 The topology of Polbooks network.

图 7 Polbooks 网络拓扑图

3) 图 8 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)进行规模控制参数,由于当 $\omega>2$ 时已无明显的社区结构,因此本节实验选择区间(0.01~2)作为参数 ω 的测试区间,3 组数据的社区个数和 SQ 值如图 9 和图 10 所示,其中图 9 说明了 ω 取值增大时,社区个数会逐渐增大且单一社区规模逐渐变小,且当 $\omega<0.4$ 时无明显社区结构. 图 10 为 ω 不同取值下 3 组数据的 SQ 值比较,其中 3 组数据的最优 ω 取值分别为 0.92, 1.02, 0.89, SQ 最

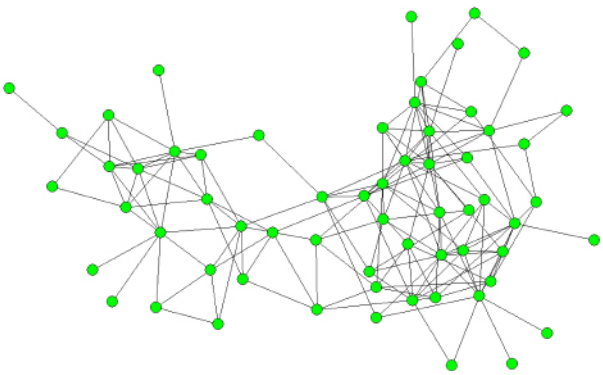


Fig. 8 The topology of Dolphins network.

图 8 Dolphins 网络拓扑图

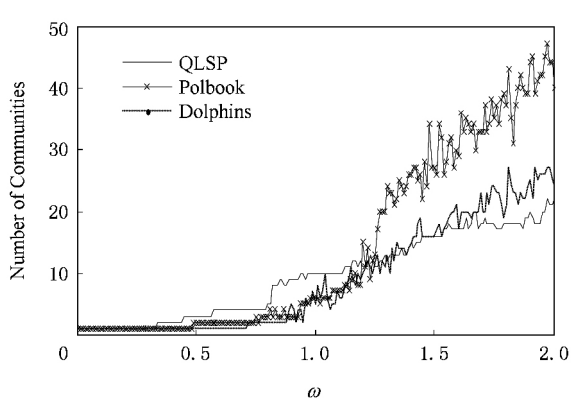


Fig. 9 The comparison chart on community number.

图 9 3 组数据的社区个数对比图

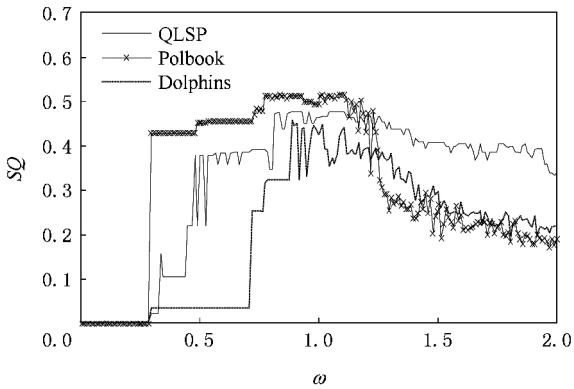


Fig. 10 The comparison chart on SQ for the three datasets.

图 10 3 组数据的 SQ 值对比图

大值分别为 0.4766,0.5159,0.4548.从图 10 的 3 组折线对比可知,当 $0.8<\omega<1.2$ 时 LSC 算法的 SQ 值最高,所划分的社区结构最合理.

为对比 ω 不同取值的语义社区划分结果,图 11 分别选取了 3 组数据在参数 $\omega=(0.6,1.0,1.4)$ 下的社区划分结果,由于 Dolphins 网络 $\omega=0.6$ 时社区个数为 1,因此 Dolphins 网络选择 $\omega=(0.8,1.0,1.4)$,图 11 直观验证了当 ω 过小时,单一社区规模较大,结合图 9 和图 10 可知由于规模较大的社区其局部信息紧密度越低,因此社区划分结果 SQ 值较低;当 ω 过大时单一社区规模较小,社区的结构特性不明显 SQ 值同样较低.综合本小节实验分析结果,LSC 算法的规模控制参数 ω 以 0.8～1.2 为最佳,过大或过小均会影响社区划分质量.

