基于话题概率模型的语义社区发现方法研究

辛 宇 [1, 2] 谢志强 [1] 杨 静 [2]

摘 要 语义社会网络 (Semantic social network, SSN) 是一种由信息节点及社会关系构成的复杂网络, 也是语义信息时代社会网络技术研究的热点, 相较于传统社会网络更具实用价值. 其研究内容包含了社会网络的语义分析及社会关系分析, 因此, 语义社会网络的社区挖掘建模具有一定的复杂性. 在语义社会网络的社区挖掘研究方面, 本文分析了当前基于话题概率模型的语义社区发现方法, 并在综述其内容的同时总结了各方法的优缺点, 为后续研究提供了理论基础. 在语义社会网络社区挖掘结果的评判方面, 本文归纳了相关的评价模型, 并通过实验分析对比了各模型对拓扑相关性和语义相关性的倾向性.

关键词 语义社会网络, 话题概率模型, 社区挖掘, 社区结构

引用格式 辛宇, 谢志强, 杨静. 基于话题概率模型的语义社区发现方法研究. 自动化学报, 2015, **41**(10): 1693−1710

DOI 10.16383/j.aas.2015.c150136

Semantic Community Detection Research Based on Topic Probability Models

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

Abstract Semantic social network (SSN) is a complex network consisting of textual node information and social relationships, and has more valuable applications than the traditional social network which focuses on social network analysis (SNA) with semantic information. As it contains the semantic analysis and social relationship analysis, there is some complexity for the modeling in mining the SNA community. In the SNA community mining aspect, the advantages and disadvantages of each method based on topic probability models are summarized to provide a theoretical basis for the further research. In the evaluation aspect of SNA community mining, relevant evaluation models are summarized. The tendency of each evaluation model toward topological and semantic relevance is compared by experimental analysis.

Key words Semantic social network (SSN), topic probability model, community mining, community structure

Citation Xin Yu, Xie Zhi-Qiang, Yang Jing. Semantic community detection research based on topic probability models. *Acta Automatica Sinica*, 2015, **41**(10): 1693−1710

现实世界中人与人的交互活动是人类社会活动的基础. 在社会学研究中将社会关系和社会个体所构成的网络称为社会网络[1]. 随着网络技术及通讯技术的发展, 人与人交互活动的途径呈数字化趋势, 对社会网络的研究也从传统社会学研究转变为数据挖掘研究, 从社会行为及社会关系研究转变为网络数理统计及量化分析研究[2]. 社会网络的研究方法也从以链接关系为主体, 发展到了以语义链接关系为主体, 从而产生了语义社会网络 (Semantic social network, SSN) 的概念. 文献 [3] 利用 FOAF (Friend of a friend project) 模型, 对语义社会网络进行了如下定义: 语义社会网络是在传统社会关系基础上, 通过对网络中的 "知识" 进行表达、关联及推理, 建立社会网络的语义相关性模型, 从而实现社会网络的语义化. 随着早期语义社会网络的资源标注模型 (如 RDF (Resource description framework)[4]、RDFS (Resource description framework schema)[5]、OWL (Web ontology language)[6]、SPARQL (Simple protocol and RDF query language)[7]) 的出现, 对语义社会网络中社区发现的研究成为语义社会网络研究的重要方向. 由此使得社区发现研究从传统非语义社区发现过渡到了语义社区发现. 非语义社区即为传统关系社区, 其中的 "关系" 是所有交互活动的统一表达, 不对具体的关系内容进行区分; 语义社区是在传统关系社区 (非语义社区) 中加入了关系内容的约束, 更强调社区的语义相关性. 图 1 表达了传统社区发现方法与

收稿日期 2015-03-31 录用日期 2015-06-24
Manuscript received March 31, 2015; accepted June 24, 2015
国家自然科学基金 (61370083, 61370086), 国家教育部博士点基金 (20122304110012), 黑龙江省自然科学基金 (F201101), 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) 资助
Supported by National Natural Science Foundation of China (61370083, 61370086), National Research Foundation for the Doctoral Program of Higher Education of China (20122304110012), Natural Science Foundation of Heilongjiang Province (F201101), Educational Commission of Heilongjiang Province (12531105), Postdoctoral Science Foundation of Heilongjiang Province (LBH-Q13092)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080 2. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001
1. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080 2. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001

语义社区发现方法的区别.

图 1 中传统非语义社区发现与语义社区发现的结果存在一定偏差, 且以语义分析为主导的社区发现结果较传统社区发现结果, 具有更合理的内部一致性, 其有效性更高.

非语义社区发现研究即为仅以拓扑关系为研究对象的传统社区发现研究, 其研究内容为社会网络的拓扑关系, 并从中挖掘出内部关系紧密的社区结构. 非语义社区发现的研究过程经历了硬社区划分和重叠社区划分阶段. 其中硬社区划分即所划分的社区不包含相交的用户节点, 其代表算法有: GN (Girvan Newman)[1]、FN (Fast Newman)[8]、FUA (Fast unfolding algorithm)[9] 等. 由于实现社会生活中, 社会个体的生活领域并不单一化, 某一社会个体可同时具备多重社会身份, 从而引发了硬社区划分结果对现实模拟具有偏差性的讨论. 为此, Palla 等[10] 提出了重叠社区的概念, 即社会网络中存在某些用户节点隶属于不同社区. 至此, 社区发现研究进入了重叠社区发现的时代, 各类经典算法孕育而生, 如 EAGLE (Agglomerative hierarchical clustering based on maximal clique)[11]、LFM (Lancichinetti-Fortunato method)[12]、RAK (Raghavan Albert Kumara)[13] 等.

语义社区发现以拓扑关系和语义内容作为研究对象, 即在传统社区发现方法中加入了语义分析, 其语义关系包含了工作、生活、娱乐和兴趣等多方面内容, 因此为使社区发现结果更加合理化, 需要在社区挖掘过程中加入对社会个体的交流内容和行为言论的分析 (即语义分析). 目前, 语义社区发现方法包含本体 – 社区挖掘[14−15]、多元属性建模[16] 及话题概率关系建模等. 在基于话题概率模型的社区挖掘研究方面, 其研究目标在于借用文本分析模型实现对语义社会网络中社会个体的话题相似性 (即语义相似性) 挖掘, 并建立社会个体或社会团体的相似性模型. 其主要的研究工作在于以各类话题发现模型 (如 LDA (Latent dirichlet allocation)[17]、PLSA (Probabilistic latent semantic analysis)[18]、HPC (Hierarchical poisson convolution)[19] 等) 为基础, 融合社会网络的拓扑关系特性, 建立话题 – 社会个体关系模型, 将话题关系相近且拓扑相关性强的社会个体聚合为社区.

本文对以话题概率关系模型为基础的语义社区发现算法进行了归纳, 在论述各方法的实现过程及理论意义的同时, 对其存在的缺点及改进方法进行了总结, 并利用图示表达法直观地阐述了各算法的内在关联. 另外, 本文对传统社区发现算法的评价指标及度量方法进行了改进, 使其满足语义社区评判的要求, 并在实验中对比了各方法的有效性.

1 主要研究方法

语义社会网络是语义信息与网络拓扑结构相结合的一类复杂系统, 且文本分析与社会网络结构均具有统计学特性 (如文本具有潜在话题分布[17], 社会网络用户节点具有幂率分布[20]), 因此, 基于概率模型的语义社区建模方法是语义社区发现的主要方法. Ding[21] 对比了面向拓扑与话题分析的社区发现算法的区别, 在利用 AT (Author-topic) 模型[22] 对话题进行抽取和分析后, 阐明了利用话题分析发现的社区比利用拓扑分析发现的社区具有更紧密的内在关联性. 继 AT 模型后, 学术界对语义社区挖掘的研究大量利用了统计学模型, 通过在文本、潜在话题、链接和社区间建立贝叶斯关联模型, 估计潜在变量的取值. 目前面向社区发现的话题概率模型 (如 AT 模型、ART 模型[23−24] 及 CART 模型[25] 等) 大多以 LDA 的建模方法为基础, 本文根据各类算法的 LDA 建模特征, 将基于概率模型的语义社区挖掘方法分为以下 4 类: 节点 – 话题概率模型、D2D (Document to document) – 话题概率模型, 链接 –

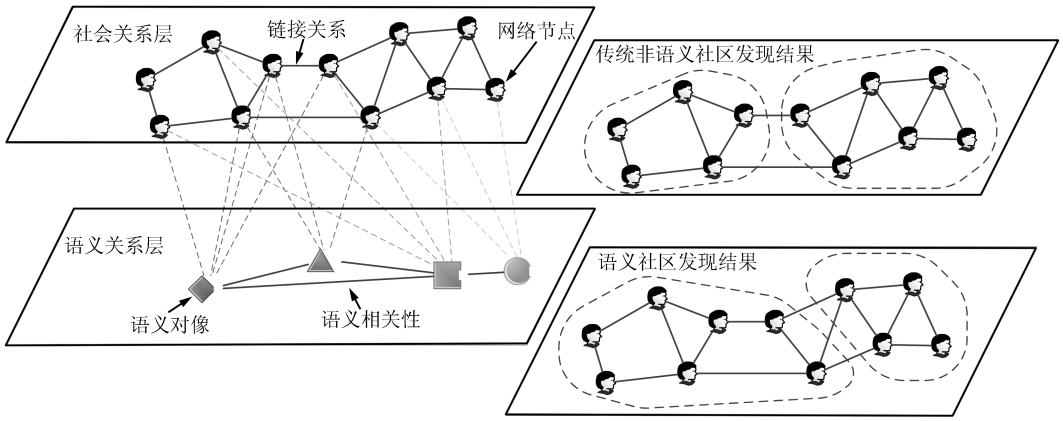


图 1 语义社区发现与传统非语义社区发现对比

Fig. 1 The comparison of samentic community detection with traditional non-semantic community detection

话题概率模型和社区–话题概率模型, 以下为 4 类模型的主要研究内容及区别.

1) 节点–话题概率模型, 此类模型以用户的言论、情感及兴趣趋向作为语义信息, 并将语义相似性高的一类用户作为基本的社区结构. 此类模型不考虑用户的社会关系层.

2) D2D 网络–话题概率模型, 此类模型以文本–文本网络作为社会网络, 将文本内容作为语义信息, 将文本内容相似的文本划分为同一社区. 此类模型是简化的链接–话题概率模型.

3) 链接–话题概率模型, 此类模型假设每个用户是一个文本集合, 即每个用户具有多重语义, 通过对用户–用户的链接关系与语义关系进行综合分析, 得到语义相似度高且结构紧密的社区.

4) 社区–话题概率模型, 此类模型假设每个用户的语义趋向受其所在社区的影响, 在进行语义分析时, 将用户的拓扑环境作为调控语义信息分布的因素. 此类模型所得到的社区结构的语义相似度最高, 但复杂度较高且结果的波动性较大.

