软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(2):363−380 [doi: 10.13328/j.cnki.jos.004810]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种面向语义重叠社区发现的 Link-Block 算法*

辛 宇[1,2], 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,黑龙江 哈尔滨 150001)

[2](哈尔滨理工大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

通讯作者: 杨静, E-mail: yangjing@hrbeu.edu.cn, http://www.hrbeu.edu.cn

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,传统语义社会网络分析算法在进行社区挖掘时需要预先设定社区个数,且无法发现重叠社区.针对这一问题,提出一种面向语义社区发现的 link-block 算法.该算法首先以 LDA 模型为语义信息模型,创新性地建立了以 link 为核心的 block 区域 LBT(link-block-topic)取样模型;其次,根据 link-block 语义分析结果,建立可度量 link-block 区域的语义链接权重方法,实现了语义信息的可度量化;最后,根据语义链接权重建立了以 link-block 为单位的聚类算法以及可评价语义社区的 SQ 模型,并通过实验分析,验证了该算法及 SQ 模型的有效性及可行性.

关键词: 语义社会网络;重叠社区;语义社区;LDA;link-block

中图法分类号: TP311

中文引用格式: 辛宇,杨静,谢志强.一种面向语义重叠社区发现的 Link-Block 算法.软件学报,2016,27(2):363−380. http://www.jos.org.cn/1000-9825/4810.htm

英文引用格式: Xin Y, Yang J, Xie ZQ. Link-Block method for the semantic overlapping community detection. Ruan Jian Xue Bao/Journal of Software, 2016,27(2):363−380 (in Chinese). http://www.jos.org.cn/1000-9825/4810.htm

Link-Block Method for the Semantic Overlapping Community Detection

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

Abstract: Since the semantic social network (SSN) is a new kind of complex networks, the traditional community detection algorithms which require presetting the number of the communities, cannot detect the overlapping communities. To solve this problem, an overlapping community structure detecting algorithm in semantic social networks based on the link-block is proposed. First, the measurement of the semantic weight of links for the link-block is established depending on the analysis of LBT. Secondly, a method to measure the semantic links weight of link-block area is developed to provide the measurement of semantic information. Thirdly, the overlapping community detection cluster method is designed, based on the semantic weight of links, with the link-block as the element. Finally, the SQ modularity for the measurement of semantic communities is obtained. The efficiency and feasibility of the algorithm and the semantic modularity are verified by experimental analysis.

Key words: semantic social network; overlapping community; semantic community; LDA; link-block

* 基金项目: 国家自然科学基金(61370083, 61370086); 教育部博士点基金(20122304110012); 黑龙江省博士后基金(LBH-Z15096); 黑龙江省教育厅科技项目(12531105); 黑龙江省博士后科研启动项目(LBH-Q13092)

Foundation item: National Natural Science Foundation of China (61370083, 61370086); Ph.D. Programs Foundation of Ministry of Education of China (20122304110012); Postdoctoral Foundation of Heilongjiang Province of China (LBH-Z15096); Science and Technology Program of Education Bureau of Heilongjiang Province of China (12531105); Postdoctoral Scientific Research Staring Foundation of Heilongjiang Province of China (LBH-Q13092)

收稿时间: 2014-04-20; 修改时间: 2014-06-06, 2014-11-17; 定稿时间: 2014-12-26

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.为了丰富用户的 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 等人提出了具有重叠(overlapping)特性的社区结构,并设计了面向重叠社区发现的 CPM 算法[4].此后,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.

由于社会网络的交流需要以文本作为载体,文本舆情的分类结果可指导社区的划分[11].为此,语义社区划分的出发点在于,根据社会网络中节点语义信息内容(如微博、社会标签等)的分类结果,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[12]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 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[13].由于 Henderson 等人在 SSN-LDA 模型的基础上融入了 IRM(infinite relational model)[14]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[15].随后,Henderson 等人在 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[16].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[17]及面向层次化分的 HSN-PAM[18]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.此类算法大多以 AT 模型[19]作为 LDA 分析的基本模型,代表算法有 McCallum 等人[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 等人在 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[21].Zhou 等人在 AT 模型中加入了 user 分布取样,提出了 CUT 模型[22].Cha 等人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[23].Nagarajan[24]通过采集用户间传递的信息以及用户间的共享信息,建立了一种 content sharing network,并提出了 Link-Content model 模型.该模型是一种以链接为研究对象的 AT 模型,在 LDA 取样过程中,根据链接关系加入了 user 选择 community 的过程.Sebastian[25]以传统社会网络中的标签传播算法为基础,提出了 SLTA 算法.SLTA 算法将话题作为标签的内容,并在标签传播的“听-说”过程中加入了 LDA 分析与节点关联性分析.Hu[26]将 AT 衍生出 FT (feature topic)和 ST(social topic)模型,以用户的话题倾向属性进行社区倾向性分析.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[27]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[28]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[29].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,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[30,31]及 LCTA 模型[32],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.Zhao 等人[33]提出的 EWKM (entropy weighting *K*-means)方法建立了节点的语义相似度矩阵,将语义社会网络量化为有权网络.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.

图 1 为各类基于 LDA 模型算法的关联关系,其中,ART 强调邻居的语义信息对节点的影响,而 HLDA 强调全局语义信息对节点的影响.本文在二者的基础上强调社区对节点的影响,为此提出了 LBT(link-block-topic)算法.LBT 算法融合了 ART 算法 recipient 取样及 HLDA 的 hierarch 取样方式.由于 ART 算法 recipient 取样的影响范围较小且 HLDA 没有考虑距离对取样的影响,为此,本文在 ART 算法与 HLDA 算法的基础上增加了 Link 取样,以扩大取样的影响范围;在 HLDA 基础上增加了 Block 取样,以满足社区的局部区域特性.

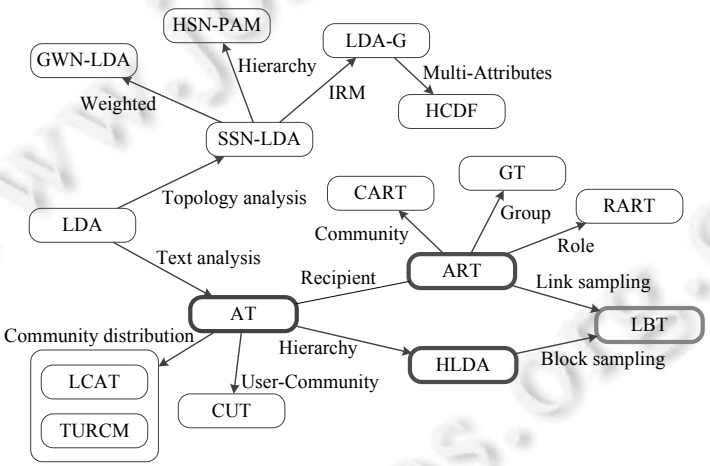


Fig.1 Relevance of various LDA-based models

图 1 各类基于 LDA 的模型关系

语义社会网络的语义社区发现算法需要兼顾两个方面的条件:

(1) 语义社区内部链接关系紧密;

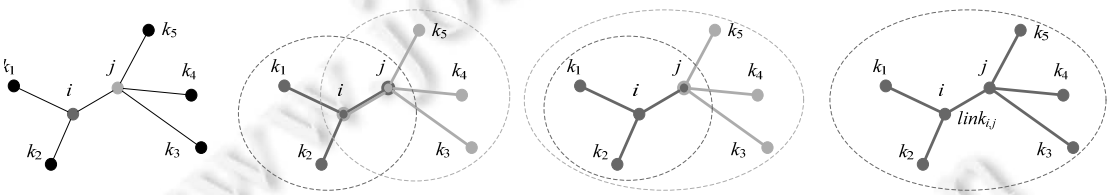
(2) 语义社区内部节点的语义信息相似度高.

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(LSC)及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的有效参数选取及 SQ 评价性能.