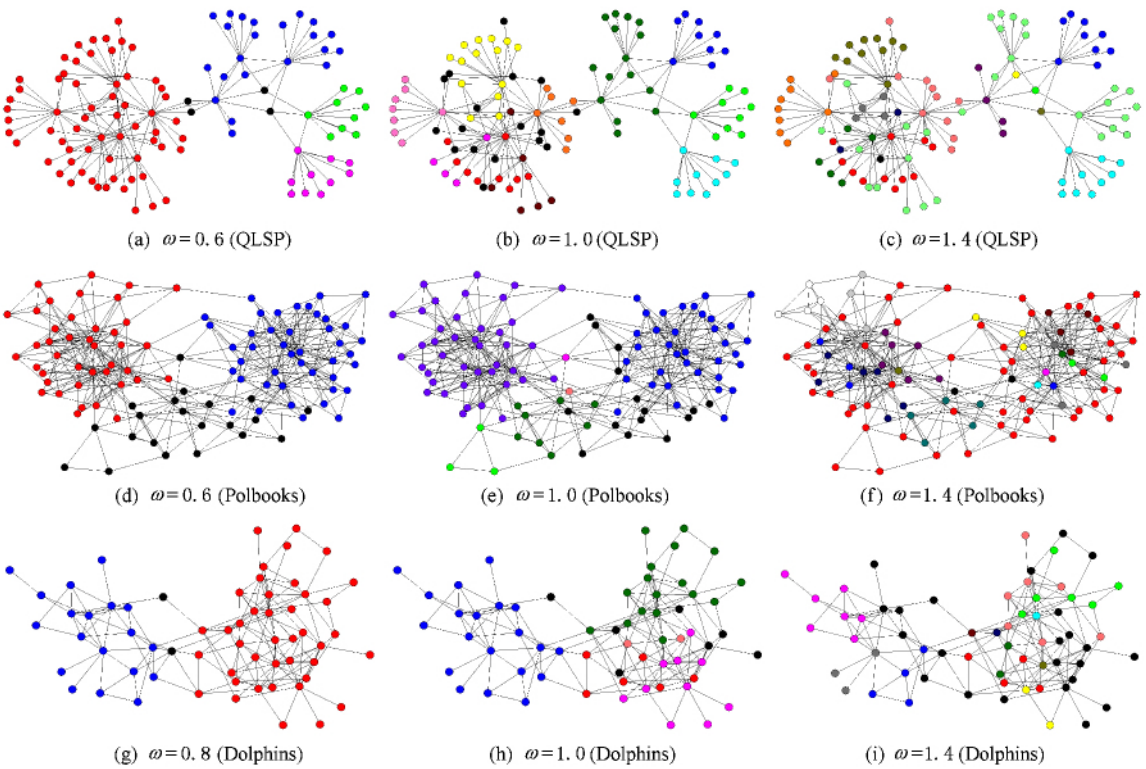


Fig. 11　The community structures with different ω.

图 11　ω 在不同取值下的网络划分结果

5.2　SQ 与 EQ 的比较分析

本节实验计算了实验 1 中 3 组数据 ω 不同取值条件下的 EQ 值,3 组数据的 EQ 和 SQ 值的对比如图 12 所示.从式(14)和式(15)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.QLSP 数据集的 EQ 最大值为 0.6008 所对应的 $\omega=1.11$,与 SQ 的最大值为 0.4766(对应的 $\omega=0.92$)存在偏差,而 $\omega=0.92$ 所对应的社区结构的 EQ 值为 0.5236.对于 Polbooks 和 Dolphins 数据集的 EQ 最大值分为 0.5137($\omega=0.84$)和 0.5276($\omega=1.02$),与 SQ 的最大值 0.5159($\omega=1.02$)和 0.4548($\omega=0.89$)同样

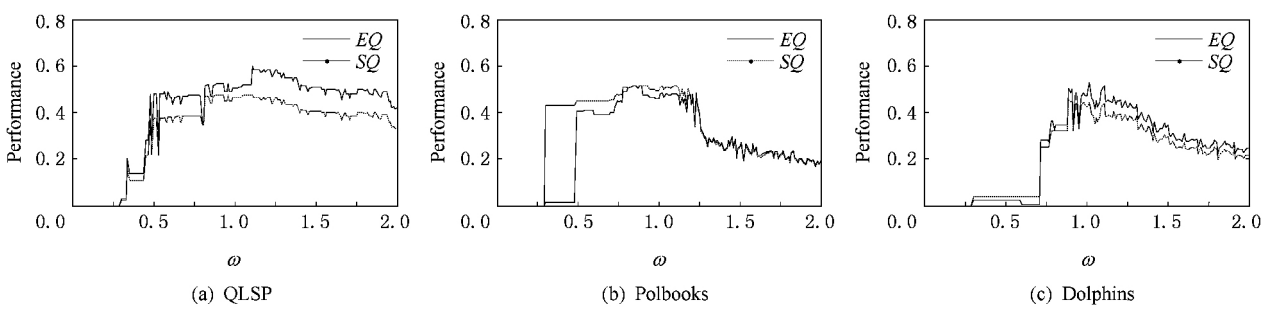


Fig. 12　The comparison figure of EQ and SQ.