为论述上述模型的区别及联系, 本文利用 “盘图” 对上述 4 类模型的贝叶斯模型框架进行表达, 其关系如图 2 所示. 图 2 中各符号表示如下:

G 表示全局网络, G_i 表示网络中的节点 i;

a_d 表示被取样点 G_d 的相邻节点集合;

$Community_d$ 表示被取样点 G_d 所隶属的社区集合;

x 表示集合 a_d 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;

w 表示关键字的集合, w_i 为集合 w 中第 i 个关键字所对应的编号;

z 表示与关键字的集合 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;

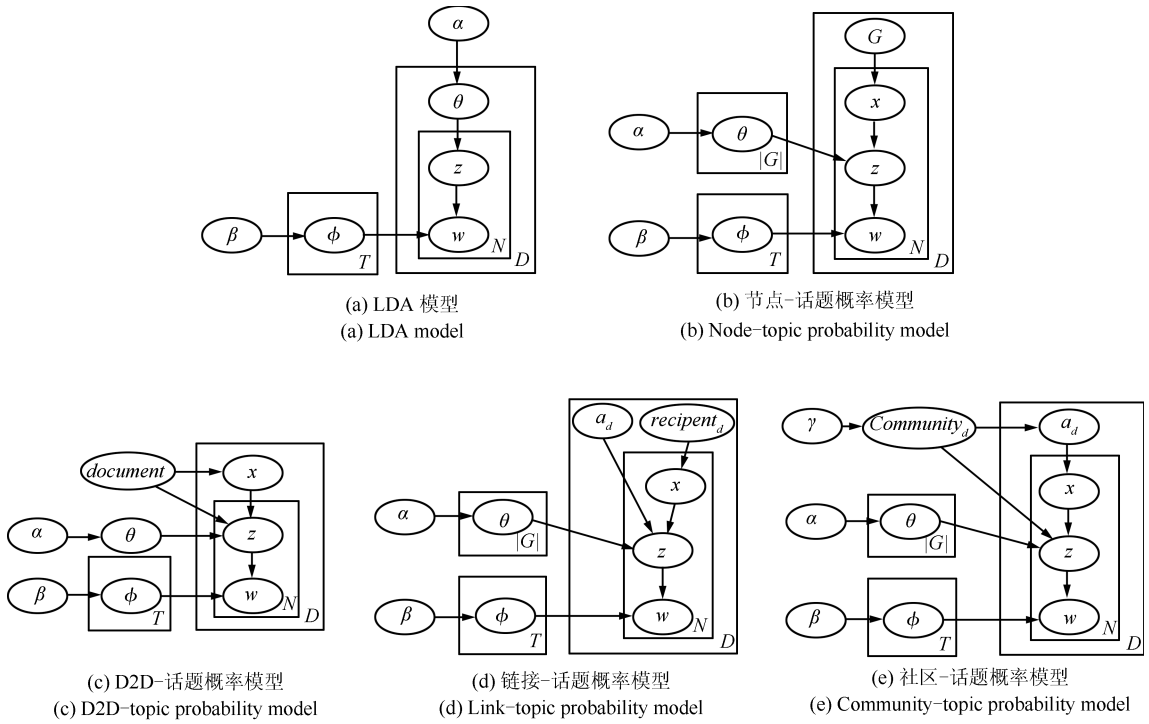
Φ 表示关键字的分布;

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;

γ 表示社区分布的先验参数.

从图 2 的对比可知: 节点–话题概率模型即在 LDA 模型的基础上, 考虑了节点的分布状态; D2D–话题概率模型考虑了节点的关联性; 链接–话题概率模型考虑了节点的关联性及节点的语义多重性; 社区–话题概率模型考虑了节点的社区属性. 利用话题概率模型进行语义社区挖掘的相关研究, 大多以上述 4 种模型为基础, 通过建立新的求解思路及关联规则, 不断充实、完善及扩展各模型的应用需求. 本文分别对上述 4 种模型的相关研究进行了总结, 并归纳了各算法的优缺点及可改进之处.



(a) LDA 模型
(a) LDA model

(b) 节点–话题概率模型
(b) Node–topic probability model

(c) D2D–话题概率模型
(c) D2D–topic probability model

(d) 链接–话题概率模型
(d) Link–topic probability model

(e) 社区–话题概率模型
(e) Community–topic probability model

图 2　各类话题概率模型的贝叶斯模型框架

Fig. 2　The framework of Bayes models for various topic probability models

1.1 节点 – 话题概率模型

用户节点是语义社会网络中话题的发起者, 由于用户节点的角色具有多面性, 因此用户节点可看作话题的集合, 且其包含的话题不唯一. 语义社会网络中用户节点的这一特性符合了话题概率模型的基本假设, 对此, 2004 年, Steyvers 等[22] 提出了 AT 模型. 该模型首次将 LDA 模型引入社会网络中, 实现了用户节点的潜在话题提取, 并开创了对用户节点的话题分布进行 LDA 建模的方法. 由于 AT 模型仅为 LDA 模型的套用, 缺少对社会网络链接关系及社区结构的考虑, 为此许多研究者借由 AT 模型的原理和模型基础, 提出了各类节点 – 概率模型. 图 3 为面向社区发现的节点 – 话题概率建模方法的主要模型, 本文根据各模型的特点对其进行了分类.

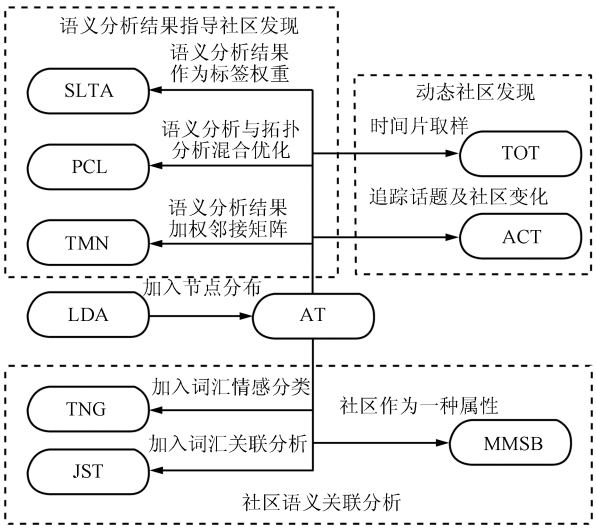


图 3 各类节点 – 话题概率模型关系图

Fig. 3 The relationship of various node-topic probability models

2006 年, Airoldi 等[26] 提出了 MMSB (Mixed membership stochastic block) 模型. 该模型首先假设每个用户节点可隶属多个社区, 藉此将社区作为 LDA 中的 “topic”, 将用户节点作为 LDA 中的 “document”; 其次, 预先设定若干社区作为初始社区, 具有多个社区属性的用户节点作为重叠节点; 最后, 在对 LDA 模型进行求解时, 利用网络的拓扑关系对同一社区中的用户节点进行取样, 从而优化所划分的社区结构. 其优点在于: 以发现潜在话题的形式发现潜在社区, 符合了 Barabási 等[20] 对社会网络中用户节点和社区之间满足一定分布关系的假设. 其缺点在于: 初始社区结构的设定对 MMSB 的结果影响很大, 因此, 初始社区结构的设定优化问题是 MMSB 需要面对的挑战.

2006 年, Wang 等[27] 假设 “随着时间累加, LDA 模型中话题先验参数的变化服从 Beta 分布”, 从而在 LDA 模型中加入了话题先验参数层, 建立了 TOT (Topics over time) 模型, 以实现动态话题变化的模拟. 在进行求解时, 对一定时间段内的语义信息进行 Gibbs 取样[28], 并将求得的参数结果作为下一次取样的初始值. 其优点在于: 对话题的建模结果较好, 其建模及求解过程满足 Gibbs 方法的求解特征. 其缺点在于: 忽略了网络的关系特性, 即话题的动态改变没有考虑人与人的交流过程及话题的扩散特性.

2007 年, Wang 等[29] 在二元 LDA 模型 BTM (Bigram topic model)[30] 的基础上加入了话题与关键字的关联关系, 提出了 TNG (Topical N-gram) 模型. 该模型构建了关键字的情感倾向性与关键字间的关联, 并直接影响了话题的分类结果, 可直接应用于情感分类中. 其优点在于: 由于 TNG 包含了 BTM 的关键字关联假设, 使得 TNG 在面向语义社会网络研究时, 可依据用户节点的链接关系追踪关键字的扩散对社区的影响. 其缺点在于: TNG 的模型结构过于复杂, 其中隐含变量过多, 导致 Gibbs 取样的结果不精确.

2008 年, Mei 等[31] 利用谐波正则化的建模方法, 将 LDA 或 PLSA 的 Likelihood 值与社区划分结果的谱分解相混合, 提出了 TMN (Topic modeling with network) 模型也称 NetPLSA. TMN 以 LDA 或 PLSA 的分析结果作为网络中链接的权重, 建立了语义社会网络的有权邻接矩阵, 并将该有权邻接矩阵作为谱分解的输入矩阵. 其优点在于: 充分利用了谱分解及谱聚类在社区发现中的有效性[32−33]. TMN 构建目标函数时采用了谐波正则化函数, 对考虑语义与结构双向优化的社区发现问题具有一定的启示作用. 其缺点在于: TMN 直接套用了 LDA 或 PLSA 的分析结果, 因而在进行语义建模时缺少对网络拓扑结构的考虑.

2009 年, Lin 等[34] 在 LDA 模型基础上, 为每个话题增加了一个情感取向先验参数, 提出了 JST (Joint sentiment topic) 模型. JST 将 LDA 的话题发现结果附加了情感特征, 从而实现了情感主题的挖掘. 其优点在于: 利用文本的情感属性细化了话题的分类结果, 增加了各话题的差异性, 有利于精确定位用户节点的社会属性. 其缺点在于: JST 需要预先量化各关键字的情感值, 由于大量关键字无法指定确切的情感指标, 因此 JST 的实现过程依赖于人为经验.

2009 年, Yang 等[35] 提出了 PCL (Popularity-based conditional link model) 模型. 该模型将语义社区建模分为语义建模及拓扑关系建模. 其中在语义建模方面采用了 DiscLDA 话题提取模型[36], 在

拓扑关系建模方面利用了社会网络的用户节点与链接的分布关系, 建立了 Conditional link model, 在进行社区划分时以 DiscLDA 和 Conditional link model 的混合函数作为目标函数进行 EM (Expectation maximization) 迭代求值. 其优点在于: 其目标函数的混合形式可直接套用各类语义分析模型及拓扑分析模型, 具有较好的可扩展性. 其缺点在于: EM 算法的复杂度较高, 对此, 在进行实际应用时需要考虑换用 Gibbs 取样算法.

2010 年, Li 等[37] 在利用 TTR-LDA-Community 模型[38] 发现社区基础上, 利用 ACT 分析模型[39−40] 分析了以时间片为单位的话题变化, 并得出了以下结论: 1) 同一社区内话题的变化具有趋同性; 2) 话题在变化时, 其自身的子话题会扩散到与其相邻的用户节点或社区中. 文章在进行语义社区评价时改进了一般社会网络的 NCP 评价指标[41], 为其加入了语义相似度因子, 使之可以度量语义社区结构. 其优点在于: 所得出的结论可作为语义社会网络话题传播及动态演变的实现基础, 其实验方法的说明及对 ACT 模型的运用十分恰当, 是对 ACT 模型的扩展研究. 其缺点在于: 对话题的动态跟踪过程中, 隐含假设了社区结构无变化, 因此, 其本质上是对静态社区结构的语义分析.

2012 年, Ríos 等[42] 提出了 SLTA (Speaker-listener topic propagation algorithm) 算法. SLTA 在 LDA 分析后, 对每一个用户节点选择一个权重最高的话题, 以此作为用户节点的标签, 再套用非语义社区发现的 SLPA 算法[43−44] 实现语义社区划分. 其优点在于: 将语义分析后的标签作为话题传播中的标签, 强化了标签的语义相关性, 为 SLPA 的模糊标签研究提供了方法. 其缺点在于: 语义社会网络中, 话题间具有一定的相关性, 且每个用户需要用多个标签描述, 只选择一个话题进行传播存在语义相关性偏差的弊端.

2013 年, Jang 等[45] 以舆论领袖的影响力及影响范围决定了社区的大小及紧密性为指导思想, 首先利用 LDA 模型将语义社会网络中的文本信息进行统一挖掘, 并将其挖掘结果作为社会个体的得分. 其次通过对得分及网络相关性的综合评价, 确定舆论领袖及其引领的话题社区. 其优点在于: 以 “舆论领袖” 作为社区核心的假设符合社区的结构特性. 其缺点在于: 进行全局文本信息的 LDA 挖掘时, 没有考虑局部的文本差异性及网络拓扑关联性, 使得零散的社会个体享有与核心个体同样的 “话语权”, 容易产生对全局文本挖掘结果的误导.

1.2 D2D-话题概率模型

D2D 网络由早期引文网络发展而来, 其研究对象是以 Document 作为用户节点的关系网络. 随着文本挖掘技术的发展, 其研究内容已从传统的拓扑关系挖掘发展到文本语义–拓扑关系综合挖掘. 2004 年, Erosheva 等[46] 将 D2D 网络看作一类特殊的语料库, 并将 LDA 模型直接应用于 D2D 网络, 从中挖掘相关话题. Erosheva 等的研究开创了利用概率模型挖掘 D2D 网络的先河. 此后, 各类以 D2D 网络为研究对象的算法层出不穷, 其研究方向逐步从话题挖掘过渡到了社区挖掘. 图 4 为当前面向社区发现的 D2D – 话题概率建模方面的主要模型, 本文根据各模型的特点对其进行了分类.