1 Link-Block-Topic 关系建模

对于有代表性半监督语义社区发现算法,如 AT,ART 及 HLDA 分别为“点、面、放射”的形式对语义网络中的节点进行文本取样,其文本话题生成过程(以节点 i,j 为例)的差别如图 2 所示,其中,

- 图 2(a)为 AT 模型的文本生成过程.在全局文本分析时,分别对节点 i,j 进行单独取样,不考虑网络拓扑的相关性,其文本生成过程以 author 为单位.
- 图 2(b)为 ART 模型的文本生成过程.由于 ART 模型以 recipient 作为 author 的桥梁,在对节点 i 进行文本生成取样时,加入了 i 相邻节点 k_1,k_2,j 的取样,并在对节点 i 进行文本生成时加入了 j 相邻节点 k_3, k_4,k_5,i 的取样,其文本生成过程以 author 的中心区域为单位.
- 图 2(c)为 HLDA 模型的文本生成过程.由于 HLDA 模型以分层的方式进行文本生成取样,在对节点 i 进行文本生成时,将与节点 i 直接相邻的节点 k_1,k_2,j 作为一次取样,与节点 i 间接相邻的节点 k_3,k_4,k_5 进行二次取样,其文本生成过程是以 author 的放射区域为单位;ART 模型和 HLDA 模型是 AT 在网络拓扑关系中的应用,ART 模型的文本生成取样区域半径为 1,导致以 author 为中心的区域规模较小,文本生成取样结果仅代表直接邻接关系,不具有社区的规模关系特性;HLDA 模型的文本生成取样过程是对放射区域的无权取样,忽略了社会网络中关系距离对社区成员间的影响.为此,本文将 ART 与 HLDA 的取样思想相结合,以 link 为中心的关系区域(block)作为取样区域,设计了 LBT(link-block- topic)文本取样模型.
- 图 2(d)为本文 LBT 模型的文本生成过程.在进行文本生成时,以 link 作为中心(如图 2(d)中 $link_{i,j}$),并将与 $link_{i,j}$ 直接相邻的节点 k_1~k_5 作为取样节点,其文本生成过程是以 link-block 为单位.



(a) AT (b) ART (c) HLDA (d) LBT

Fig.2 Context sampling process

图 2 文本取样过程

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 LBT 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 表示网络中的节点 i.
- $|G|$表示语义社会网络中节点的个数.
- L 表示网络 G 中的边(链接)集合,$L_{i,j}$ 表示节点 G_i 与节点 G_j 之间的边.
- $|L|$表示语义社会网络中边(链接)个数.
- a_L 表示与本文取样链接 L 相邻的 author 节点集合.
- x 表示集合 a_L 中抽取的一个元素.
- 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.
- D 表示语义社会网络中语料信息个数.
- $\boldsymbol{w}$ 表示关键字的集合,w_i 为集合 $\boldsymbol{w}$ 中第 i 个关键字所对应的编号.
- $\boldsymbol{z}$ 表示与关键字的集合 $\boldsymbol{w}$ 对应的话题编号集合,z_i 表示 w_i 所隶属的话题编号.
- T 表示话题个数.
- θ表示话题分布概率.
- φ表示关键字的分布.
- α表示各节点的话题分布先验参数.
- β表示某一话题内部,关键字分布的先验参数.

图 3 分别为 LDA,AT,ART 及 LBT 模型的对比关系图.

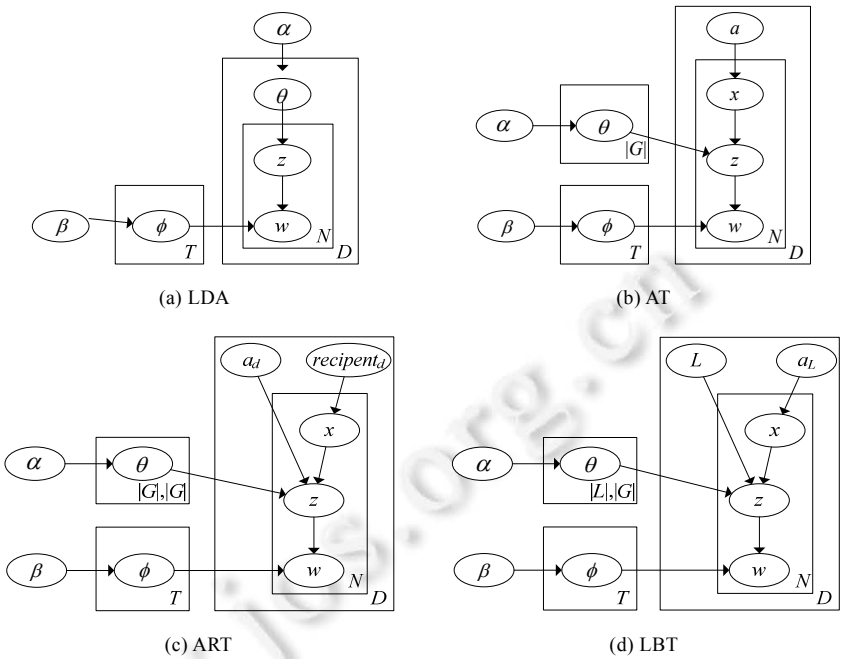


Fig.3 Comparison on semantic models

图 3 语义模型对比

其概率生成关系式如下:

- $x|a_L$~*Uniform*(a_L):表示从 a_L 集合中选作一个元素作为 L 的扩展 author.
- $z|\theta$~*Multinomial*(α):表示从给定的 L 及选定的扩展 author 本文信息中抽取一个话题,该话题服从以θ为参数的多项式分布.
- $\theta|\alpha$~*Dirichlet*(α):表示θ服从以α为参数的狄利克雷分布.
- $w|\varphi$~*Multinomial*(φ):表示话题中的关键字 w 服从以φ为参数的多项式分布.
- $\varphi|\beta$~*Dirichlet*(β):表示φ服从以β为参数的狄利克雷分布.

$$p(\theta,\varphi,x,z,w\mid\alpha,\beta,L,a_L)=\prod_{i=1}^{|L|}\prod_{j=1}^{|G|}p(\theta_{ij}\mid\alpha)\prod_{t=1}^{T}p(\varphi_t\mid\beta)\prod_{d=1}^{D}\prod_{n=1}^{N}(p(x_{dn}\mid a_L)p(z_{dn}\mid\theta_{L,x})p(w_{dn}\mid\varphi_{dn})) \tag{1}$$

其中,x_{dn}表示第 d 个语料信息中,第 n 个关键字所隶属的与 L 直接邻接的 author 号;z_{dn}表示在 L 直接相邻的节点中,第 d 个语料信息的第 n 个关键字所隶属的话题号;w_{dn}表示在 L 直接相邻的节点中,第 d 个语料信息的第 n 个关键字在语义字典中的编号.$\theta_{L,x}$和φ_{dn}分别表示生成 d 个语料信息的第 n 个关键字时,话题 z_{dn}及关键字 w_{dn}出现的概率.公式(1)中,$p(\theta_{ij}|\alpha)$表示在 link-block 区域中话题出现的概率,$p(\varphi_t|\beta)$表示被选择的话题的先验参数出现的概率,$p(x_{dn}|a_L)$表示在 link-block 区域中某一节点的文本信息的概率,$p(z_{dn}|\theta_{L,x})$表示从给定的 L 及选定的扩展 author 本文信息中抽取一个话题的概率,$p(w_{dn}|\varphi_{dn})$表示以φ为参数的关键字 w 出现概率.对公式(1)用积分表示形式,以消除变量θ和φ,得出 LBT 模型的 w 生成公式:

$$p(w\mid\alpha,\beta,L,a_L)=\iint\prod_{i=1}^{|L|}\prod_{j=1}^{|G|}p(\theta_{ij}\mid\alpha)\prod_{t=1}^{T}p(\varphi_t\mid\beta)\prod_{d=1}^{D}\prod_{n=1}^{N}\sum_{x_{dn}=1}^{|G|}(P(x_{dn}\mid a_L))\sum_{z_{dn}=1}^{T}(p(z_{dn}\mid\theta_{L,x})p(w_{dn}\mid\varphi_{dn}))\mathrm{d}\varphi\mathrm{d}\theta \tag{2}$$

经文献[21]的推导过程可知:

$$P(x_{dn},z_{dn}\mid x_{-dn},z_{-dn},w,\alpha,\beta,L,a_L)\propto\frac{\alpha_{z_{dn}}+n_{L,x_{dn},z_{dn}}-1}{\sum_{t=1}^{T}(\alpha_t+n_{L,x_{dn},t})-1}\frac{\beta_{w_{dn}}+m_{z_{dn},x_{dn}}-1}{\sum_{v=1}^{V}(\beta_v+m_{z_{dn},v})-1} \tag{3}$$

参数θ及φ后验估计为

$$\hat{\theta}_{ijz}=\frac{\alpha_z+n_{i,x,z}}{\sum_{t=1}^{T}(\alpha_t+n_{L,x,t})},\hat{\varphi}_{tw}=\frac{\beta_w+m_{t,w}}{\sum_{v=1}^{V}(\beta_v+m_{t,v})} \tag{4}$$

其中,V 为语义字典中关键字的个数,$n_{i,x,t}$ 表示第 i 号 Link-Author 对中属于话题 t 的关键字个数,m_{tv} 表示关键字 v 属于话题 t 的个数.θ_{ijz} 表示第 i 号 link 与 author G_j 的混合文本中话题 z 出现的概率.Gibbs 取样过程如下:

initialize the author and topic assignments randomly
repeat
  for d=1 to D do
    for i=1 to N do
      draw x_{dn} and z_{dn} from $P(x_{di},z_{di}|x_{-di},z_{-di},w,\alpha,\beta,L,a_L)$
      update $n_{L,x_{dn},z_{dn}}$ and $m_{z_{dn},w_{dn}}$
    end for
  end for
until the Markov chain reaches its equilibrium
calculate the posterior estimates of $\boldsymbol{\theta}$ and φ

2 语义链接权重的量化映射

本文为避免 LDA 语义分析模型中需要预先设定社区个数的问题,实现社区发现的无监督化,需要在语义网络中建立节点的量化关系,并依据量化关系建立启发式社区发现算法.为此,本节的主要内容是根据 LBT 模型的结果建立网络链接的语义链接权重度量,从语义关系及网络拓扑关系角度量化文本分析结果.