图 12　EQ 和 SQ 对比结果

存在偏差. 其结果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效果较差.

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面向语义社会网络时,划分结果存在偏差,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[2], FN[3], LFM[6], COPRA[7], UEOC[8], EAGLE[5], CPM[4], 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 13 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. 以上经典算法以链接关系优化划分为导向,从表 1 中的结果可分析出,经典算法的 *EQ* 值高于本文算法(0.523 6),但 *SQ* 值均低于本文算法(0.476 6,$\omega=0.92$),由此验证了,传统面向链接关系的社区划分算法 *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.

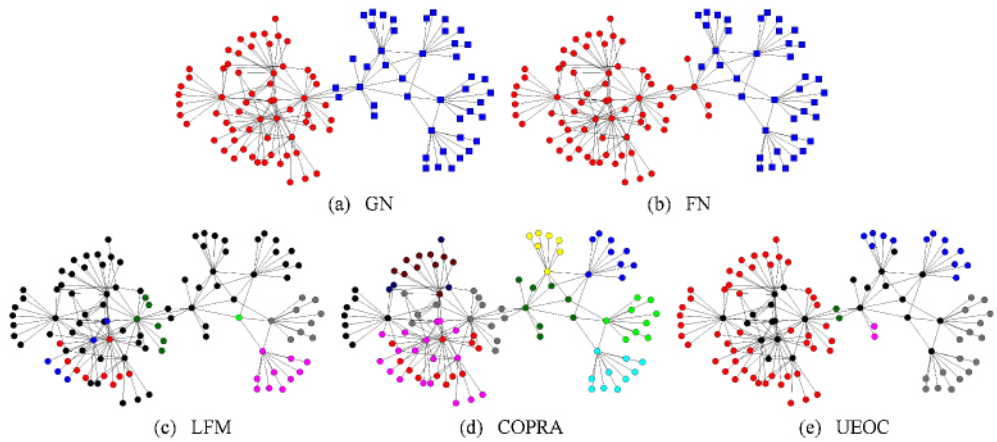


Fig. 13 The community results from classical algorithms.

图 13 各算法的社区划分结果

Table 1 *SQ* and *EQ* of Classical Algorithms

表 1 经典算法的 *SQ* 和 *EQ* 值

Method	*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
LSC	0.476 6	0.523 6

5.4 真实数据集比较

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(April 12,2006)数据集(1 200 000 个节点)中分别截取 1 500 个节点数据集和 2 000 个节点数据集作为实验数据,分别记作 DBLP(A)和 DBLP(B). 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,其中本文 LSC 算法的运行

Table 2 The Results of Classical Algorithms on Various Datasets

表 2 各数据集的执行结果

Method	Measurement	QLSP	AFD	CND	DBLP (A)	DBLP (B)
GN	*EQ*	0.31	0.13	0.19	0.28	0.32
	SQ	0.23	0.16	0.18	0.21	0.28
	CS	10	25	39	17	16
FN	*EQ*	0.42	0.15	0.22	0.32	0.26
	SQ	0.32	0.14	0.17	0.29	0.25
	CS	10	27	37	19	16
LFM	*EQ*	0.37	0.15	0.24	0.41	0.36
	SQ	0.32	0.13	0.22	0.33	0.31
	CS	12	24	33	22	12
COPRA	*EQ*	0.42	0.32	0.11	0.38	0.41
	SQ	0.29	0.22	0.12	0.29	0.32
	CS	13	21	35	21	13
UEOC	*EQ*	0.38	0.23	0.26	0.36	0.32
	SQ	0.32	0.22	0.23	0.29	0.20
	CS	12	24	30	22	14
LSC	*EQ*	0.32	0.24	0.20	0.35	0.30
	SQ	0.35	0.26	0.27	0.36	0.37
	CS	14	25	34	23	15

参数为 $\omega=0.96$(在 0.8～1.2 中随机选择),表 2 包括 EQ,SQ 及社区个数 CS,图 14 和图 15 分别为各算法的 EQ 和 SQ 直方图. 由于不同的数据集的社区结构不同,因此,图 14 和图 15 所示的各数据集的 EQ 和 SQ 不一致. 在同类算法的 EQ 和 SQ 对比中,图 14 的结果表示本文 LSC 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 LSC 算法的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 LSC 算法更适合处理语义社会网络的社区发现问题.