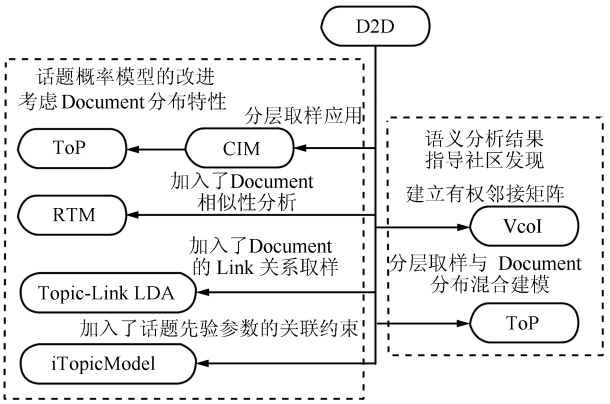


图 4　各类 D2D – 话题概率模型关系图

Fig. 4　The relationship of various D2D-topic probability models

2007 年, Dietz 等[47] 提出了 CIM (Citation influence model) 模型. 该模型以 LDA 模型为基础对引文网络 (有向 D2D 网络) 的有向链接进行文本分析, 因此 CIM 属于层次 LDA 模型, 其取样过程以有向链接作为基本单元. CIM 在引用与被引用的 Document 之间创新建立了取样平衡参数, 藉此修正了层次 LDA 存在的引用 Document 对被引用 Document 的完全包含假设. 其优点在于: 对有向链接进行取样的方式有利于 CIM 面向连接预测应用, 所建立的平衡参数为利用 LDA 模型处理有向语义社会网络社区挖掘问题提供了参考. 其缺点在于: 平衡参数需要人为预先设定.

2009 年, Chang 等[48] 在 LDA 模型的基础上建立了 D2D 网络的话题发现模型 RTM (Relational topic model). RTM 将两个有关联的 Document 的话题分布进行关联分析, 从而将 Document 的相似性引入 LDA 模型. 其优点在于: RTM 在改进 LDA 的同时既保持了 LDA 文本分析的有效性, 又在文本分析的同时加入了拓扑关联性, 为后续的语义社区挖掘研究开拓了新的研究方法, 许多语义社区研究方法藉此而生. 其缺点在于: RTM 仅以直接链接

关系作为网络的关联, 忽略了网络的区域性及规模性.

2009 年, Sun 等[49] 提出了用于 D2D 网络社区发现的 iTopicModel 模型. 该模型利用 MRF (Markov random field) 方法[50] 将 LDA 模型中的话题先验参数进行关联, 改进了 MRF 方法无法处理有权网络的问题, 并提出了可用于度量语义社区发现结果的 NMI 模型. 其优点在于: 在 MRF 中加入了社区因素, 使其具有社区发现功能. 其缺点在于: 其建模方式是在原有的三层 LDA 模型中增加了话题先验参数分布层, 且话题的先验参数分布不够明确, 使得其实现过程复杂化, 精确度较差.

2009 年, Liu 等[51] 提出了 Topic-Link LDA 模型. 该模型为简化 ART 模型的取样频数, 将 Author 所共享的所有 Document 综合为一个 Document, 从而将语义社会网络转化为 D2D (Document-to-document) 网络, 并在假设的社区内对 Author 进行取样以优化社区结构. 其优点在于: 采用简化的 D2D 形式, 适合处理大规模语义社会网络数据. 其缺点在于: 需要预先设定初始社区结构及社区个数.

2009 年, L′Huillier 等[52] 首先在 D2D 网络中建立了 Topic-document 关联矩阵 SWM (Semantic weights matrix), 以此作为 LDA 模型的输入; 其次将 LDA 模型所得出的话题分布参数作为阈值, 据此为话题相似度接近的用户节点建立一个虚拟链接, 从而形成虚拟网络 VcoI (Virtual communities of interests). 重新整理后的有权网络结构可作为输入, 套用非语义社区发现算法即可实现语义社区划分. 其优点在于: 采用了话题相似度进行虚拟链接补全的形式, 适用于语义结构洞分析及动态演化分析. 其缺点在于: 对虚拟链接的考虑仅以话题相似度为出发点, 缺少对网络结构的考虑.

2011 年, Zheng 等[53] 指出: Net-PLSA 等模型所采用的 D2D 网络, 将用户节点的所有 Document 拼凑为一个 Document 而忽略了 Document 的分布特性. 针对这一问题, 提出了 ToP (Topics on participations) 算法. ToP 在 CIM[47] 基础上, 将社区发现模型分成了 Hierarchical Dirichlet process 和 Document modeling 阶段, 在进行社区发现的同时考虑了 Document 的分布性. 其优点在于: 实现过程中, 通过对社区个数的增量训练避免了 LDA 模型需要预先设定社区个数的问题. 其缺点在于: 增量式训练过程的运算代价较大.

2013 年, Chahal 等[54] 利用词汇的相似性匹配关系建立了 Documents 相似性度量方法, 即将 Documents 看作词汇的集合, 通过词汇相似性匹配关系的累加关系得出 Documents 的相似性匹配关系; 其次, 以 Document 相似性匹配关系为输入, 通过对 Documents 的邻接关系聚类实现了 D2D 网络的社区划分. 其优点在于: 根据 Documents 相似性关系, 在 LDA 模型中加入 Documents 相关性参数, 提高了 LDA 挖掘模型的分类精度. 其缺点在于: Documents 的相似性分类过程中需要人工分组, 由于人为经验而产生的偏差直接影响了分类结果.

1.3 链接 – 话题概率模型

1999 年, Kleinberg[55] 提出网络通讯以话题作为信息单元以链接作为途径. 其理论依据在于: 由于用户节点具有多重社会角色, 使用户节点所包含的话题复杂化, 而用户节点大多使用同一种身份与另一用户节点进行交流, 因此, 同一链接所包含的话题内容相似. Kleinberg 所得出的结论是面向语义链接建模的理论基础, 后续的研究者将用户节点的语义信息看作其链接所包含的话题信息的集合, 从而开拓了以链接为研究对象的语义社区挖掘方法. 本节综述了此类方法与概率模型相结合的相关研究. 图 5 为当前面向社区发现的链接 – 话题概率建模方面的主要模型, 本文根据各模型的特点对其进行了分类.

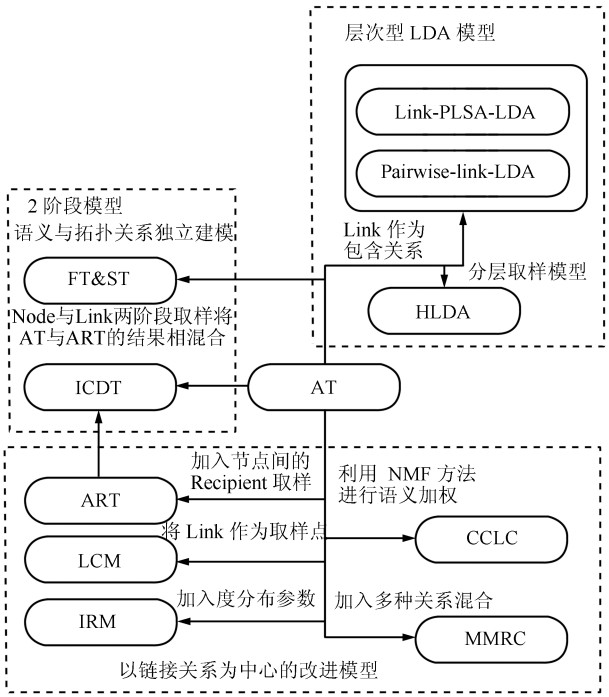


图 5 各类链接 – 话题概率模型关系图

Fig. 5 The relationship of various link-topic probability models

2005 年, McCallum 等[23−24] 提出了 ART (Author recipient topic) 模型. 该模型在 AT 模型的基础上加入了邻居节点的取样过程. 其实现过程为: 对

某一用户节点进行文本取样时, 随机选择该取样对象的一个邻居, 将该用户节点与其邻居节点的取样结果作为该用户节点的取样结果. ART 模型的取样过程考虑了用户节点对其周围节点的语义影响. 其优点在于: ART 模型将以用户为中心的局部小区域的语义信息作为用户节点的语义信息, 使得取样后的话题分布融合了网络的拓扑特性. 另外, 由于每次取样都随机选择一个邻居作为取样节点的扩展, 保障了取样节点的语义支配地位. 其缺点在于: 利用了均匀分布的形式对邻居进行选择, 即假设所有邻居对取样点的影响力均相同, 没有考虑邻居对取样用户节点的拓扑影响力.

2006 年, Kemp 等[56] 考虑了网络度分布对节点关系的影响, 提出了以链接存在概率为研究对象的 IRM (Infinite relational models) 模型, IRM 将已存在的链接看作概率分布的一个样本, 并假设链接服从用户节点对度的 Beta 分布, 计算任意用户节点间的链接概率. 在进行社区发现时, IRM 以链接概率矩阵为输入, 建立了社区内部链接概率最优化模型, 通过最优划分社区内部的链接概率实现社区发现. 其优点在于: 以链接概率矩阵的形式泛化了邻接矩阵, 从而使网络的拓扑结构可以应用于各类概率关系模型. 其缺点在于: 其链接概率的计算过程缺少了社会网络度的幂率分布假设, 因此其链接概率矩阵不能准确表达社会网络的特征.

2007 年, Long 等[57] 为实现社会网络中的多种关系混合建模, 提出了 MMRC (Mix membership relational clustering) 算法. 其在 AT 模型中加入了节点–链接的混合层, 以链接的种类 (如一般朋友关系、挚友关系、上下级关系等) 作为节点分布的权重, 强调链接在网络分布中的主导作用. 另外, 算法在进行局部拓扑关系挖掘方面采用了结构简单的 Sort relational clustering 方法以降低计算代价, 在全局范围的聚类划分方面采用了精确性高的 Hard relational clustering 方法. MMRC 属于两阶段聚类方法, 其优点在于: 合理地分配了 2 阶段的计算代价问题, 适用于处理大规模数据及云计算数据. 其缺点在于: 各聚类簇的内部紧致性不高, 将链接关系作为语义同样无法解决社区个数需要预先设定的问题.

2007 年, Zhu 等[58] 利用 NMF[59−60] 的非语义社区发现形式进行语义社区划分, 提出了 CCLC (Combining content and link for classification) 算法, 其实现过程为: 首先, 对具有语义关联的链接进行语义加权, 由此将语义社会网络转化为有权社会网络; 其次, 将网络加权邻接矩阵按 NMF 的形式构建以加权邻接矩阵近似为目标的社区–特征分解表达式; 最后, 以社区–特征分解表达式为目标函数, 建立参数估计的最优化方法, 以此计算社区划分结果. 其优点在于: 设计了 NMF 与语义社区发现相结合的平滑参数, 且其目标函数的构建方法值得借鉴. 其缺点在于: 继承了 NMF 的高计算复杂度, 不利于算法的推广.

2008 年, Nallapati 等先后提出了 Link-PLSA-LDA[61] 和 Pairwise-link-LDA[62] 取样模型对具有引用关系的语义网络进行建模, 以提取其中的潜在话题. 其假设被引用者的话题完全包含于引用者的话题结构内, 从而在 LDA 模型中加入了话题–话题关系. 其优点在于: 对引文类型网络的挖掘效果较好, 通过加入了引用者与被引用者的社会关系挖掘方法 MMSB[26] 实现了引文网络的社区挖掘. 其缺点在于: 语义社会网络的话题关系大多不具备完全包含关系, 限制了 Link-PLSA-LDA 和 Pairwise-link-LDA 的应用范围.

2012 年, Yin 等[63] 提出了 ICDT (Integration of community discovery with topic modeling) 模型. 该模型将取样过程分为用户节点取样和链接取样. 其过程为在对某一用户节点进行取样后, 对该用户节点的所有连接进行一次取样, 将这 2 阶段的取样结果作为该用户节点的取样结果, 即将 AT 与 ART 的结果相混合. 另外, ICDT 为每个用户节点及链接均分配了社区隶属度属性, 在取样过程结束后可根据社区隶属度对用户节点分配社区. 其优点在于: 2 阶段的取样过程形成了以用户节点为核心的取样区域, 该区域可模拟用户节点对周边链接的语义影响力. 其缺点在于: ICDT 在进行社区隶属度假设时, 没有考虑社会网络的链接关系, 导致所划分的社区出现不连通的情况.