根据上一节的分析,对 LBT 模型进行 Gibbs 迭代取样后,可根据公式(4)计算 link-author 的 3 维话题概率分布 θ_{lat},其中,l 表示 link 维度,共有|L|个不同元素;a 表示与 l 直接相邻的 author 维度,共有|G|个不同元素;t 表示话题维度,包含了 T(话题个数)维向量.根据上一节的 LDA 模型分析可知,该 T 维向量表达了 link-author 的话题隶属度.由于 LBT 的文本生成过程是以 link 为核心的中心区域为单位,3 维话题概率分布 θ_{lat} 中,author 维度可作为 link 维度的从属(如图 2(d)中 $\text{link}_{i,j}$ 与 k_1~k_5 的关系),因此,可以以 link 维度作为 θ_{lat} 的主属性对 topic 维度进行加和消除 author 维度,从而将 3 维话题概率分布 θ_{lat} 转化为 2 维 θ'_{lt},即 $\theta'_{lt}=\sum_{a=1}^{|G|}\theta_{lat}$.

2 维 link-topic 矩阵 $\boldsymbol{\theta}'$ 中的 l 行元素 $\theta'_{l,\cdot}$,可看作以链接 l 为核心的区域(l-block)在语义社会网络整体结构中的话题隶属度.为衡量各 l-block 的语义凝聚力,需要建立链接 l 的语义链接权重度量,语义链接权重较大的 l-block 紧致性更强,更能体现社区的结构性.在语义链接权重度量方面,链接 l 的话题隶属度 $\theta'_{l,\cdot}$ 体现了链接 l 在全局网络中的 T 维权重.为实现 T 维权重的可度量化,本文利用 PCA 主成分加权法,将矩阵 $\boldsymbol{\theta}'$ 各行向量的相关矩阵的 T 个特征值所构成的向量 $\boldsymbol{\Lambda}=(\lambda_1,\ldots,\lambda_T)$ 作为权重向量,将 $\theta'_{l,\cdot}$ 与 $\boldsymbol{\Lambda}$ 的内积作为链接 l 的权重 W_l,由 W_l 所形成的邻接矩阵 $\boldsymbol{W}$ 即为语义网络 G 的语义链接权重邻接矩阵.

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).本文分别在每篇论文的摘要中抽取 5 个关键字作为论文节点的语义信息,以话题个数 T 为 5 进行 Gibbs 取样迭代(所提取的话题见表 1),并利用 PCA 加权法计算语义链接权重邻接矩阵 $\boldsymbol{W}$,其关系拓扑图及权重邻接矩阵分别如图 4 和图 5 所示.

Table 1 Topic groups of QLSP data set

表 1 QLSP 数据集的话题分组

Topic	1	2	3	4	5
Word	Protocol	Words	Graph	Image	Distinguish
	Route	Vocabulary	Structure	Vector	Classifier
	Topological	Model	Analysis	Retrieval	Semantic
	Asynchronously	Context	Theory	Combination	Matrix
	Homogeneous	Candidate	Engineering	Detection	Measure

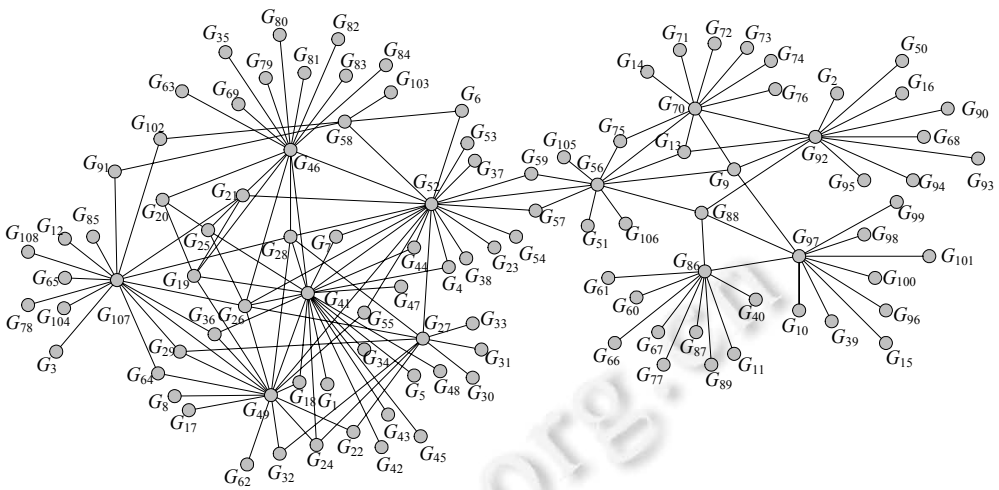


Fig.4 Weighted Topology of QLSP

图 4 QLSP 的有权拓扑

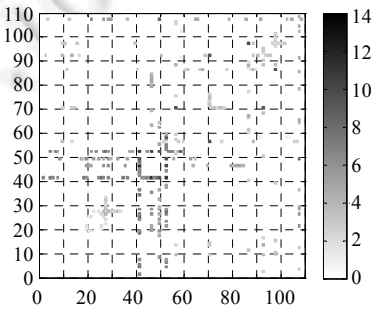


Fig.5 Weighted Topology of QLSP

图 5 QLSP 的有权拓扑

3 Link-Block 聚类算法

本节根据上一节的 LBT 量化分析结果,建立了以 block 为聚类单位的重叠社区发现算法 LBTC(LBT cluster).由上一节的分析可知,link 权重的大小表现了以 link 为核心的局部区域的语义凝聚力大小.因此,可将权重较大的 link-block 作为社区的基本局部块,并可将彼此关系紧密的局部块聚合为社区,从而实现社区发现.因此,经语义链接权重的量化映射所得出的语义网络 G 的有权邻接阵 $\boldsymbol{W}$ 可作为语义凝聚力大小的度量,即,W_{ij} 表示 $link_{i,j}$ 的 link-block 的语义凝聚力.量化聚类算法的描述过程,对 link-block 作如下定义:

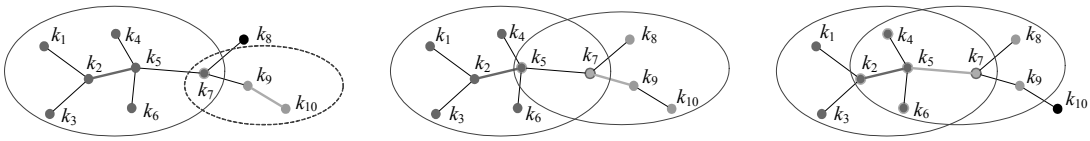
定义 1(核心链接). 核心链接是指 link-block 的中心链接如图 2(d)中的链接 $link_{i,j}$,*link-block*(*i*,*j*)表示以链接 $link_{i,j}$ 作为核心链接的 link-block.

定义 2(核心节点). 核心节点是指核心链接的端点.

定义 3(边缘链接). 边缘链接是指 link-block 中的非核心链接.

定义 4(边缘节点). 边缘节点是指 link-block 中的非核心节点.

图 6 为 link-block 的 3 种相交关系示例,其中,图 6(a)中 *link-block*(2,5)与 *link-block*(9,10)的相交部分仅含边缘节点 k_7,这种情况称为边缘节点相交;图 6(b)中 *link-block*(2,5)与 *link-block*(7,9)的相交部分包含了二者的边缘链接 $link_{5,7}$,这种情况称为边缘链接相交;图 6(c)中 *link-block*(2,5)与 *link-block*(5,7)的相交部分包含了二者的核心链接及核心节点,这种情况称为核心相交.