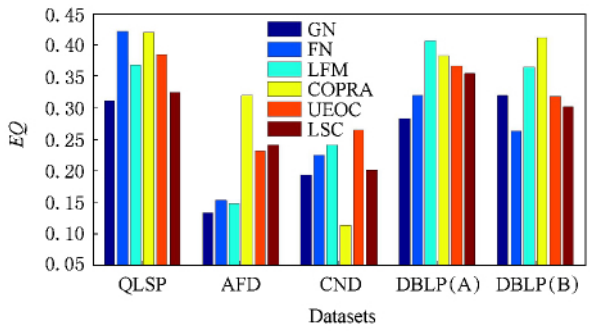


Fig. 14　The histogram of EQ under different classical algorithms.

图 14　各算法的 EQ 直方图

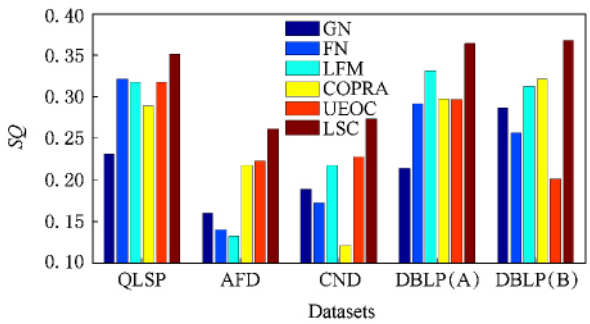


Fig. 15　The histogram of SQ under different classical algorithms.

图 15　各算法的 SQ 直方图

5.5　语义社区发现算法对比分析

本节实验对比各类需要预先设定社区个数的语义社区发现算法,以语义社区发现算法中通用的 Enron 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 MB 数据量. 表 3 为经 LDA 分析后从 Enron 数据集中抽取的 4 组话题. 表 4 和表 5 分别为 Enron 数据集分别在 TURCM,CART,CUT 和 LCTA 算法下的 EQ 值及 SQ 值,表中社区个数表示各算法执行前的社区预设数. 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.

Table 3　Topics Extracted from Enron

表 3　Enron 数据集的话题分组

Topic	Power	Gas	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) LSC 算法的最优参数取值为 $\omega\in(0.8,1.2)$;

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,LSC 相对于经典重叠社区发现算法更有效;

4) LSC 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,LSC 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 LSC 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)以节点间语义坐标的相对熵(KL 散度)作为节点语义相似度的度量;3)提出语义社会网络的局部社区结构 S-fitness 模型,并根据 S-fitness 模型建立了局部语义聚类算法 LSC;4)提出了评价语义社区划分结果的 SQ.

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,LSC 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 Bo, Liu Dayou, Jin Di, et al. Complex network clustering algorithms [J]. Journal of Software, 2009, 20(1): 54-66 (in Chinese)
(杨博, 刘大有, 金弟, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66)

[2] 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

[3] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 1-12

[4] 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

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

[6] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 1-10

[7] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2010, 12(10): 1-11

[8] 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(5): 1-11

[9] Jin Di, Yang Bo, Liu Jie, et al. Ant colony optimization based on random walk for community detection in complex networks [J]. Journal of Software, 2012, 23(3): 451-464 (in Chinese)
(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[J]. 软件学报, 2012, 23(3): 451-464)

[10] Gan Wenyan, He Nan, Li Deyi. Community discovery method in networks based on topological potential [J]. Journal of Software, 2009, 20(8): 2241-2254 (in Chinese)
(淦文燕, 赫南, 李德毅. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254)

[11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 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 [C] //Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2007: 200-207

[13] 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: SIAM (Society for Industry and Applied Mathematics), 2010: 754-765

[14] 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

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

[16] 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

[17] 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

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

[19] 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: 249-272

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

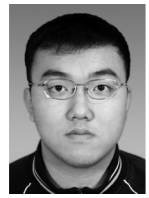
[26] Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 331-340

[27] 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 interests include database and knowledge engineering, enterprise intelligence computing.

Yang Jing, born in 1962. Professor at Harbin Engineering University and PhD supervisor. Senior member of China Computer Federation. Her research interests include database and knowledge engineering, enterprise intelligence computing.

Tang Chuheng, born in 1994. Undergraduate at Harbin Institute of Technology. Her main research interests include social network analysis, enterprise intelligence computing(Tangchuheng@hrbhit.edu.cn).



Ge Siqiao, born in 1993. Undergraduate at Harbin Engineering University. Student member of China Computer Federation. His main research interests include social network analysis, enterprise intelligence computing(gesiqiao@hrbhit.edu.cn).