2012 年, Cha 等[64] 根据社交网络中跟帖人的话题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, 提出了面向语义社会网络分析的 HLDA 模型. 该模型在进行文本生成取样时, 以取样节点为核心, 对与之距离为 2, $\cdots$, 3 的用户节点进行分层取样. 其优点在于: 可有效处理论坛类 (非熟人关系) 网站的用户节点分类问题, 此外, 其文本生成的取样方式可兼顾网络的拓扑特性. 其缺点在于: HLDA 的分层取样过程为无权取样过程, 没有考虑话题的传播受距离影响的因素.

2012 年, Hu 等[65] 以 “朋友关系越亲密其话题分布越接近” 为假设, 提出了面向语义分析的 FT (Feature topic) 模型和面向关系分析的 ST (Social topic) 模型, 其中 FT 与 ST 模型为相互独立的 LDA 模型. 在话题挖掘过程中, 利用 Gibbs 取样方法, 分别对 FT 及 ST 进行交叉取样, 从而保证了 LDA 模型参数求解时满足 FT 和 ST 的贝叶斯关系条件. 其优点在于: 通过建立两个 LDA 模型 (FT 和

ST) 的方式避免了混合模型的复杂化, 同时, 其求解方式又保持了 FT 及 ST 的约束条件. 其缺点在于: 以 FT 及 ST 的形式使用户节点的语义信息和拓扑信息相独立, 缺少了语义 – 拓扑相关性 (即语义对拓扑结构的影响力) 的考虑.

2013 年, Natarajan 等[66] 以 Link community[67] 为切入点, 建立了以 Link-content 为语义分析目标的 LCM (Link-content model) 模型. 该模型以用户节点间的共享信息及用户节点间的传递信息作为 Link-content, 在进行 LDA 建模时将 Link 作为社区分类的基本单元. 其优点在于: 构建了 Link-content 的 LDA 模型, 在发现 Link community 的同时可处理语义社会网络的链接预测问题. 其缺点在于: Link-content 属于用户节点话题概率模型的简单套用, 没能结合 Link-content 的特征进行改进, 因此 LCM 同样具有用户节点 – 话题概率模型的一般缺点.

1.4 社区 – 话题概率模型

局部区域性是社会网络有别于一般图状网络的特有拓扑特征, 传统社会网络研究中对社会网络中的局部区域性结构 (如 Clique 社区[10]、Block 社区[68−69]) 的存在性已做了充分证明. 近些年在基于概率关系模型的语义社区挖掘过程中, 将社会网络的局部区域性与概率生成关系相混合, 提出了一系列考虑局部区域性因素的语义社区挖掘方法. 图 6 为当前面向社区发现的社区 – 话题概率建模方面的主要模型, 本文根据各模型的特点对其进行了分类.

2005 年, Wang 等[70] 为了使 LDA 具有社区分类的能力, 提出了 GT (Group topic) 模型. 该模型在 LDA 模型中加入了社区元素, 并将关键字分为社区内部关键字与全局关键字. 其中全局关键字的分布仅依赖于话题分布, 社区内部关键字的分布依赖于话题与社区的混合分布. GT 模型的参数估计过程为: 先以全局关键字估计话题分布, 再以社区内部关键字及话题分布估计社区分布, 如此反复直到参数收敛. 其优点在于: GT 所建立的话题与社区独立的两阶段混合取样方式, 提高了 LDA 面向多元参数估计时的精度. 其缺点在于: GT 没有考虑用户节点间的链接关系, 应对社会网络的社区结构挖掘时效果较差.

2006 年, Zhou 等[71] 为了解决社区的不连通现象, 提出了 CUT (Community user topic) 模型. 其在 AT 模型的基础上为用户节点加入了社区分布条件. 其贝叶斯依赖关系为: 先以社区生成用户节点, 再以用户节点生成语义话题. 通过对中间层 (用户节点层) 进行估计可挖掘出以语义作为区分的用户节点簇, 从而实现社区划分. 其优点在于: CUT 模型弥补了 AT 模型不考虑用户节点间关联性的不足, 为后续的语义社区挖掘研究作了铺垫. 其缺点在于: 用户节点的社区分布没有考虑用户的链接关系, 仅在处理不强调链接关系的语义社区挖掘时有效, 在处理以网络拓扑为基础的社区挖掘问题时, 会出现社区内部不连通的情况.

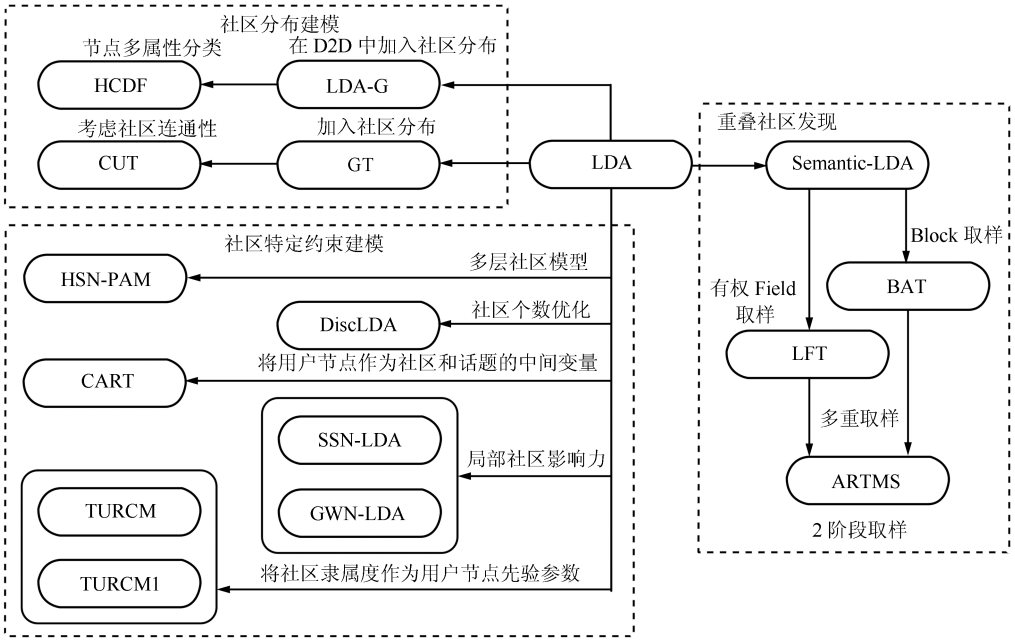


图 6 各类社区 – 话题概率模型关系图

Fig. 6 The relationship of various community-topic probability models

2007 年, Zhang 等[72] 提出了 SSN-LDA (Simple social network-LDA) 模型. 该模型建立了用户节点对所有相邻用户节点的 SIW (Social interaction weight) 分布函数, 以此模拟相邻用户节点间的影响力. 随后 Zhang 等[73] 在 SSN-LDA 的基础上考虑了权重因素, 提出了 GWN-LDA (Generic weighted network) 模型. SSN-LDA 与 GWN-LDA 在进行 LDA 模型求解时, 按 SIW 对取样元素进行加权, 从而划分出考虑用户节点相关性的社区结构. 其优点在于: SIW 的加权取样过程可作为考虑语义相关性的基础取样模型, 其建模过程简单, 具有较强的可扩展性. 其缺点在于: 语义社会网络中节点的语义信息 (如话题及标签关键字) 间同样存在着关联性, 而 SSN-LDA 与 GWN-LDA 仅面向有权网络缺少对语义信息相关性的考虑. 随后, Zhang 等[74] 又提出了 HSN-PAM (Hierarchical social network pachinko allocation) 模型, HSN-PAM 利用 LDA 模型的层次性建立了多层社区模型, 即在 LDA 的社区分布中加入了多层变量, 所求解出的各层社区间存在树状包含关系. 其优点在于: 将社区变量进行多层化, 不仅可以对 HSN-PAM 的求解过程进行分布式计算以应对大规模数据挖掘的需要, 而且树状结构使 HSN-PAM 模型具有网络拓扑整体性约束, 即下层社区内容的变更需要依赖上层的全局状态. 其缺点在于: 由于 HSN-PAM 模型的层次数较多, 导致对 LDA 进行多层求解时, 各层变量的准确性降低.

2008 年, Pathak 等[25] 在 ART 模型[23] 的基础上提出了 CART (Community author recipient topic) 模型. 该模型在 ART 模型中加入了社区元素, 将用户节点和话题作为社区的生成结果. 不同于 TURCM (Topic user recipient community model)[75] 将社区隶属度作为用户节点先验参数的建模方式, CART 将用户节点作为社区和话题的中间变量, 以挖掘潜在用户节点簇的方式实现社区发现. 其优点在于: 建立了社区与用户节点及社区与话题的双向关联, 使得所发现的语义社区结构中保持了同一社区话题相近的条件. 其缺点在于: CART 模型中预设社区个数对社区挖掘的结果影响较大, 对社区预设个数的最优化选取是 CART 模型急需解决的问题.

2009 年, Lacoste-Julien 等[36] 为解决 LDA 模型需要预先设定社区个数的问题, 提出了 DiscLDA (Discriminative variation on LDA) 模型. 该模型在 LDA 模型的话题分布的先验参数中加入了转移概率矩阵. DiscLDA 模型利用 k 维话题先验参数与转移概率矩阵的乘积形式对 k 维话题进行降维调整, 最终实现话题个数的迭代优化. 其优点在于: 利用转移概率矩阵的形式对话题进行优化迭代, 解决了 LDA 模型的分析结果依赖于话题的预设个数问题. 其缺点在于: DiscLDA 以 k 维话题的分布参数与转移概率矩阵之积作为话题的先验参数, 在进行 Gibbs 取样时无法求得准确的转移概率矩阵取值, 因此其分类结果较差.

2009 年, Henderson 等[76] 为了在 D2D 网络中挖掘潜在的 Document community, 提出了 LDA-G 模型. 该模型将网络中的所有用户节点作为 LDA 中的 Document, 将社区作为 Topic, 将用户节点间的链接作为 Word 关联. LDA-G 将网络中的拓扑信息作为文本信息, 从而将 LDA 模型中 Word-docuemt-topic 套用为 Node-link-community, 再利用 LDA 的求解方式实现社区发现. 其优点在于: LDA-G 发现了 LDA 模型与社区模型的相似之处, 从而保证了其理论可行性. 其缺点在于: 其本质上不属于面向文本话题分析的社区挖掘算法, 且 LDA-G 在直接套用 LDA 模型时, 对社区紧致性的考虑较少, 所挖掘的社区结构模块度不高.

2010 年, Henderson 等[77] 在 LDA-G 模型的基础上为用户节点加入了属性元素提出了 HCDF (Hybird community discovery framework) 模型. HCDF 通过对用户节点及属性的混合建模, 建立了考虑属性分类和用户节点分类双向优化的 LDA 模型, 其社区划分结果满足了社区内部语义的一致性. 其优点在于: 所建立的双目标混合概率模型, 满足了语义社区对拓扑关系和语义关系的双向要求. 其缺点在于: 双目标混合优化的 LDA 模型不同于单目标混合优化 LDA 模型 (如 CART[25]), 需要在求解过程中增加额外的参数估计过程, 从而降低了参数估计的准确性.

2010 年, Kumar 等[78] 建立了社区话题动态变化的度量方法, 其中包括零散用户节点、小社区和大规模社区话题变化度量. 对 3 种规模社区的话题跟踪及度量得出了以下结论: 不同规模及不同结构的社区, 其话题内容均具有一定的相关性, 即社区的划分无法隔断话题的连通性和一致性. 其研究结论为语义社区发现提供了一种新的思路: 对局部话题的分析是否需要以全局语义为基础, 以及如何解决局部语义对全局语义的依赖性建模问题.

2010 年, Kwak 等[79] 在网络节点的幂率分布假设[20] 的基础上, 分别在 Follower、Reciprocity、Degree of Separation、Homophily 4 个方面分析了用户节点之间的关联性, 并给出了用于评价用户节点权威性的 Rank 指标. Kwak 等通过实验对比得出以下结论: 1) 用户节点的行为及社会特征极大程度受与之相邻的用户节点影响, 即语义环境决定了用户节点的

属性; 2) 各用户节点的差异性分布在环境接受阈内, 其特征与环境呈同向变化的趋势. 文章从理论推导和实验验证 2 方面论证了语义相关性对社区挖掘的影响, 支持了社会网络语义层面的相关性, 是语义社会网络挖掘的重要依据.