(a) 边缘节点相交 (b) 边缘链接相交 (c) 核心相交

Fig.6 3 kinds of intersection

图 6 3 种相交关系

由于语义链接权重表达的是以该链接作为核心链接的 link-block 的整体语义凝聚力,可将语义链接权重较大的两个相交 link-block 进行聚类合并.同时,对于如图 6(a)和图 6(b)所示的边缘节点相交及边缘链接相交情况,两个 link-block 的相交部分仅含彼此的局部语义凝聚力,不体现两个 link-block 的共同语义凝聚力,不适合作为 link-block 合并的备选情况.由此,本文以核心相交的 link-block 作为聚类的基本单位.由于核心相交的 link-block 的语义凝聚力是以 link-block 为取样区域的 LDA 求解结果,其仅代表 link-block 区域的紧致性,无法代表 link-block 合并后的聚簇紧致性,为此,本文采用非增量式聚类的形式,即,聚类过程再不涉及权重的重新计算.其聚类过程如下:① 按语义链接权重对 link-block 进行降序排序,其结果为 link-block-queue;② 从 link-block-queue 中选择前 c 个 link-block,使得这 c 个 link-block 恰好覆盖整个网络;③ 以语义链接权重大小为顺序,合并 c 个 link-block 中核心相交的 link-block,将各个聚类簇作为社区,并将各社区间相交的边缘节点(非核心节点)作为重叠节点.QLSP 数据集的聚类过程如图 7 所示,其中,聚类簇 $c1$,$c2$ 及 $c3$ 即为所发现的社区,其聚类结果如图 8 所示.

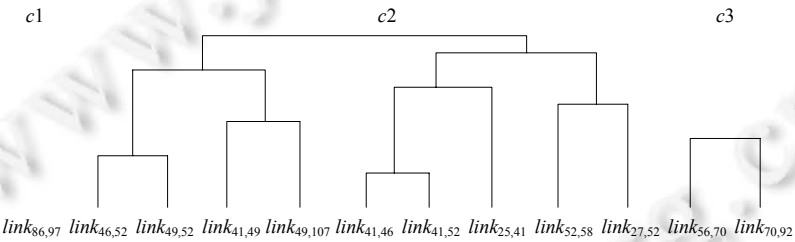


Fig.7 Clustering process of QLSP

图 7 QLSP 的聚类过程

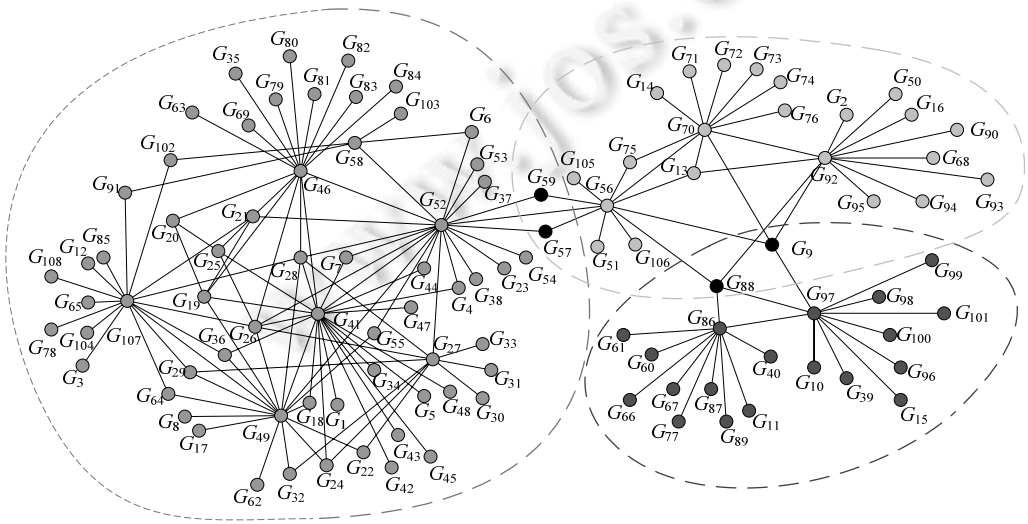


Fig.8 Clustering result of QLSP

图 8 QLSP 的聚类结果

4 语义重叠社区的评价标准

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

$$EQ=\frac{1}{X}\sum_{t}\sum_{i\in C_t,j\in C_t}\frac{1}{O_iO_j}\left[A_{i,j}-\frac{R_iR_j}{X}\right] \tag{5}$$

其中,R_i 为节点 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 i 所隶属的社区个数,C_t 表示第 t 个社区.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.由语义链接权重的量化映射分析可知,节点间的语义关系可由 2 维 link-topic 矩阵 $\boldsymbol{\theta}'$ 及语义链接权重邻接矩阵 $\boldsymbol{W}$ 表达.其中,2 维 link-topic 矩阵 $\boldsymbol{\theta}'$ 中的元素 $\theta'_{(i,j),k}$ 表示链接 $link_{i,j}$ 对第 k 个话题的隶属度,$\theta'_{(i,j),\cdot}$ 可作为 $link_{i,j}$ 在 T 维话题空间中的坐标;$W_{i,j}$ 表示 $link_{i,j}$ 的语义链接权重,可作为节点 i 与节点 j 的相似度.为此,本文根据 EQ 模型分别以 $\boldsymbol{\theta}'$ 及 $\boldsymbol{W}$ 为参数,建立可评价标语义重叠社区的模块度模型 SQ1 及 SQ2,其表达式分别为

$$SQ1=\frac{1}{Y}\sum_{t}\sum_{\substack{link(i,j)\in C_t\\ link(i',j')\in C_t}}\frac{U\left(\theta'_{(i,j),\cdot},\theta'_{(i',j'),\cdot}\right)}{B_{(i,j)}B_{(i',j')}}\left[L_{(i,j),(i',j')}-\frac{P_{(i,j)}P_{(i',j')}}{Y}\right] \tag{6}$$

其中,$P_{(i,j)}$为 $link_{i,j}$ 的度数,Y 为网络链接的总度数,$\boldsymbol{L}$ 为网络链接的邻接矩阵,$B_{(i,j)}$为 $link_{i,j}$ 所隶属的社区个数,$U(\theta'_{(i,j),\cdot},\theta'_{(i',j'),\cdot})$ 为向量 $\theta'_{(i,j),\cdot}$ 与 $\theta'_{(i',j'),\cdot}$ 的余弦相似度函数;

$$SQ2=\frac{1}{2Z}\sum_{t}\sum_{i\in C_t,j\in C_t}\frac{1}{O_iO_j}\left[W_{i,j}-\frac{S_iS_j}{2Z}\right] \tag{7}$$

其中,S_i 为与节点 i 相连的所有边的语义链接权重之和,Z 为网络中所有 link 的语义链接权重之和,$W_{i,j}$ 为 $link_{i,j}$ 的语义链接权重,O_i 为节点 i 所隶属的社区个数,C_t 表示第 t 个社区.

5 实验分析

5.1 话题个数T取值分析

话题个数 T 是本文算法(LBTC)的输入参数,为验证话题个数 T 对语义社区划分结果的影响,本文选用如下 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 9 所示的 Krebs 建立的美国政治之书网络(Krebs polbooks network),其中,每本书的政治倾向分为 3 类,每类只有 0 或 1 两种选择,因此,为实现语义化模拟,将与某一节点 i 具有直接相邻关系(距离为 1)的节点 j 和间接相邻关系(距离为 2)节点 k 的信息向量之和作为节点 i 的信息向量;③ 图 10 所示的 Newman 建立的海豚家族(dolphins network)关系网络,为模拟语义社会网络的特性,本文实验借用 polbooks 网络及 dolphins 网络的社会关系特性,并为每个节点生成 6 维随机数作为节点的语义坐标.

