计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):257−265 [doi: 10.15888/j.cnki.csa.008915]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于社区与结构熵的异质网络影响力最大化①

徐智敏, 周丽华, 刘　超

(云南大学 信息学院, 昆明 650500)
通信作者: 周丽华, E-mail: lhzhou@ynu.edu.cn

摘　要: 影响力最大化的目的是在网络中发现能够触发最大数量的剩余节点参与到信息传播过程的一小群节点. 目前异质信息网络中影响力最大化的研究通常从网络中抽取同质子图、或基于节点局部结构的元路径进行节点影响力的评估, 没有考虑节点的全局特征和网络中高影响力节点间的集群现象给种子集合最终扩散范围造成的影响损失. 文中提出了一种基于社区与结构熵的异质信息网络影响力最大化算法, 该算法能够有效地从局部和全局两个方面度量节点的影响. 首先, 通过构建元结构保留节点在网络中的局部结构信息和异质信息度量节点的局部影响; 其次, 利用节点所属社区在整个网络中的权重占比对节点的全局影响进行度量; 最后, 综合求出节点的最终影响并选出种子集合. 在真实数据集上进行的大量实验结果表明所提算法有较好的有效性和效率.

关键词: 异质信息网络; 影响力最大化; 富人俱乐部现象; 结构熵; 社区

引用格式: 徐智敏,周丽华,刘超.基于社区与结构熵的异质网络影响力最大化.计算机系统应用,2023,32(1):257–265. http://www.c-s-a.org.cn/1003-3254/8915.html

Influence Maximization of Heterogeneous Networks Based on Community and Structure Entropy

XU Zhi-Min, ZHOU Li-Hua, LIU Chao

(School of Information Science and Engineering, Yunnan University, Kunming 650500, China)

Abstract: The purpose of influence maximization is to find a small group of nodes in a network that can trigger the maximum number of remaining nodes to participate in the process of information transmission. At present, the research on the influence maximization of heterogeneous information networks usually extracts homogeneous subgraphs from the network or evaluates the influence of nodes according to the meta-path of local node structure. However, it does not consider the global features of nodes and the influence loss of the final spread range of the seed set caused by the clustering phenomenon among highly influential nodes. This study proposes an influence maximization algorithm for heterogeneous information networks based on community and structure entropy, which can effectively measure the influence of nodes locally and globally. Firstly, the local structure information and heterogeneous information of nodes in the network are retained by the construction of meta-structure to measure the local influence of nodes. Secondly, the global influence of nodes is measured by the weight ratio of the community to which the nodes belong to the whole network. Finally, the final influence of nodes is calculated, and the seed set is selected. Many experiments on real data sets indicate that the proposed algorithm is effective and efficient.

Key words: heterogeneous information network; influence maximization; rich-club phenomenon; structure entropy; community

① 基金项目: 国家自然科学基金 (62062066, 61762090, 61966036); 云南省基础研究计划重点项目 (202201AS070015); 云南省高校物联网技术及应用重点实验室项目; 国家社会科学基金 (18XZZ005); 云南省教育厅科学研究基金 (2021Y026); 云南大学研究生科研创新项目 (2021Y024)
收稿时间: 2022-06-07; 修改时间: 2022-07-06; 采用时间: 2022-07-20; csa 在线出版时间: 2022-09-01
CNKI 网络首发时间: 2022-11-15

随着互联网技术的创新与进步, 在线社交网络作为社会网络中一种重要的表现形式, 已经在人们的生活中发挥了愈发重要的作用. 而挖掘网络中有影响力的节点成为企业和组织进行病毒营销、目标广告、舆论控制和社会推荐等领域的主要选择之一, 有着极大的商业价值. 比如在病毒营销模式中, 考虑到企业用于新产品广告预算的有限和直接联系所有客户人群的不可能性, 应当选择有限数量的在线社交网络用户来形成初始子集以启动传播消息的过程, 而确定网络中有影响力的成员成为这种模式能付诸行动的关键因素, 影响力最大化 (influence maximization, IM) 问题也因此被提出.