2011 年, Sachan 等[75] 提出了 TURCM 模型. 该模型以一次文本交互中的链接关系作为文本取样的对象, 并在取样过程中加入了用户节点从属多个社区的假设. 通过对 TURCM 的取样分析, 可得到各个用户节点对各个社区的隶属关系, 从而将隶属度大的社区作为用户节点所在的社区. 其优点在于: 将用户节点对社区的隶属度作为用户节点分布的先验参数, 通过 Gibbs 取样法对 TURCM 进行求解, 可直接得出用户节点对社区的隶属关系. 其缺点在于: 没有考虑社区与话题间的关联, 其概率模型结构不稳定. 为了弥补 TURCM 的这一缺点, Sachan 等[80] 随后提出了 TURCM 的改进模型 TURCM1, 在 TURCM 的基础上增加了社区与话题的关联. 然而 TURCM 与 TURCM1 的共同缺点在于, 在将社区作为用户节点属性的同时缺少对社区连通性的考虑, 所挖掘出的社区结构会出现不连通的情况.

2014 年, 辛宇等在 Semantic-LDA[81] 基础上, 分别以节点的 Block 区域及有权 Field 区域为中心, 以 Block 区域及有权 Field 区域的语义分析作为该取样节点的语义信息, 提出了 BAT (Block author topic) 模型[82] 及 LFT (Link field topic) 模型[83], 并随后提出了可加速 Gibss 取样过程的多重取样 ARTMS (ARTs multiple sampling)[84] 方法. 此类算法采用语义分析与社区发现相分离的 2 阶段混合策略. 其优点在于: 可对第 1 阶段的语义社区划分结果套用传统社区发现算法, 易于直接应用于动态社区发现及重叠社区发现且无需预先设定社区个数; 其缺点在于: Semantic-LDA、BAT、LFT 及 ARTMS 在处理节点间语义相关性较低的问题时, 各节点的语义分析结果差异较大, 导致所划分的同一社区中存在与社区语义相关性低的噪音节点.

2 语义划分结果的度量方法分析

语义社区结构在传统社区结构中加入了语义信息, 因此, 语义社区划分的评价标准需要兼顾社区内部的语义相关性及社区结构的紧致性. 本文综述了语义社区评价中常用的方法及模型, 并设计了对比各模型评价性能的实验方法.

2.1 度量方法综述

本文对各类评价方法进行了统一描述, 所涉及的数学符号如下:

G 表示全局网络, $|G|$ 表示网络中用户节点的数量, G_i 表示网络 G 中的第 i 个用户节点;

L 表示网络 G 中的链接集合, $|L|$ 表示网络中的链接数量, $l(i,j)$ 表示连接用户节点 G_i 和 G_j 的链接;

A 表示网络 G 的邻接矩阵;

degree_i 表示用户节点 G_i 的度;

C 表示所划分的社区集合, $|C|$ 表示社区个数, $|C_i|$ 表示社区 C_i 内的用户节点个数;

$\boldsymbol{m}_i$ 表示用户节点 G_i 的属性 (话题) 向量, 其维数为 k.

各语义社区评价模型如下:

1) Qov 模型[42−85]: 该模型是模块度 Q[86] 的广义模型, 可用于评价有权有向网络, 适用于评价将语义相似性量化后的语义社区结构.

其表达式如式 (1) 所示, 其中 $f(x)=2px-p$, $\alpha_{i,c}$ 是节点 i 对社区 c 的贡献度, $\beta_{l(i,j),c}$ 是社区 c 的内部链接 $l(i,j)$ 对社区 c 的贡献度, $\beta^{\text{in}}_{l(i,j),c}$ 和 $\beta^{\text{out}}_{l(i,j),c}$ 分别为有向边缘链接对社区 c 的贡献度, Qov 取值越大社区结构越合理.

2) ACS (Average cluster similarity) 模型[87]: 该模型利用 RBF (Radial bias function) 核模型[88] 对语义社区结构进行评价, 其表达式为

$$ACS=\sum_{G_i\in c,G_j\in c'}\frac{K(\boldsymbol{m}_i,\boldsymbol{m}_j)}{k} \tag{2}$$

其中, $K(\boldsymbol{m}_i,\boldsymbol{m}_j)$ 表示 RBF 核模型, 其表达式为

$$\begin{aligned}
&Qov=\frac{1}{2\,|L|}\sum_{c\in C}\sum_{i,j\in G}\left[\beta_{l(i,j),c}A_{ij}-\frac{\beta^{\text{out}}_{l(i,j)}\text{degree}^{\text{out}}_i\beta^{\text{in}}_{l(i,j)}\text{degree}^{\text{in}}_j}{2\,|L|}\right] \quad \sum_{c=1}^{|C|}\alpha_{i,c}=1\\
&\beta^{\text{out}}_{l(i,j),c}=\frac{\sum\limits_{j\notin c,i\in c}F\left(\alpha_{i,c},\alpha_{j,c}\right)}{|G|},\quad \beta^{\text{in}}_{l(i,j),c}=\frac{\sum\limits_{j\notin c,i\in c}F\left(\alpha_{i,c},\alpha_{j,c}\right)}{|V|}\\
&\beta_{l(i,j),c}=\frac{\sum\limits_{i,j\in c}F\left(\alpha_{i,c},\alpha_{j,c}\right)}{|V|},\quad F\left(\alpha_{i,c},\alpha_{j,c}\right)=\frac{1}{\left(1+\mathrm{e}^{-f(a_{i,c})}\right)\left(1+\mathrm{e}^{-f(a_{j,c})}\right)}
\end{aligned} \tag{1}$$

$$K\left(\boldsymbol{m}_i, \boldsymbol{m}_j\right)=\exp\left(\frac{-\left\|\boldsymbol{m}_i-\boldsymbol{m}_j\right\|^2\lambda^2}{2}\right) \tag{3}$$

其中, λ 为控制因子, ACS$\in(0,1]$, ACS 越大社区内部的语义相似性越高, 社区划分结果越合理.

3) PWF (Paier wise F-measure) 模型[35]: 该模型建立了语义和社区相关性度量方法, 其表达式为

$$PWF=\frac{2\left|R\cap T\right|}{|R|+|T|} \tag{4}$$

其中, T 表示具有相同标签的相邻用户节点的集合, R 表示划分为同一社区的相邻用户节点的集合. PWF 的取值越大说明语义社区划分的结果越好.

4) PurQ 模型[89]: 该模型建立了模块度 Q 与语义话题相关性相混合的度量方法, 其表达式为

$$\begin{aligned} PurQ&=(1+\gamma^2)\frac{Purity\cdot Q}{\gamma^2\cdot Purity+Q}\\ Purity&=\frac{1}{|C|\sum\limits_{i=1}^{|C|}\max\limits_{1\le j\le k}\left\{\dfrac{n_{ij}}{|C_i|}\right\}}\end{aligned} \tag{5}$$

其中, Q 为社区划分结果的模块度, n_{ij} 表示属于社区 C_i 和话题 j 的用户节点个数, $Purity$ 表达了社区的话题关联性, 其取值越高话题的划分效果越好, γ 为 $Purity$ 与 Q 的平衡参数, 当 $0<\gamma<1$ 时, $PurQ$ 偏重于 $Purity$, 当 $1<\gamma<\infty$ 时, $PurQ$ 偏重于 Q.

5) NR (Network regulaziation) 模型[31]: 该模型在 GHF (Graph harmonic function)[90] 的基础上, 建立了基于矩阵分解模型的语义社区度量方法, 其表达式为

$$Reg=\frac{1}{2}\sum_{c\in C}\sum_{j=1}^{|C|}\boldsymbol{f}_j^{\mathrm{T}}\left(c\right)\left(D\left(c\right)-W(c)\right)\boldsymbol{f}_j\left(c\right) \tag{6}$$

其中, $W(c)$ 为社区 c 的权重邻接矩阵, $D(c)$ 为社区 c 的对角阵, 其中, $D_{u,u}\left(c\right)=\sum_v W_{u,v}\left(c\right)$, $\boldsymbol{f}_j(c)$ 为由社区 c 的所有用户节点对话题 i 的隶属度所构成的向量.

6) SCE (Semantic community entropy) 模型[91]: 该模型建立了以语义相似性量化为前提的语义社区度量方法, 其表达式为

$$\begin{aligned} H_C&=\sum_{i=1}^{|C|}H\left(C_i\right)\\ H\left(C_i\right)&=\sum_{j=1}^{|C_i|-1}\sum_{l=j+1}^{|C_i|}s_{jl}\ln s_{jl}+\\ &\quad\left(1-s_{jl}\right)\ln\left(1-s_{jl}\right)\end{aligned} \tag{7}$$

其中, s_{jl} 为用户节点 G_j 与 G_l 的相似性, H_C 的取值越小, 则社区内部的语义相似度越高, 语义相关性越强.

7) 全局语义相关性模型[49]: 该模型可表达网络 G 的语义相似性与拓扑结构的相关度, 其表达式为

$$Corr=\frac{\sum\limits_{\langle G_i,G_j\rangle\in L}w_{ij}\cos\left(\boldsymbol{m}_i,\boldsymbol{m}_j\right)}{\sqrt{\sum\limits_{\langle G_i,G_j\rangle\in L}w_{ij}^2}\sqrt{\sum\limits_{\langle G_i,G_j\rangle\in L}\cos\left(\boldsymbol{m}_i,\boldsymbol{m}_j\right)}} \tag{8}$$

其中, w_{ij} 为 G_i 与 G_j 间链接的权重, $\cos(\boldsymbol{m}_i,\boldsymbol{m}_j)$ 为 G_i 与 G_j 的话题向量 $\boldsymbol{m}_i$ 与 $\boldsymbol{m}_j$ 的余弦相似度. 本文在该理论的基础上建立了语义社区度量的语义 NMI 模型 (记作 $s\text{-}NMI$), 该模型以传统度量方法中的 NMI 模型[92] 为基础, 其表达式为

$$s\text{-}NMI\left(c,c'\right)=\frac{\sum\limits_{i\in c}\sum\limits_{i\in c'}p\left(i,j\right)\log\left(\dfrac{p\left(i,j\right)}{p_c\left(i\right)p_{c'}\left(j\right)}\right)}{\sqrt{\sum\limits_{i\in c}pc\left(i\right)\log p_c\left(i\right)\times\sum\limits_{j\in c'}p_{c'}\left(j\right)\log p_{c'}\left(j\right)}} \tag{9}$$

式中, p $(i,\ j)$ 为用户节点 G_i 和 G_j 的 1-dis 社区 (用户节点与其自身的邻居所构成的社区) 的交集内某些话题出现的概率, $p_c(i)$ 表示在 c 社区中的用户节点 G_i 所构成的 1-dis 社区内某些话题出现的概率. $s\text{-}NMI$ $(c,\ c')$ 表达了社区间的话题相关性.

8) OFS (Occurrence frequency similarity)[93]: 该模型利用了 Document 中关键字的出现频率及关键字间的相关性, 在关键字的分类记录中分析用户节点间的语义相似性. 其表达式如下:

$$\begin{aligned} of\left(i_u,i_x\right)=\frac{1}{|N|}\times\sum_{n\in N}&\left(1+\log\left(\frac{V}{f\left(i_un\right)}\right)\times\right.\\ &\left.\log\left(\frac{V}{f\left(i_un\right)}\right)\right)\end{aligned} \tag{10}$$

式中, N 为 Document 属性构成的集合, $|N|$ 为属性总数, V 为记录中关键字的总数. i_un 表示节点 i_u 的第 n 个属性, $f(i_un)$ 是节点 i_u 的第 n 个属性在总体记录中出现的次数, OFS$\in(0,1]$ 且 OFS 越大则户节点间的语义相似性越高.