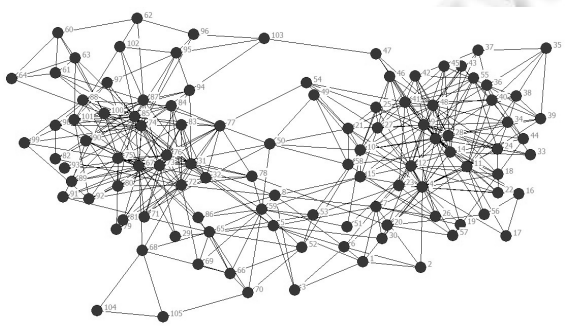


Fig.9 Topology of polbooks

图 9 Polbooks 的拓扑

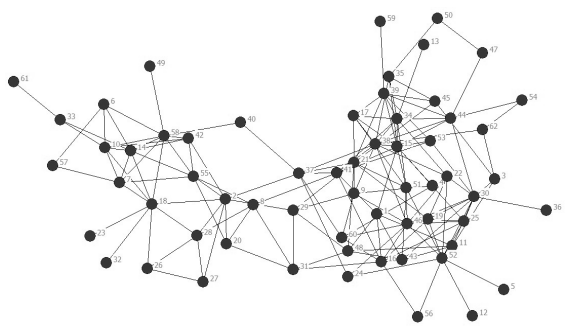


Fig.10 Topology of dolphins

图 10 Dolphins 的拓扑

图 11 为话题个数 T 在不同取值条件下,3 组数据的社区个数、EQ、SQ1 及 SQ2 对比结果.为对比话题个数 T 不同取值的语义社区划分结果,图 12 分别选取了 3 组数据在 T=(6,12,18)下的社区划分结果,其中,黑色节点为重叠节点.

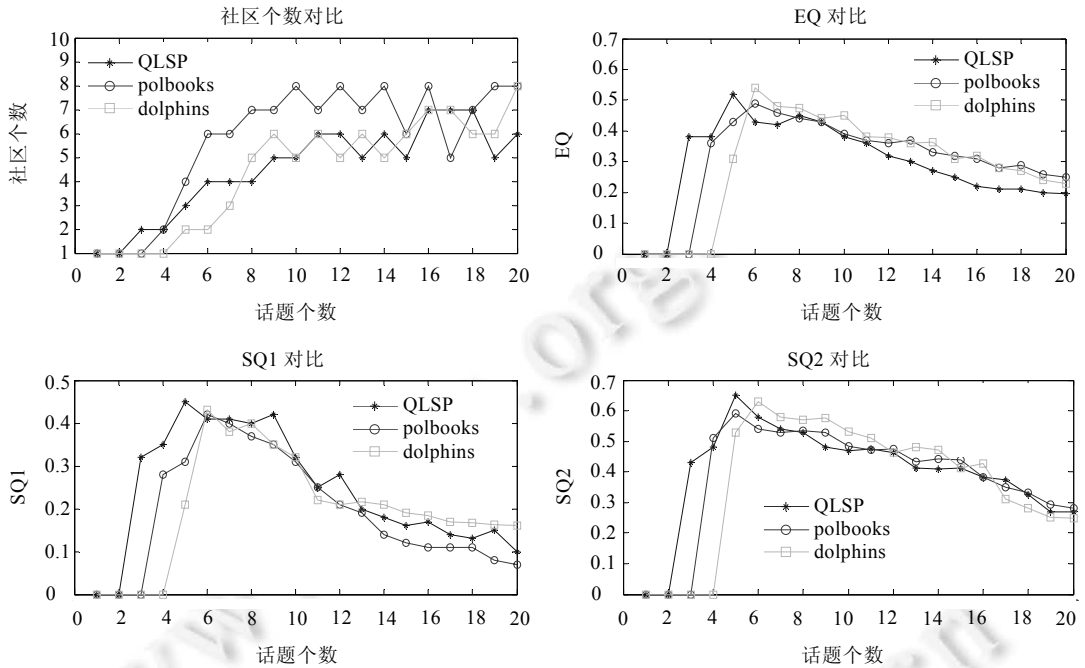
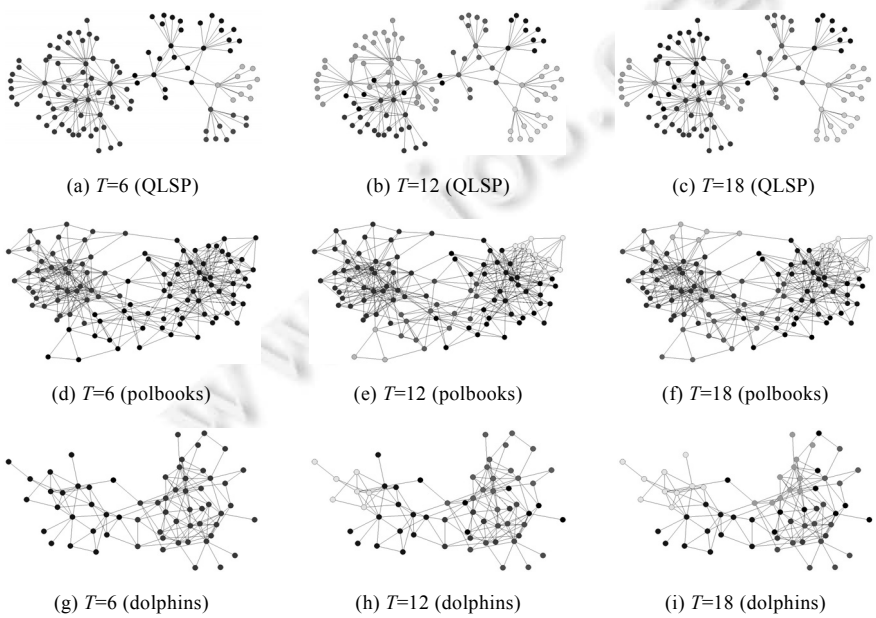


Fig.11 Comparison on the 3 datasets for topics (1~20)

图 11 3 组数据集的在话题个数为(1~20)下的对比



(a) T=6 (QLSP) (b) T=12 (QLSP) (c) T=18 (QLSP)

(d) T=6 (polbooks) (e) T=12 (polbooks) (f) T=18 (polbooks)

(g) T=6 (dolphins) (h) T=12 (dolphins) (i) T=18 (dolphins)

Fig.12 Detected communities for various T

图 12 T 不同取值下发现的社区

本实验分析了话题个数对 LBTC 的结果所产生的影响,从图 11 的对比可知:

1) 当话题个数增加时,所划分的社区个数增加.其原因在于:话题个数的增加,导致 LDA 取样分析的结果差异性增加,即全局的语义相似度的可区分性增加,有利于局部社区识别.
2) 当话题个数大于某一临界值时,EQ,SQ1 及 SQ2 出现下降趋势,说明 LDA 分析结果存在最优取值;当话题个数小于最优值时,随着话题个数的增加,全局语义分析的结果逐渐充分;当话题个数大于最优值时,随着话题个数的增加,语义分析的结果出现“过拟合”现象,其有效性下降.
3) 不同数据集的最优话题个数取值不同,即 LDA 类算法的最优参数取值需要人为经验确定,该问题为 LDA 类算法的普遍缺陷.

5.2 SQ1和SQ2的比较分析

本节实验以文献[34]的有权 benchmark 为标准,分别生成 4 组节点个数为 3 000,4 000,5 000,6 000 的网络,其社区个数分别为 215,282,348,405,并对每个链接分配 5 维权重数据,构造 2 维 link-topic 矩阵$\boldsymbol{\theta}'$,以模拟经 LDA 分析后的语义量化数据.为对比 SQ1 及 SQ2 的评价性能,分别在以上 4 组数据中增加每个社区的内部链接(所有邻接链接均与其在同一社区的链接)的权重,以相对减少社区的边界链接(非内部链接)的权重.通过以上的权重变化,可逐步增加社区语义凝聚力,其 EQ,SQ1,SQ2 取值如图 13 所示,其中,x 轴表示内部链接相对边界链接的权重倍数.