IM 问题旨在发现网络中的一小群节点, 这些节点能在信息传播模型下触发最大数量的剩余节点参与信息传播过程. 文献 [1] 最开始从理论的角度提出了 IM 问题, 并从算法的角度对 IM 问题进行了建模. 文献 [2] 证明了 IM 问题是一个离散优化问题, 并通过贪心算法求得了一个近似解. 然而, 由于利用贪心算法进行 IM 问题的求解有着高昂的时间代价, 因此学者们纷纷在贪心算法的基础上加以完善, 或提出新的策略来提高运行的效率. 如基于社区的算法[3,4]、基于启发式策略的算法[5,6] 或利用信息熵度量节点影响力[7,8] 的算法等.

然而这些传统的关于 IM 问题的研究大多将社会网络建模为同质信息网络, 即在网络建模的过程中仅考虑了一种关系类型和一种对象类型. 但在实际生活中, 对象之间的关系往往更加复杂且多变. 采用同质信息网络建模的方法无法保留对象之间的复杂关系, 会造成信息的不完整或缺失.

较于同质信息网络, 包含有丰富的关系和对象类型的异质信息网络能有效地融合多种对象类型的节点, 保留更加完整的网络结构和语义信息, 实现对实际生活中的社会关系更加真实和完整的抽象. 基于信息融合的优势, 在异质信息网络上进行 IM 问题的研究更有利于舆论控制、目标广告等应用的开发和更精确地度量节点的影响. 文献 [9] 提出了一种关于熵的异质信息网络 IM 算法, 通过限制对象路径的长度构造同质子图以评估节点的影响. 文献 [10] 利用元路径从异质信息网络中抽取多个同质子图, 通过综合目标对象在不同子图中的影响力以求解 IM 问题.

尽管在异质信息网络中关于 IM 问题的研究已经取得了一些初步成果, 但在异质信息网络中节点之间的连接关系更加丰富、网络结构也更加全面, 因此 IM 问题也还有进一步研究的空间. 如: 目前有关于异质信息网络中 IM 问题的研究, 大多基于元路径提取节点的连接关系, 并未考虑全局结构对节点的影响, 同时也忽略了大部分网络所具有的“富人俱乐部现象”对节点影响评估时所造成的影响损失. “富人俱乐部现象”是大多数网络中所具有的一种重要的结构特征, 在网络中主要体现为具有较高影响力的节点之间连接关系往往更加紧密, 例如在论文学术网络中, 具有较大影响力的作者之间往往倾向于彼此合作.

基于以上不足, 本文提出了一种基于社区与结构熵的影响力最大化算法 (community and structure entropy influence maximization, CMIM). CMIM 算法首先通过元结构提取并保留节点在网络中的局部结构和异质关系, 以度量节点的局部影响. 然后引入对节点影响传播有着重要作用的社区结构, 通过评估节点所属社区在整个网络中的权重占比作为节点全局影响力的度量标准, 同时提升了社区边缘节点的全局影响以减少“富人俱乐部现象”带来扩散范围的损失, 本文主要贡献总结如下.

(1) 提出了基于社区与结构熵的异质信息网络影响力大化算法, 从全局和局部两个角度对节点的影响进行评估, 更有利于提升节点影响力度量的准确性.

(2) 提出了一种基于社区的节点全局影响力度量方式, 通过基于元路径的拓扑势社区划分算法将异质信息网络划分为多个重叠社区, 并提升社区边缘节点的权重占比减少了“富人俱乐部现象”带来的影响损失.

(3) 通过在真实数据集上进行影响范围对比、参数分析等实验验证了所提算法的有效性和效率.

1 相关工作

1.1 同质信息网络的 IM 问题研究

由于利用贪心算法进行 IM 问题的求解需要进行大量的蒙特卡洛模拟, 时间复杂度过高. 为了能在接受的时间范围内求得一个可行解, 学者们提出了许多新的策略来提高运行的效率. 文献 [11] 提出了一种新的贪婪框架, 以 1/3 近似比有效的解决多重同质网络的 IM 问题, 但该算法需要消耗极大的内存资源. 文献 [6] 提出了一种局部两跳搜索算法, 该算法的核心是既接受单跳邻居节点的影响也接受两跳邻居的影响. 文献 [12]

基于多标准决策的元启发式方式来进行 IM 问题的研究. 文献 [13] 基于社区同质性, 考虑了社区中非活跃节点的影响, 提出了关于影响竞争的 IM 算法, 通过利用社区边界影响的两阶段种子节点选择算法来解决社区同质性引起的信息阻塞问题. 文献 [14] 通过以两种不同的竞争信息传播案例为例, 在竞争信息种子节点集合已知的前提下, 设计了一种自利益信息的解决方案, 提出了一种基于用户兴趣的节点回避影响 IM 算法. 然而这些算法都仅考虑单一的对象类型和关系类型, 并未考虑到实际生活中不同对象间的异质连接关系所带来的影响差异和损失.