9) PMI (Pointwise mutual information) 模型[94]: 该模型是在 Network similarity 模型[95] 作为拓扑结构相似性度量的基础上结合语义相似度分析, 对用户与社区相关性概率分布进行互信息建模. 若两个社区对同一用户的相关性相近, 则说明这

两个社区相似, 藉此实现语义社区间的相似性度量. Network similarity 模型的表达式如式 (11) 所示.

$$NS(u,x)=\frac{\log(|MFG(u,x)E|)}{\log(2|FG(u)E|)} \tag{11}$$

其中, $FG(u)E$ 为 G_u 的 1-dis 社区的内部边数. $MFG(u,x)E$ 为 G_u 的 1-dis 社区和 G_x 的 1-dis 社区所构成的社区交集的内部边数. 在 NS 模型基础上, PMI (Pointwise mutual information) 模型的表达式为

$$PMI(u,s)=P(F_u,F_s)\log\frac{P(F_u,F_s)}{P(F_u)\cdot P(F_s)} \tag{12}$$

其中, F_u 表示 G_u 的朋友圈, $P(F_u)$ 表示某一用户节点成为 G_u 的朋友的概率, $P(F_u,F_s)$ 表示某一用户节点成为 G_u 和 G_s 的共同朋友的概率.

10) 传统社区评价模型的语义改进模型: 语义社会网络的分析结果可量化为用户节点的多维属性 $\boldsymbol{m}_i$, 且用户节点 G_i 和 G_j 间的相似性 $U(\boldsymbol{m}_i,\boldsymbol{m}_j)$ 可代表 $l(i,j)$ 的权重, 因此, 可利用 $U(\boldsymbol{m}_i,\boldsymbol{m}_j)$ 建立网络 G 的有权邻接矩阵 S, $S_{ij}=U(\boldsymbol{m}_i,\boldsymbol{m}_j)$.

表 1 为传统社会网络的社区评价指标, 其中 $C_i^{\rm out}=\sum_{p\in C_i,q\notin C_i}A_{pq}$, $C_i^{\rm in}=\sum_{p,q\in C_i}A_{pq}$, $|L|=\sum_{p,q\in G}A_{pq}$, $C_i^{\rm in}(j)=\sum_{p\in C_i}A_{jp}$, $C_i^{\rm out}(j)=\sum_{p\notin C_i}A_{jp}$, 在语义社会网络的评价方面, 本文借用传统社会网络评价模型的表达形式, 对其中的参数进行了如下改造 $C_i^{\rm out}=\sum_{p\in C_i,q\notin C_i}S_{pq}$, $C_i^{\rm in}=\sum_{p,q\in C_i}S_{pq}$, $|L|=\sum_{p,q\in G}S_{pq}$, $C_i^{\rm in}(j)=\sum_{p\in C_i}S_{jp}$, $C_i^{\rm out}(j)=\sum_{p\notin C_i}S_{jp}$, 使传统社会网络评价模型可评价语义社区划分结果. 经过改造后的模型记为 "*s-model*" (如 *s-EQ*).

2.2 度量方法的实验对比

本文利用实验分析对各类语义社区度量方法进行了社区结构相关性测试及语义相关性测试. 由于某些度量方法 (如 AF 模型) 需要预先设定参数, 且参数的取值决定了模型的度量性能. 由于对各模型的参数进行对比讨论需要大量的篇幅. 对此, 本文仅从对比分析的实验方法出发, 对各类语义社区度量方法进行方法性比较, 不作参数取值的对比, 其中模型的参数根据相关文献进行设定.

1) 数据集与度量方法

在实验数据集方面, 本文利用 LFR Benchmark[96−97] 生成模拟数据集 G_1000 作为语义社会网络的关系拓扑. 其参数设置为 ($|G|=1\,000$, $ad=4$, $dmax=16$, $cmin=15$, $cmax=50$, $on=80$, $om=5$, $mi=2.5$), 其中参数 $|G|$ 表示用户节点的个数; ad 和 $dmax$ 分别表示网络中用户节点的平均度和最大度; $cmin$ 和 $cmax$ 分别表示最小社区和最大社区中用户节点的数量; on 表示重叠用户节点个数; om 表示每个重叠用户节点连接的社区个数; mi 为混合系数, 表示用户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, 网络的社区结构不明显.

在语义数据模拟方面, 本文随机选择 30 组话题, 每组话题包含 200 个关键字. 选择 G_1000 中度数最大的 30 个用户节点作为话题源用户节点, 为每个

表 1 各类社区评价模型

Table 1 Various measurement models for community detection

算法	表达式
EQ[11]	$EQ=\frac{1}{2\|L\|}\sum_{i=1}^{\|C\|}\sum_{v,w\in C_i}\frac{1}{O_vO_w}\left(A_{vw}-\frac{\text{degree}(v)\text{degree}(w)}{2\|L\|}\right)$
s-EQ	$s\text{-}EQ=\frac{1}{2\|L\|}\sum_{i=1}^{\|C\|}\sum_{v,w\in C_i}\frac{\cos(\boldsymbol{m}_i,\boldsymbol{m}_j)}{O_vO_w}\left(A_{vw}-\frac{\text{degree}(v)\text{degree}(w)}{2\|L\|}\right)$
Average condutance[41]	$AC=\frac{1}{\|C\|}\sum_{i=1}^{\|C\|}\frac{C_i^{\rm out}}{\min\left(\left(C_i^{\rm out}+\frac{1}{2}C_i^{\rm in}\right),\left(2\|L\|-C_i^{\rm out}-\frac{1}{2}C_i^{\rm in}\right)\right)}$
MinMaxCut[98]	$MMC=\sum_{i=1}^{\|C\|}\frac{2C_i^{\rm out}}{C_i^{\rm in}}$
Silhouette[99]	$Sil=\frac{1}{\|C\|}\sum_{j=1}^{\|C\|}\left(\frac{1}{\|C_j\|}\sum_{i\in C_j}\frac{a_i-b_i}{\max(b_i,a_i)}\right)$, $a_i=\frac{1}{\|C_k\|}\sum_{i,j\in C_k}A_{ij}$, $b_i=\max\left(\frac{1}{\|C_r\|}\sum_{i\notin C_r,j\in C_r}A_{ij}\right)$
Ductance[100]	$Duc=\sum_{i=1}^{\|C\|}\frac{C_i^{\rm out}}{C_i^{\rm in}+C_i^{\rm out}}$
Expansion[101]	$Exp=\sum_{i=1}^{\|C\|}\frac{C_i^{\rm out}}{\|C_i\|}$
NCut[102]	$NCut=\sum_{i=1}^{\|C\|}\frac{C_i^{\rm out}}{C_i^{\rm in}+C_i^{\rm out}}+\frac{C_i^{\rm out}}{2\left(\|L\|-\frac{1}{2}C_i^{\rm in}\right)+C_i^{\rm out}}$
Average fitness[12]	$AF=\frac{1}{\|C\|}\sum_{i=1}^{\|C\|}\sum_{j\in C_i}\frac{C_i^{\rm in}(j)}{\left(C_i^{\rm in}(j)+C_i^{\rm out}(j)\right)^r}$

话题源用户节点分配一组话题, 并记录各话题与关键字的对应关系. 网络中各用户节点同时随机选择自身的 80 个关键字向邻居用户节点进行扩散, 如此循环直到网络中每个用户节点分配到的关键字数量均大于 200. 此过程模拟话题的传播特性、各节点所分配的关键字可作为其语义信息. 在多维话题向量生成方面, 本文利用关键字与话题的对应关系, 统计话题分配次数. 将每个用户节点的话题分配次数作为话题向量 $\boldsymbol{m}$, 并将话题向量进行整理 $m_i = m_i/\max(\boldsymbol{m})$, 所得到的多维属性作为各用户节点的话题隶属度.

在度量方法的选择及参数设置方面, 选用余弦相似度作为多维相似度度量, 对比模型为: Qov ($p=0.5$), ACS ($\lambda=1.5$), PWF, $PurQ$ ($\gamma=1$), NR, SCE, $s\text{-}NMI$, PMI, $s\text{-}EQ$, $s\text{-}AC$, $s\text{-}MMC$, $s\text{-}Sil$, $s\text{-}Duc$, $s\text{-}Exp$, $s\text{-}Ncut$ 和 $s\text{-}AF$ 共 16 个模型 (由于 OFS 模型不含社区结构度量, 不对其进行比较). 其中, Qov, ACS, PWF, $PurQ$, NR, $s\text{-}NMI$, $s\text{-}EQ$, $s\text{-}Sil$, $s\text{-}Exp$ 和 $s\text{-}AF$ 的取值越大语义社区划分的结果越优, SCE, PMI, $s\text{-}AC$, $s\text{-}MMC$, $s\text{-}Duc$ 和 $s\text{-}Ncut$ 的取值越小语义社区划分的结果越优. 语义社区评价模型需要综合考虑社区结构的拓扑相关性及语义相关性, 由于各模型的表达式不同, 其对拓扑相关性及语义相关性的敏感度不同. 为此, 本实验设计了社区结构相关性测试实验及语义相关性测试实验.

2) 拓扑相关性测试, 分析语义数据不变的条件下各评价模型对网络拓扑结构的依赖性.

本实验首先利用 G_1000 的网络生成方法及语义数据模拟方法, 随机生成 20 组 G_1000 数据, 并在各用户节点的语义数据不变的条件下, 对每组数据的 2 个社区之间随机加边, 由于 LFR Benchmark 具有固定的社区结构, 社区间的随机加边会导致社区的结构变得模糊, 从而导致模块度 EQ 下降; 其次, 在社区的模糊化时, 跟踪 20 组数据中各语义评价指标的变化情况, 以此来评价各模型对拓扑相关性的依赖; 最后, 记录 20 组 G_1000 数据 EQ 从 0.6 到 0.2 的变化过程中各评价模型的取值变化, 并将评价模型的取值映射在 (0, 1) 区间内以增加可对比性. 图 7 为 16 个模型在 20 组 G_1000 数据中的社区结构相关性对比.

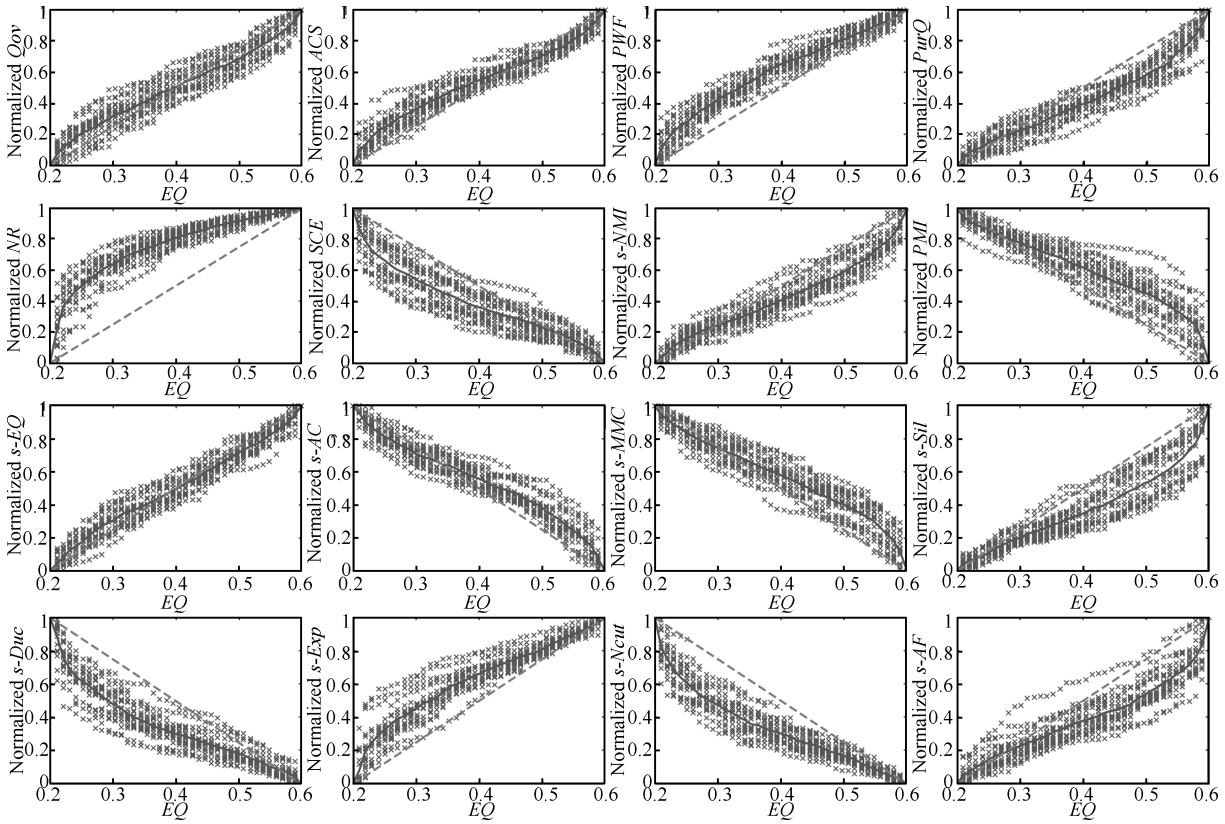


图 7　各评价模型的社区结构相关性测试

Fig. 7　The correlation test on community structure for each evaluation model

图 7 中虚线为 EQ 的线性变化辅助线, 实线为 20 组 G_1000 数据的均值. 各评价模型与 EQ 的相关系数分别为: $Qov(0.98)$, $ACS(0.88)$, $PWF(0.83)$, $PurQ(0.82)$, $NR(0.68)$, $SCE(-0.73)$, $s\text{-}NMI(0.80)$, $PMI(-0.75)$, $s\text{-}EQ(0.98)$, $s\text{-}AC(-0.90)$, $s\text{-}MMC(-0.89)$, $s\text{-}Sil(0.84)$, $s\text{-}Duc(-0.76)$, $s\text{-}Exp(0.80)$, $s\text{-}Ncut(-0.72)$ 和 $s\text{-}AF(0.76)$, 说明 NR, SCE, PMI, $s\text{-}Duc$, $s\text{-}Duc$, $s\text{-}Ncut$ 和 $s\text{-}AF$ 与模块度 Q 的线性变化偏差较大, 当社区拓扑相关性 (非语义结构) 发生改变时, 上述模型的变化量不与拓扑相关性的变化量线性相关, 即上述评价模型更强调语义相关性, 适合用于评价语义社区划分.