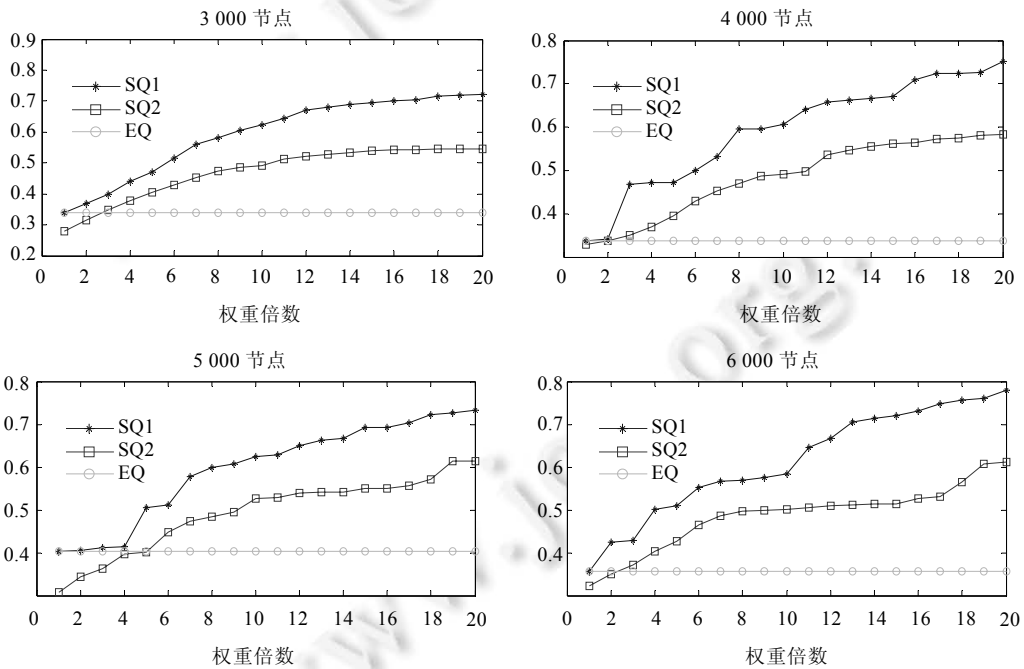


Fig.13 Comparison of EQ, SQ1 and SQ2 on 4 groups of benchmark datasets

图 13 4 组 benchmark 数据的 EQ,SQ1 及 SQ2 对比

本次实验仅增加了链接权重,各组数据的社区结构无变化,因此 EQ 取值不变.从图 13 的 4 组数据对比可知:SQ1 随着权重的增加呈收敛趋势,SQ2 呈线性递增趋势,即相对于 SQ1,SQ2 对权重的变化更加敏感,且更加符合权重的线性递增变化.因此,从对比分析可知,SQ2 相对于 SQ1 更适合评价语义社区结构.

5.3 重叠社区发现算法比较分析

本节实验的目的在于分析经典社区发现算法(非语义社区发现算法)在面向语义社会网络时划分结果存在

的偏差,为此,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,其中,LFM,COPRA,UEOC,EAGLE,CPM 为重叠社区发现算法.由于 QLSP 数据集仅包含 1 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 14 为以上各算法的社区划分结果,其中,黑色节点为重叠节点,各算法的 SQ1,SQ2 和 EQ 值见表 2.

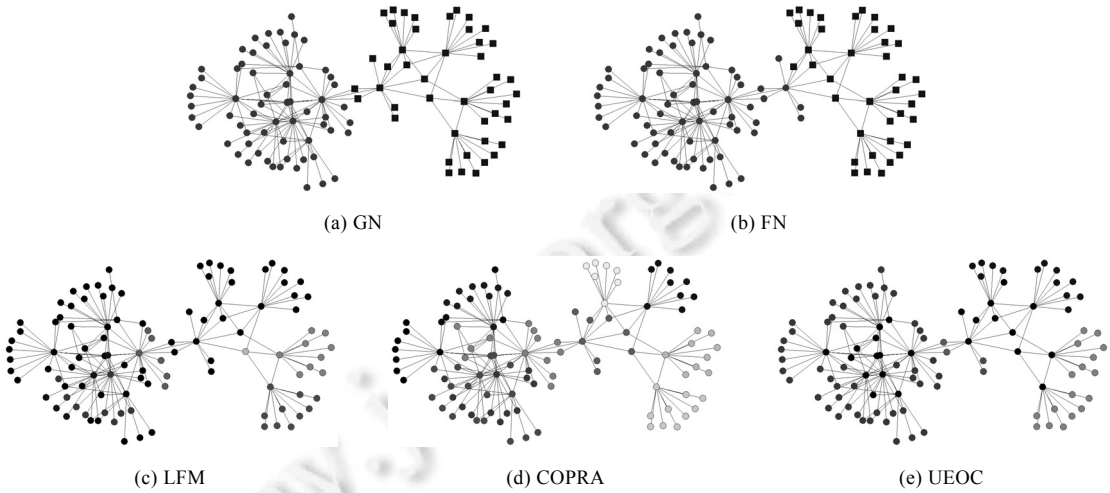


Fig.14 Detected communities by various algorithms

图 14 各算法发现的社区

Table 2 EQ and SQ values of classical algorithms

表 2 经典算法的 EQ 和 SQ 值

Algorithms	EQ	SQ1	SQ2
GN	0.461 7	0.358 4	0.573 4
FN	0.406 1	0.315 7	0.464 9
LFM	0.325 4	0.232 9	0.431 4
COPRA	0.541 0	0.400 3	0.569 8
UEOC	0.441 0	0.387 1	0.456 4
LBTC	0.524 6	0.326 6	0.647 6

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 值高于本文算法(0.524 6),但 SQ2 值均低于本文算法(0.647 6).由此验证了传统面向链接关系的社区划分算法(EQ 值较高)在处理语义社区划分问题时 SQ(SQ1,SQ2)值较低,所划分的社区结果与语义社区的理想结果偏差较大.

5.4 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)、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)、DBLP(April 12,2006)数据集(1 200 000 个节点)中分别截取:(A) 15 000 个节点数据集和(B) 20 000 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 LBTC 算法的运行参数为 T=5.

表 3 包括 EQ,SQ1,SQ2 及社区个数 CS,图 15~图 17 分别为各算法的 EQ,SQ1 和 SQ2 直方图,其中,图 15 的结果表示本文 LBTC 算法结果在 EQ 标准下的结果较差,而图 16 及图 17 的结果充分验证了本文 LBTC 算法的语义社区划分结果更精确.从图 15~图 17 的比对可知:相较于传统经典算法,本文的 LBTC 算法更适合处理语义社会网络的社区发现问题.

Table 3 Execution results of various datasets

表 3 各数据集的执行结果

Algorithms		QLSP	AFD	CND	DBLP (A)	DBLP (B)
GN	EQ	0.310 8	0.132 5	0.192 8	0.282 3	0.319 2
	SQ1	0.231 0	0.159 7	0.189 1	0.213 9	0.286 5
	SQ2	0.340 6	0.325 4	0.331 3	0.348 5	0.323 0
	CS	10	25	39	17	16
FN	EQ	0.421 6	0.152 5	0.223 5	0.319 1	0.261 8
	SQ1	0.331 4	0.139 2	0.172 1	0.291 6	0.256 1
	SQ2	0.478 6	0.324 1	0.340 1	0.389 2	0.334 3
	CS	10	27	37	19	16
LFM	EQ	0.366 8	0.147 3	0.240 6	0.405 2	0.364 1
	SQ1	0.317 2	0.132 1	0.217 2	0.331 7	0.313 3
	SQ2	0.385 0	0.296 8	0.317 6	0.423 8	0.441 9
	CS	12	24	33	22	12
COPRA	EQ	0.419 8	0.318 6	0.111 9	0.383 0	0.411 3
	SQ1	0.289 1	0.217 7	0.120 2	0.297 1	0.321 7
	SQ2	0.481 2	0.343 4	0.271 1	0.432 3	0.476 4
	CS	13	21	35	21	13
UEOC	EQ	0.384 9	0.2312	0.2648	0.3658	0.3183
	SQ1	0.317 7	0.2218	0.2271	0.2964	0.2011
	SQ2	0.492 2	0.3291	0.3	0.4132	0.4614
	CS	12	24	30	22	14
LBTC	EQ	0.324 8	0.241 5	0.201 5	0.354 2	0.300 1
	SQ1	0.351 2	0.261 3	0.273 4	0.364 2	0.367 9
	SQ2	0.519 2	0.388 5	0.387 1	0.484 5	0.489 3
	CS	14	25	34	23	15