1.2 异质信息网络的 IM 问题研究

在异质信息网络中, 元路径、元结构作为一种能够提取网络中对象之间异质关系和进行语义分析的手段, 已经被应用于许多的数据挖掘任务. 文献 [10] 基于元路径提取节点间的关系, 并利用信息熵来对异质信息网络中节点的影响力进行建模, 该算法通过构造多个同质子图, 并且以同质子图为基础对节点的影响力进行评估. 然而构造多个同质子图有着大量的时间花销, 并且基于同质子图进行节点影响的度量会导致原有网络中一些重要信息的损失. 文献 [15] 从不同的网络研究相同用户的影响力问题, 旨在从多个角度研究用户的影响. 文献 [16] 采用深度技术学习异质信息网络中相同类型节点间的联系特征, 用于评判节点的最终影响. 文献 [17] 提出了一种基于异质关系嵌入的多重聚合 MAHE-IM 的异质网络 IM 算法, 该算法能较好地捕获网络中的异质高阶结构和语义特征. 文献 [18] 通过探索相邻节点之间的交互性、用户的标签和兴趣相似性等特征来评估异质网络中节点的影响. 文献 [19] 为异质网络中的节点构造了一个元结构, 以存储节点在网络中的局部结构信息和异质信息, 提出了路径熵与结构熵度量节点的最终影响, 由于元结构能有效地捕获异质信息和局部结构信息, 本文通过在元结构的基础上对异质信息网络中的 IM 算法研究进行拓展, 旨在更全面的度量节点影响.

2 基本概念

2.1 相关定义

定义 1. 异质信息网络[20]. 异质信息网络是一个包含有: 对象映射函数$\partial: V \to T$和一个关系映射函数$\ell: E \to R$的有向图, 通常记为$G(V,T,E,R)$. 其中V是节点集合, T是节点类型集合, E是边集合, R是关系类型集合, 且$|T|+|R|>2$.

图 1 是一个异质信息网络示例, 它包含了作者、论文、会议等多种类型的节点, 从图 1 所示的网络中不仅可以提取到作者间的合作关系, 也可以提取出论文之间的引用关系.

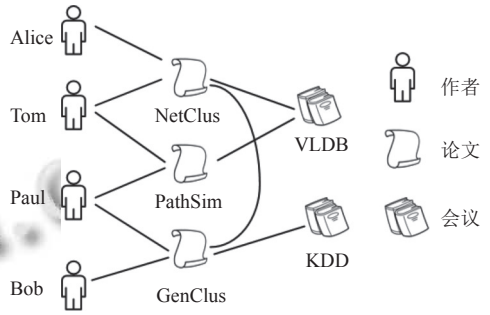

图 1 异质信息网络示例

定义 2. 网络模式[21]. 异质信息网络$G=(V,T,E,R)$的网络模式记为$T_G=(T,R)$, 是对G中所有对象类型和关系类型之间关联模式最直接的基础表达.

定义 3. 元路径、元结构[22]. 元路径是定义在异质信息网络模式$T_G=(T,R)$上的线性序列, 源对象与目标对象位于路径的两端, 可被表示为: $T_1T_2\cdots T_n$. 元结构是指定义在网络模式T_G上的一个有向无环图, 通常记为$M=(V_M,E_M)$, 其中V_M表示节点集, E_M表示边集.

定义 4. 信息熵[8]. 信息量是某个事物产生信息大小的度量指标, 而信息熵是对信息量的期望. 在 IM 问题研究中可以根据一个节点信息量的大小来度量节点的影响, 对于网络中任意节点v的信息熵可以由式 (1) 计算:

$$E_V=\sum_{u\in N_v}H_{uv}=\sum_{u\in N_v}-p_{uv}\log(p_{uv}) \tag{1}$$

其中, $p_{uv}=\dfrac{d_u}{\sum_{l\in N_v}d_l}$, N_v是节点v的邻居节点集合, d_v表示节点v的度, H_{uv}表示节点u对v的影响大小.