3) 语义相关性测试, 分析网络拓扑结构不变的条件下各评价模型对语义数据的依赖性.

本实验首先利用 G_1000 的网络生成方法及语义数据模拟方法随机生成 20 组 G_1000 数据, 并在 20 组数据的拓扑结构不变的条件下, 按高斯场的势能函数[90, 103] 在关键字的扩散过程中增加扩散系数 $\exp(-(\text{step}/\sigma)^2)$, 其中 $\text{step}=0, 1, \cdots$ 为扩散的步数; 其次以扩散系数作为关键字扩散的权重, 跟踪 20 组 G_1000 数据在 $\sigma \in (1, 5)$ 时, 各评价模型的取值, 其中 σ 较大时, 各 step 的扩散系数趋于均等, 导致社区间的语义相似性高, 即 σ 越大语义相关性越差; 最后, 记录 20 组 G_1000 数据 σ 从 1 到 5 的变化过程中各评价模型的变化状态, 并将评价模型的取值映射在 (0, 1) 区间内以增加可对比性. 图 8 为 16 个模型在 20 组 G_1000 数据中的语义相关性对比, 其中虚线为 σ 的线性变化辅助线, 实线为 20 组 G_1000 数据的均值. 各评价模型与 σ 的相关系数分别为: $Qov(-0.95)$, $ACS(-0.90)$, $PWF(-0.76)$, $PurQ(-0.78)$, $NR(-0.88)$, $SCE(0.88)$, $s\text{-}NMI(-0.76)$, $PMI(0.87)$, $s\text{-}EQ(-0.88)$, $s\text{-}AC(0.81)$, $s\text{-}MMC(0.81)$, $s\text{-}Sil(-0.82)$, $s\text{-}Duc(0.80)$, $s\text{-}Exp(-0.88)$, $s\text{-}Ncut(0.78)$ 和 $s\text{-}AF(-0.77)$, 说明 PWF, $PurQ$, $s\text{-}Ncut$ 和 $s\text{-}AF$ 与 σ 的线性变化偏差较大, 说明上述评价模型的变化量不与语义相关性的变化量线性相关, 即上述评价模型更强调拓扑相关性, 适合用于评价传统非语义社区划分. 综合拓扑相关性测试的结果, $s\text{-}Ncut$ 和 $s\text{-}AF$ 的综合评价性能较强.

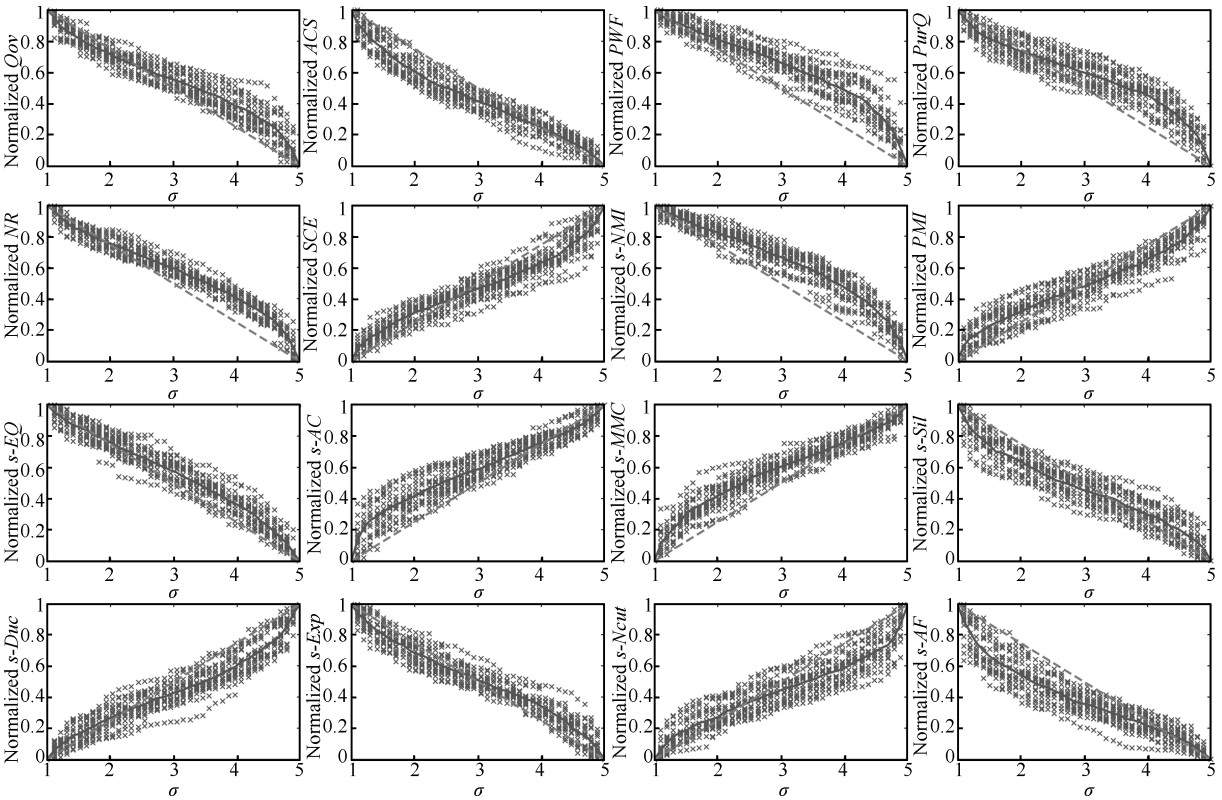


图 8 各评价模型的语义相关性测试

Fig. 8 The correlation tests on semantic for each evaluation model

3 结论

本文对基于话题概率模型的语义社区发现方面的相关文献进行了归纳, 在描述其各自实现过程的同时总结了各种算法的优缺点, 其中优点可作为后续研究的指导思想, 缺点可作为后续研究的改进方向. 语义社会网络的社区挖掘兼备了语义分析和拓扑分析 2 方面技术, 本文所列举的大多文献的共同缺点在于没有将二者进行有效结合, 如在语义分析过程中没有考虑拓扑结构对语义分析的影响, 以及在社区挖掘过程中融入语义–拓扑混合的相关性分析. 对此, 后续的工作需要借鉴以往算法的经验和不足, 在语义–拓扑层面对语义社区挖掘方法进行创新.

本文最后综述了语义社区的评价模型, 并利用实验分析对比了各模型的评价特性. 由于本文的写作目的在于综述各类评价模型的表达形式及设计思想, 其实验的目的在于方法性说明, 因此, 实验过程及参数讨论相对简略. 在后续的研究中, 可利用本文所提供的实验方法, 改进语义数据的话题标签扩散方式, 以及利用社区的结构特性指导话题的扩散方向, 实现评价模型的多角度对比.

由于基于话题概率模型的语义社区发现具有较复杂的文本训练过程, 且仅适用于静态社会网络. 对此, 在未来大数据环境下, 需要解决话题模型的简化, 并建立动态流式数据的增量挖掘模式, 实现动态语义社会网络 (包含节点的动态交互及语义信息的动态变化) 的动态语义社区挖掘, 包括动态语义社区发现、语义社区推荐、动态语义话题溯源等. 综上所述, 本文对语义社区挖掘相关文献的归纳和总结, 可作为该领域后续工作的参考.

References

1 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821−7826

2 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. *Nature*, 1998, **393**(6684): 440−442

3 Golbeck J, Rothstein M. Linking social networks on the web with FOAF: a semantic web case study. In: Proceeding of the 23rd Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence. Menlo Park, Canada: AAAI, 2008. 1138−1143

4 Klyne G, Carroll J J. Resource description framework (RDF): concepts and abstract syntax [Online], available: http://www.w3.org/TR/REC-rdf-syntax, February 1, 2006

5 Brickley D, Guha R V. RDF vocabulary description language 1.0: RDF schema. W3C Recommendation 10 February 2004, World Wide Web Consortium, 004 [Online], available: http://www.w3.org/TR/rdf-schema/, November, 2006

6 W3C OWL Working Group. OWL 2 web ontology language: document overview. W3C Recommendation 27 October 2009, World Wide Web Consortium, 2009[Online], available: http://www.w3.org/TR/2009/REC-owl2-overview-20091027/, November, 2007

7 Prud′hommeaux E, Seaborne A, Laboratories H P, Bristol. SPARQL query language for RDF. W3C Recommendation 15 January 2008, World Wide Web Consortium, 2008 [Online], available: http://www.w3.org/TR/rdf-sparql-query/, Auguest, 2007

8 Newman M E. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, **69**(6): 066133

9 Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, DOI: 10.1088/1742-5468/2008/10/P10008

10 Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, **435**(7043): 814−818

11 Shen H W, Cheng X QI, Cai K, Hu M B. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 2009, **388**(8): 1706−1712

12 Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 2009, **11**(3): 033015

13 Gregory S. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 2010, **12**(10): 103018

14 Berners-Lee T, Hendler J, Lassila O. The semantic web. *Scientific American*, 2001, **284**(5): 28−37

15 Thovex C, Trichet F. Semantic social networks analysis. *Social Network Analysis and Mining*, 2013, **3**(1): 35−49

16 Yang Jing, Xin Yu, Xie Zhi-Qiang. Semantics social network community detection algorithm based on topic comprehensive factor analysis. *Journal of Computer Research and Development*, 2014, **51**(3): 559−569
(杨静, 辛宇, 谢志强. 基于话题综合因子分析的语义社会网络社区发现算法. 计算机研究与发展, 2014, **51**(3): 559−569)

17 Blei D M, Ng A Y, Jordan M I. Latent dirichllocation. *Journal of Machine Learning Research*, 2003, **3**(8): 993−1022

18 Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1999. 50−57

19 Bischof J, Airoldi E. Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning. New York, USA: ICML, 2012. 201−208

20 Barabási A L, Albert R. Emergence of scaling in random networks. *Science*, 1999, **286**(5439): 509−512

21 Ding Y. Community detection: topological vs. topical. *Journal of Informetrics*, 2011, **5**(4): 498−514

22 Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2004. 306−315

23 McCallum A, Corrada-Emmanuel A, Wang X R. Topic and role discovery in social networks. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco, USA: ACM, 2005. 786−791

24 McCallum A, Wang X R, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 2007, **30**(1): 249−272

25 Pathak N, DeLong C, Banerjee A, Erickson K. Social topic models for community extraction. In: Proceedings of the 2nd SNA-KDD Workshop. New York, USA: ACM, 2008. 1−8

26 Airoldi E M, Blei D M, Fienberg S E, Xing E P, Jaakkola T. Mixed membership stochastic block models for relational data with application to protein-protein interactions. In: Proceedings of the 2006 International Biometrics Society Annual Meeting. Tampa, FL, USA: ENAR, 2006. 23−31