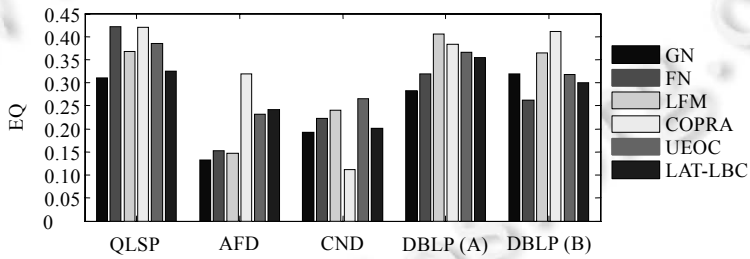


Fig.15 EQ histogram of various algorithms

图 15 各算法的 EQ 直方图

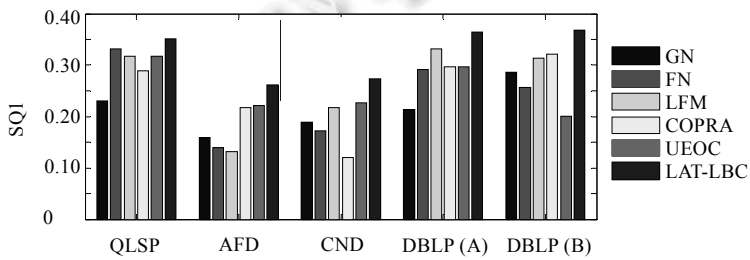


Fig.16 SQ1 histogram of various algorithms

图 16 各算法的 SQ1 直方图

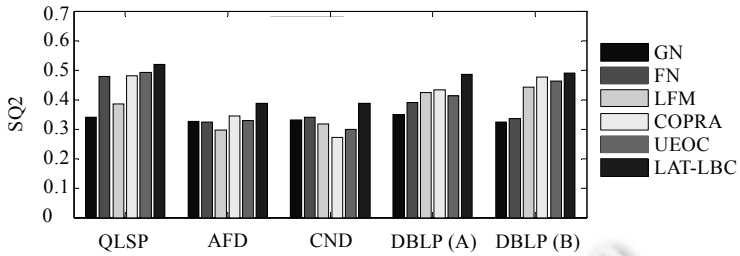


Fig.17 SQ2 histogram of various algorithms

图 17 各算法的 SQ2 直方图

5.5 语义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 0.5M 条数据,423M 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 为 Enron 数据集分别在 TURCM,CART,CUT,LCTA 算法下的 EQ 值及 SQ(SQ1,SQ2)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 数据集社区的最佳个数为 10.本文算法的社区个数为 11,EQ,SQ1 和 SQ2 取值分别为 0.332,0.318 和 0.393.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 4 Topic groups of Enron data set

表 4 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

Table 5 EQ, SQ1 and SQ2 values of various semantic community detection algorithms

表 5 各语义社区发现算法的 EQ,SQ1 及 SQ2 值

Algorithms		*CS*=6	*CS*=8	*CS*=10	*CS*=12	*CS*=14
TURCM	EQ	0.198	0.271	0.339	0.331	0.283
	SQ1	0.173	0.231	0.281	0.310	0.261
	SQ2	0.231	0.311	0;381	0.346	0.336
CART	EQ	0.152	0.249	0.302	0.294	0.255
	SQ1	0.122	0.226	0.256	0.268	0.226
	SQ2	0.284	0.293	0.387	0.365	0.334
CUT	EQ	0.133	0.231	0.266	0.278	0.227
	SQ1	0.126	0.215	0.233	0.235	0.202
	SQ2	0.247	0.281	0.299	0.352	0.363
LCTA	EQ	0.164	0.239	0.278	0.311	0.249
	SQ1	0.161	0.208	0.243	0.279	0.215
	SQ2	0.219	0.283	0.341	0.384	0.315

5.6 运行时间分析

为了分析 LBTC 的执行效率,本节实验对比了话题个数 T 与节点个数$|G|$对运行时间的影响.在数据集选择方面,本文利用人工数据集分别在网络拓扑关系及文本信息两方面进行模拟,其数据集生成过程如下:

(1) 在网络拓扑关系模拟方面,本实验利用 LFR benchmark[34]设计生成了 21 组实验数据,其参数为($|G|$={1500,3000,…,31500},ad={4,5,…,24},$d$max={30,32,70},$c$min={10,15,…,110},$c$max={100,110,…,300},on={100,200,…,2100},om={4,5,…,24},mi=2.5).其中,参数$|G|$表示节点的个数;ad 和 $d$max 分别表示网络中节点的平均度和最大度;$c$min 和 $c$max 分别表示最小社区和最大社区包含节点的数量;on 表

示重叠节点个数;*om* 表示每个重叠节点连接的社区个数;*mi* 为混合系数,表示节点与社区外部连接的概率.随着 *mi* 值的增大,网络社区结构越来越模糊;当 *mi*>0.5 时,网络的社区结构非常模糊.

(2) 在文本信息模拟方面,将各节点的节点编号作为节点的关键字,以 COPRA 算法[7]的方式对节点的关键字进行传播.为模拟文本信息传播的衰减过程,以公式(8)所示的拓扑势公式对节点的关键字进行加权.文本信息的模拟过程保持了节点间的文本信息的关联性与独立性,使得隶属同一社区的节点语义坐标相近似,符合语义社区结构的特性.通过文献[10]的分析可知,公式(8)中σ的最优取值为$\sigma\in(1,2)$.为此,本实验中σ=1.5.

$$weight_{i,j}=\exp\left(-\left(\frac{dis(G_i,G_j)}{\sigma}\right)^2\right) \tag{8}$$

本文所采用的实验环境为:Intel(R) Pentium(R)处理器,3.0GHz CPU,4.0GB 内存,160GB 硬盘,Microsoft Windows 7 操作系统,编程语言为 Matlab R2012b.由于算法的运行时间与实验环境的相关性较大,因此,本实验将|*G*|={1500,3000,…,31500},*T*={5,10,…,55}的运行时间与|*G*|=1500,*T*=5 的运行时间的比值作为分析对象,其对比分析结果如图 18 所示.其分析如下:

1) 节点个数对比为 *T*={10,20,30,40,50}时,运行时间随节点个数的变化对比.从对比可知:当 *T* 不变时,其运行时间随着节点个数的增加而增加;话题个数对比为|*G*|={6000,12000,18000,24000,30000}时,运行时间随话题个数的变化对比,其运行时间随着话题个数的增加而增加.结合节点个数对比与话题个数对比可知:当|*G*|>20000,*T*>30 时,运行时间的增加率较高,运行效率下降较快.

2) 话题时间δ为 *T* 增加 5 且|*G*|不变时,时间比值的增加量,节点时间δ为|*T*|增加 1500 且 *T* 不变时,时间比值的增加量.从图 18 中的话题时间δ及节点时间δ统计直方图可知:话题时间δ的分布集中在(400,1000)区间内,节点时间δ的分布集中在(0,600)区间内.由此可知,话题个数的变化对总体运行时间的影响更大.

3) 图 18 的 3 维话题个数-节点个数对比关系图直观表达了当|*G*|>20000,*T*>30 时,运行时间较高且变化强烈.话题个数维度的斜率高于节点个数维度,即,话题个数 *T* 对总体的运行时间影响更大.

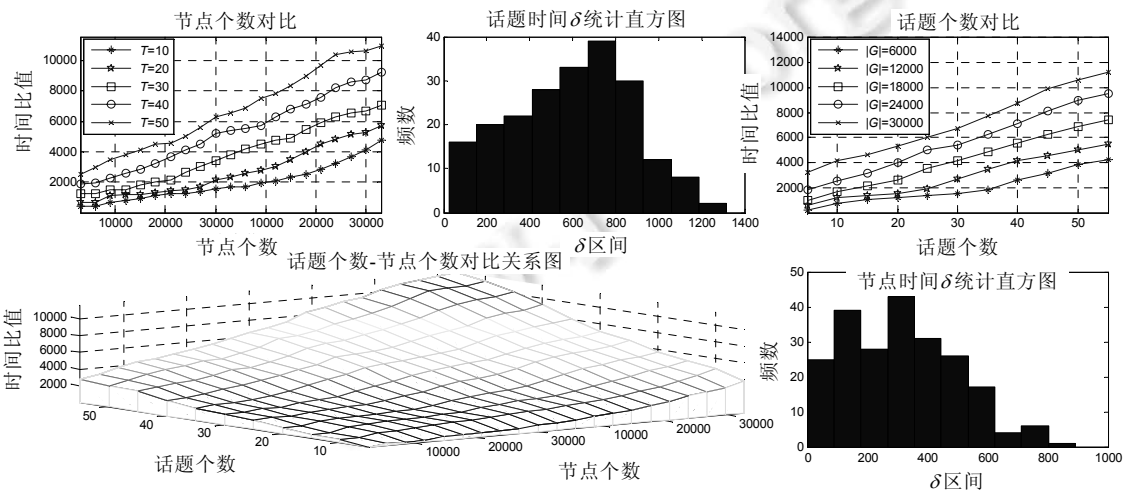


Fig.18 Comparison charts of running time

图 18 运行时间对比图

为对比各算法的运行效率,本文选取 21 组实验数据中的 5 组数据作为实验数据,分别对 TURCM,CART,CUT,LCTA 及 LBTC 算法在 C 语言平台下的时间运行进行了对比.

表 6 为各算法的运行时间对比结果(单位:h).从对比结果可知:由于 TURCM,CART,CUT,LCTA 及 LBTC 算法的建模方式相类似,其运行时间也基本相同.

Table 6 Comparison of running time on various semantic community detection algorithms

表 6 各语义社区发现算法的运行时间对比

Algorithms	$\|G\|$=1000, T=10	$\|G\|$=5000, T=20	$\|G\|$=10000, T=30	$\|G\|$=15000, T=40	$\|G\|$=20000, T=50
TURCM	0.12	0.63	1.58	3.62	4.19
CART	0.11	0.68	1.72	3.94	4.31
CUT	0.13	0.57	1.56	3.38	4.77
LCTA	0.11	0.53	1.63	3.41	4.62
LBTC	0.12	0.69	1.76	3.73	4.74

6 结束语

本文针对一般语义社会网络社区划分需要预先设定社区个数的问题,提出了 LBTC 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,可有效地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提出 LBT 模型,并建立了以 link 为核心的 block 区域取样方法;建立了可度量 link-block 区域的语义链接权重方法;提出以 link-block 为单位的重叠社区发现聚类算法(LBTC);提出了评价语义社区划分结果的 SQ1 及 SQ2 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,LBTC 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.另外,LDA 模型的缺点在于求解过程(Gibbs 取样过程)复杂度较高,对大数据的应对能力较弱.为此,下一步工作拟从并行取样角度出发,结合语义社会网络的结构特性,对 LDA 的求解过程进行优化.

References:

[1] Yang B, Liu DY, Jin D, MA HB. Complex network clustering algorithms. Ruan Jian Xue Bao/Journal of Software, 2009,20(1): 54−66 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[2] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of National Academy of Science, 2002, 9(12):7921−7826. [doi: 10.1073/pnas.122653799]

[3] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[4] Palla G, Derenyi I, Farkas I, Vicsde T. Uncovering the overlapping community structures of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[5] Shen HW, Cheng XQ, Cai K, Hu MB. Detect overlapping and hierarchical community structure in networks. Physica A, 2009,388 (8):1706−1712. [doi: 10.1016/j.physa.2008.12.021]

[6] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks. New Journal of Physics, 2009,11(3):033015. [doi: 10.1088/1367-2630/11/3/033015]

[7] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):103018. [doi: 10.1088/1367-2630/12/10/103018]

[8] Jin D, Yang B, Baquero C, Liu DY, He DX. A Markov random walk under constraint for discovering overlapping communities in complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2011,2011(5):75−98. [doi: 10.1088/1742-5468/2011/05/P05031]

[9] Jin D, Yang B, Liu J, Liu DY, He DX. Ant colony optimization based on random walk for community detection in complex networks. Ruan Jian Xue Bao/Journal of Software, 2012,23(3):451−464 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3996.htm [doi: 10.3724/SP.J.1001.2012.03996]

[10] Gan WY, He N, Li DY, Wang JM. Community discovery method in networks based on topological potential. Ruan Jian Xue Bao/Journal of Software. 2009,20(8):2241−2254 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

[11] Fu XH, Liu G, Guo YY, Wang ZQ. Multi-Aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems, 2013,37(1):186−195. [doi: 10.1016/j.knosys.2012.08.003]

[12] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3(1):993−1022.

[13] Zhang HZ, Qiu BJ, Giles CL, Foley HC, Yen J. An LDA-based community structure discovery approach for large-scale social networks. In: Proc. of the Intelligence and Security Informatics. Piscataway: IEEE, 2007. 200−207. [doi: 10.1109/ISI.2007.379553]

[14] Kemp C, Tenenbaum JB, Griffiths TL, Yamada T, Ueda U. Learning systems of concepts with an infinite relational model. In: Proc. of the Association for the Advancement of Artificial Intelligence. Palo Alto: AAAI, 2006. 5−13.

[15] Henderson K, Eliassi RT. Applying latent dirichlet allocation to group discovery in large graphs. In: Proc. of the 2009 ACM Symp. on Applied Computing. New York: ACM Press, 2009. 1456−1461. [doi: 10.1145/1529282.1529607]

[16] Henderson K, Eliassi RT, Papadimitriou S, Faloutsos C. HCDF: A Hybrid community discovery framework. In: Proc. of the 10th SIAM Int'l Conf. on Data Mining. Philadelphia: SDM, 2010. 754−765.

[17] Zhang HZ, Giles CL, Foley HC, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proc. of the Association for the Advancement of Artificial Intelligence. Palo Alto: AAAI, 2007. 663−668.

[18] Zhang HZ, Li W, Wang XR, Giles CL, Foley HC. HSN-PAM: Finding hierarchical probabilistic groups from large-scale networks. In: Proc. of the 7th Int'l Conf. on Data Mining Workshops. Piscataway: IEEE, 2007. 27−32. [doi: 10.1109/ICDMW.2007.115]

[19] Steyvers M, Smyth P, Rosen ZM, T Griffiths. Probabilistic author-topic models for information discovery. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 306−315. [doi: 10.1145/1014052.1014087]

[20] McCallum A, Corrada EA, Wang X. Topic and role discovery in social networks. Computer Science Department Faculty Publication Series, 2005,29(3):1−7.

[21] 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(4):249−272.

[22] Zhou D, Manavoglu E, Li J, Giles CL, Zha HY. Probabilistic models for discovering e-communities. In: Proc. of the 15th Int'l Conf. on World Wide Web. New York: ACM Press, 2006. 173−182. [doi: 10.1145/1135777.1135807]

[23] Cha Y, Cho J. Social-Network analysis using topic models. In: Proc. of the 35th ACM SIGIR Int'l Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2012. 565−574. [doi: 10.1145/2348283.2348360]

[24] Nagarajan N, Sen P, Chaoji V. Community detection in content-sharing social networks. In: Proc. of the 2013 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. Niagara Falls: IEEE, 2013. 82−89. [doi: 10.1145/2492517.2492546]

[25] Rios S, Munoz R. Dark Web portal overlapping community detection based on topic models. In: Proc. of the ACM SIGKDD Workshop on Intelligence and Security Informatics. New York: ACM Press, 2012. 1−7. [doi: 10.1145/2331791.2331793]

[26] Hu B, Song Z, Martin E. User features and social networks for topic modeling in online social media. In: Proc. of the 2012 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. Istanbul: IEEE, 2012. 202−209. [doi: 10.1109/ASONAM.2012.43]

[27] Wang XR, Mohanty N, McCallum A. Group and topic discovery from relations and text. In: Proc. of the 3rd Int'l Workshop on Link Discovery. New York: ACM Press, 2005. 28−35. [doi: 10.1145/1134271.1134276]

[28] Pathak N, Delong C, Banerjee A, Erickson K. Social topic models for community extraction. In: Proc. of the 2nd SNA-KDD Workshop. New York: ACM Press, 2008. 1−8.

[29] Mei QZ, Cai D, Zhang D, Zhai CX. Topic modeling with network regularization. In: Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 101−110. [doi: 10.1145/1367497.1367512]

[30] Sachan M, Contractor D, Faruquie T, Subramaniam V. Probabilistic model for discovering topic based communities in social networks. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2011. 2349−2352. [doi: 10. 1145/2063576.2063963]

[31] Sachan M, Contractor D, Faruquie TA, Subramaniam LV. Using content and interactions for discovering communities in social networks. In: Proc. of the 21st Int'l Conf. on World Wide Web. New York: ACM Press, 2012. 331−340. [doi: 10.1145/2187836.2187882]

[32] Yin ZJ, Cao LL, Gu QQ, Han JW. Latent community topic analysis: Integration of community discovery with topic modeling. ACM Trans. on Intelligent Systems and Technology, 2012,3(4):67−83. [doi: 10.1145/2337542.2337548]

[33] Zhao ZY, Feng SZ, Wang Q, Huang JZ, Williams GJ, Fan JP. Topic oriented community detection through social objects and link analysis in social networks. Knowledge Based Systems, 2012,26:164−173. [doi: 10.1016/j.knosys.2011.07.017]

[34] Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E, 2009,80(1):016118. [doi: 10.1103/PhysRevE.80.016118]

附中文参考文献:

[1] 杨博,刘大有,金弟,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66. http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[9] 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. http://www.jos.org.cn/1000-9825/3996.htm [doi: 10.3724/SP.J.1001.2012.03996]

[10] 淦文燕,赫南,李德毅,王建民.一种基于拓扑势的网络社区发现方法.软件学报,2009,20(8):2241−2254. http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

辛宇(1987－),男,黑龙江哈尔滨人,博士,讲师,CCF 学生会员,主要研究领域为企业智能计算,SNA.

谢志强(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为企业智能计算.

杨静(1962－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为企业智能计算.