27 Wang X R, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2006. 424−433

28 Griffiths T. Gibbs sampling in the generative model of latent dirichlet allocation. *Standford University*, 2002, **518**(11): 1−3

29 Wang X R, McCallum A, Wei X. Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: Proceeding of 7th IEEE International Conference on Data Mining. Omaha, USA: IEEE, 2007. 697−702

30 Wallach H. Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM, 2006. 977−984

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

32 White S, Smyth S. A spectral clustering approach to finding communities in graphs. In: Proceedings of the 5th SIAM International Conference on Data Mining. Houston, USA: SIAM, 2005. 76−84

33 Azran A, Ghahramani Z. Spectral methods for automatic multiscale data clustering. In: Proceeding of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2006. 190−197

34 Lin C H, He Y L. Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2009. 375−384

35 Yang T B, Jin R, Chi Y, Zhu S H. Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2009. 927−936

36 Lacoste-Julien S, Sha F, Jordan M. DiscLDA: discriminative learning for dimensionality reduction and classification. In: Proceeding of the 2009 Advances in Neural Information Processing Systems. Piscataway, USA: NIPS, 2009. 897−904

37 Li D F, He B, Ding Y, Tang J, Sugimoto C, Qin Z, Yan E J, Li J Z, Dong T X. Community-based topic modeling for social tagging. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2010. 1565−1568

38 Leskovec J, Lang K J, Mahoney M. Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM, 2010. 631−640

39 Tang J, Jin R M, Zhang J. A topic modeling approach and its integration into the random walk framework for academic search. In: Proceeding of 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008. 1055−1060

40 Tang J, Zhang J, Yao L M, Li J Z, Zhang L, Su Z. ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008. 990−998

41 Leskovec J, Lang K J, Dasgupta A, Mahoney M W. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 2009, **6**(1): 29−123

42 Ríos S A, Muñoz R. Dark Web portal overlapping community detection based on topic models. In: Proceedings of the 2012 ACM SIGKDD Workshop on Intelligence and Security Informatics. New York, USA: ACM, 2012. 1−7

43 Xie J R, Szymanski B K, Liu X M. Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Proceeding of the 11th IEEE International Conference on Data Mining Workshops. Piscataway, USA: IEEE, 2011. 344−349

44 Barber M J, Clark J W. Detecting network communities by propagating labels under constraints. *Physical Review E*, 2009, **80**(2): 026129.1-026129.11

45 Jang J, Myaeng S H. Discovering dedicators with topic-based semantic social networks. In: Proceeding of the 7th International Conference on Weblogs and Social Media. Boston, USA: ICWSM, 2013. 1−12

46 Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(S1): 5220−5227

47 Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences. In: Proceedings of the 24th International Conference on Machine learning. Piscataway, USA: ACM, 2007. 233−240

48 Chang J, Blei D W. Relational topic models for document networks. In: Proceedings of the 2009 International Conference on Artificial Intelligence and Statistics. Piscataway, USA: IEEE, 2009. 81−88

49 Sun Y Z, Han J W, Gao J T. Itopicmodel: information network-integrated topic modeling. In: Proceeding of 9th IEEE International Conference on Data Mining. Miami, FL, USA: IEEE, 2009. 493−502

50 Perez P. Markov random fields and images. *CWI Quarterly*, 1998, **11**(4): 413−437

51 Liu Y, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning. Piscataway, USA: ACM, 2009. 665−672

52 L′Huillier G, Ríos S A, Alvarez H, Aguilera F. Topic-based social network analysis for virtual communities of interests in the dark web. In: Proceeding of the 2010 ACM SIGKDD Workshop on Intelligence and Security Informatics. New York, USA: ACM, 2010. 23−31

53 Zheng G Q, Guo G W, Yang L C, Xu S L, Bao S H, Su Z, Han D Y, Yu Y. Mining topics on participations for community discovery. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2011. 445−454

54 Chahal P, Singh M, Kumar S. An ontology based approach for finding semantic similarity between web documents. *International Journal of Current Engineering and Technology*, 2013, **3**(5): 1925−1931

55 Kleinberg J M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, **46**(5): 604−632

56 Kemp C, Tenenbaum J B. Learning systems of concepts with an infinite relational model. In: Proceeding of the 21st National Conference on Artificial Intelligence. Menlo Park, Canada: AAAI, 2006. 1−15

57 Long B, Zhang Z M, Yu P S. A probabilistic framework for relational clustering. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2007. 470−479

58 Zhu S H, Yu K, Chi Y, Gong Y H. Combining content and link for classification using matrix factorization. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007. 487−494

59 Xu W, Liu X, Gong Y H. Document clustering based on nonnegative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2003. 267−273

60 Wang F, Li T, Wang X, Zhu S G, Ding C. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 2011, **22**(3): 493−521

61 Nallapati R, Cohen W. Link-PLSA-LDA: a new unsupervised model for topics and influence in blogs. In: Proceeding of the 2nd International Conference on Weblogs and Social Media. Piscataway, USA: AAAI, 2008. 1−12

62 Nallapati R M, Ahmed A, Xing E P, Cohen W W. Joint latent topic models for text and citations. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008. 542−550

63 Yin Z J, Cao L L, Gu Q Q, Han J W. Latent community topic analysis: integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 2012, **3**(4): 63−63

64 Cha Y, Cho J. Social-network analysis using topic models. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2012. 565−574

65 Hu B, Song Z, Ester M. User features and social networks for topic modeling in online social media. In: Proceeding of the 2012 International Conference on Advances in Social Networks Analysis and Mining. Washington D. C., USA: ACM, 2012. 202−209

66 Natarajan N, Sen P, Chaoji V. Community detection in content-sharing social networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, USA: ACM, 2013. 82−89

67 Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 2009, **80**(1): 016105

68 Nowicki K, Snijders T A B. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 2001, **96**(455): 1077−1087

69 Jamali M, Huang T L, Ester M. A generalized stochastic block model for recommendation in social rating networks. In: Proceedings of the 5th ACM Conference on Recommender Systems. New York, USA: ACM, 2011. 53−60

70 Wang X R, Mohanty N, McCallum A. Group and topic discovery from relations and text. In: Proceedings of the 3rd International Workshop on Link Discovery. New York, USA: ACM, 2005. 28−35

71 Zhou D, Manavoglu E, Li J, Giles C L, Zha H Y. Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web. New York, USA: ACM, 2006. 173−182

72 Zhang H Z, Qiu B J, Giles C L, Foley H C, Yen J. An LDA-based community structure discovery approach for large-scale social networks. In: Proceeding of the 2007 IEEE Intelligence and Security Informatics. New Brunswick, NJ, USA: IEEE, 2007. 200−207

73 Zhang H Z, Giles C L, Foley H C, Yen H. Probabilistic community discovery using hierarchical latent gaussian mixture model. In: Proceeding of the 22nd National Conference on Artificial Intelligence. Menlo Park, Canada: AAAI, 2007. 663−668

74 Zhang H Z, Li W, Wang X R, Giles C L, Foley H C, Yen J. HSN-PAM: finding hierarchical probabilistic groups from large-scale networks. In: Proceeding of the 7th IEEE International Conference on Data Mining Workshops. Omaha, NE, USA: IEEE, 2007. 27−32

75 Sachan M, Contractor D, Faruquie T, Subramaniam V. Probabilistic model for discovering topic based communities in social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2011. 2349−2352

76 Henderson K, Eliassi-Rad T. Applying latent dirichlet allocation to group discovery in large graphs. In: Proceedings of the 2009 ACM Symposium on Applied Computing. New York, USA: ACM, 2009. 1456−1461

77 Henderson K, Elisssi-Rad T, Papadimitriou S, Faloutsos C. HCDF: a hybrid community discovery framework. In: Proceedings of the 2010 SIAM International Conference on Data Mining. Columbus, Ohio, USA: SIAM, 2010. 754−765

78 Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks. *Link Mining: Models, Algorithms and Applications*. New York: Springer, 2010. 337−357

79 Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: Proceedings of the 19th International World Wide Web Conference Series. New York, USA: ACM, 2010. 591−600

80 Sachan M, Contractor D, Faruquie T A, Subramaniam L V. Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web. New York, USA: ACM, 2012. 331−340

81 Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping semantic community structure detecting algorithm by label propagation. *Acta Automatica Sinica*, 2014, **40**(10): 2262−2275
(辛宇, 杨静, 谢志强. 基于标签传播的语义重叠社区发现算法. 自动化学报, 2014, **40**(10): 2262−2275)

82 Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping community structure detecting algorithm in semantic social network based on the block field. *Acta Automatica Sinica*, 2015, **41**(2): 362−375
(辛宇, 杨静, 谢志强. 一种面向语义重叠社区发现的 Block 场取样算法. 自动化学报, 2015, **41**(2): 362−375)

83 Xin Y, Yang J, Xie Z Q. A semantic overlapping community detection algorithm based on field sampling. *Expert Systems with Applications*, 2015, **42**(1): 366−375

84 Xin Y, Yang J, Xie Z Q, Zhang J P. An overlapping semantic community detection algorithm base on the ARTs multiple sampling models. *Expert Systems with Applications*, 2015, **42**(7): 3420−3432

85 Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009, (3): P03024

86 Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**(2): 026113

87 Java A, Joshi A, Finin T. Detecting commmunities via simultaneous clustering of graphs and folksonomies. In: Proceedings of the 10th Workshop on Web Mining and Web Usage Analysis. New York, USA: ACM 2008. 23−32

88 Chen S, Cowan C F N, Grant P M. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 1991, **2**(2): 302−309

89 Zhao Z Y, Feng S Z, Wang Q L, Huang Z X, Graham J W, Fan J P. Topic oriented community detection through social objects and link analysis in social network. *Knowledge-Based Systems*, 2012, *26*: 164−173

90 Zhu X J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20nd International Conference on Machine Learning. Washington D. C., USA: MIT Press, 2003. 912−919

91 Cruz J D, Bothorel C, Poulet F. Entropy based community detection in augmented social networks. In: Proceedings of the 2011 International Conference on Computational Aspects of Social Network. Salamanca, Apain: IEEE, 2011. 163−168

92 Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 2003, **3**: 583−617

93 Rawashdeh A, Rawashdeh M, Díaz I, Ralescu A. Measures of semantic similarity of nodes in a social network. In: Proceeding of the 15th Information Processing and Management of Uncertainty in Knowledge-Based Systems. Piscataway, USA: IEEE, 2014. 76−85

94 Akcora C G, Carminati B, Ferrari E. Network and profile based measures for user similarities on social networks. In: Proceeding of the 2011 IEEE International Conference on Information Reuse and Integration. Las Vegas, NV, USA: IEEE, 2011. 292−298

95 Deshpande M, Karypis G. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, 2004, **22**(1): 143−177

96 Demir G N, Uyar A S, Ögüdücü S G. Graph-based sequence clustering through multiobjective evolutionary algorithms for web recommender systems. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. New York, USA: ACM, 2007. 1943−1950

97 Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987, **20**: 53−65

98 Kannan R, Vempala S, Vetta A. On clusterings: good, bad and spectral. *Journal of the ACM*, 2004, **51**(3): 497−515

99 Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(9): 2658−2663

100 Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 888−905

101 Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 2008, **78**(4): 046110

102 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

103 Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011, **73**(4): 423−498

辛 宇 哈尔滨理工大学计算机科学与技术学院讲师. 2015 年获得哈尔滨工程大学计算机应用技术博士学位. 主要研究方向为数据库与知识工程.
E-mail: xinyu@hrbeu.edu.cn
(**XIN Yu** Ph. D., lecturer at the College of Computer Science and Technology, Harbin University of Science and Technology. He received his Ph. D. degree from Harbin Engineering University in 2015. His research interest covers database and knowledge engineering.)

谢志强 哈尔滨理工大学计算机科学与技术学院教授. 主要研究方向为数据库与知识工程.
E-mail: xiezhiqiang@hrbust.edu.cn
(**XIE Zhi-Qiang** Professor at the College of Computer Science and Technology, Harbin University of Science and Technology. His research interest covers database and knowledge engineering.)

杨 静 哈尔滨工程大学计算机科学与技术学院教授. 主要研究方向为数据库与知识工程. 本文通信作者.
E-mail: yangjing@hrbeu.edu.cn
(**YANG Jing** Professor at the College of Computer Science and Technology, Harbin Engineering University. Her research interest covers database and knowledge engineering. Corresponding author of this paper.)