2.2 问题定义

对于给定的异质信息网络G及目标对象类型, 本文目的是利用网络中所包含的异质信息和结构信息来度量节点的影响力, 并选择出与目标对象类型具有相同类型的节点集合S^*, 使S^*在特定的扩散模型下具有最大影响范围. S^*定义如式 (2):

$$\sigma(S^*)=\underset{S\subseteq V,|S|=k,\forall v\in S,v\to T_i,T_i\in T}{\arg\max}\{\sigma(S)\} \tag{2}$$

其中, k表示集合大小, σ表示影响激活函数.

3 基于社区与结构熵的 IM 建模

文献 [19] 提出了一种基于元结构的异质网络 IM 算法 (meta-structure-based influence maximization, MSIM). MSIM 算法通过构造元结构捕获节点在网络中局部的结构信息, 并利用路径熵和结构熵衡量节点的影响作为种子节点的选择标准. 在 MSIM 中网络中, 节点v_1的路径熵$PE_{v_1}^W$的计算如式 (3), 结构熵ME_{v_1}的计算如式 (4):

$$PE_{v_1}^W=\sum H_w=\sum -p_{v_1v_2}\cdots p_{v_{n-1}v_n}\log\left(p_{v_1v_2}\cdots p_{v_{n-1}v_n}\right) \tag{3}$$

其中, w表示元路径W的路径实例, H_w表示路径实例w的信息熵大小.

$$ME_{v_1}=\sum_{W_i\in MS_{v_1}}\lambda_i PE_{v_1}^{W_i} \tag{4}$$

其中, λ_i表示元路径W_i的影响因子.

MSIM 算法为网络中的节点构造了一个元结构以保留节点在网络中的局部结构信息和异质信息, 并基于元结构度量节点的影响. 因此 MSIM 算法所求的节点影响是一种局部影响, 并且在 MSIM 算法中未考虑“富人俱乐部现象”给节点集扩散范围所造成的损失.

社区结构是网络中普遍存在的一种结构特征, 通过对网络进行社区划分能将连接紧密的节点划分为同一个社区, 并且在基于社区度量节点影响的同时, 适当地提升社区边缘节点的权重占比进行种子节点的选择能有效地缓解“富人俱乐部”所带来的扩散损失. 基于这个观察, 本文在 MSIM 算法的基础上提出了异质信息网络中基于社区与结构熵的影响力最大化算法 (CMIM). CMIM 算法从两个方面度量节点的影响. 一方面通过 MSIM 算法计算节点的结构熵作为节点的局部影响力的评判依据. 另一方面通过社区发现算法将异质信息网络中的相同类型的节点划分为不同的重叠社区, 并计算节点所属社区在整个网络中的权重来度量节点的全局影响, 通过提升社区边缘节点的权重占比, 以减少“富人俱乐部现象”所带来的损失. 最后综合局部影响和全局影响求出节点的最终影响并选择种子节点.

3.1 节点的局部影响度量

网络中节点的局部结构指“在网络拓扑结构中节点处存在的邻域集合”. 基于局部结构度量节点的影响, 能有效地提升信息扩散初始阶段种子集合的扩散范围. CMIM 算法以 MSIM 算法为基础, 为网络中目标类型节点构造一个元结构, 以保留节点在网络中局部邻域的异质连接信息和结构信息. 然后依据节点的元结构计算节点的路径熵与结构熵, 并且将最终结构熵的计算结果作为节点的局部影响. 如式 (5):

$$LI_v=ME_v \tag{5}$$

其中, LI_v指节点的局部影响力.

3.2 节点的全局影响度量

节点全局影响是指节点在整个网络拓扑结构中影响作用的大小. 由于社区结构是网络普遍存在的一种结构特征, 在信息传播的过程中起着重要的作用, 本文通过度量节点所属社区在整个网络中的权重占比, 以度量节点的全局影响力. 考虑到社区发现算法高昂的时间复杂度, CMIM 算法选择了具有线性时间复杂度的基于节点的拓扑势[23] 的重叠社区划分算法, 然后根据划分好的社区结构度量节点的全局影响.

网络中大部分节点都不是孤立存在的, 节点之间往往存在着不同的连接关系, 而节点的拓扑势可以描述一个节点与其邻居的相互作用的大小. 网络中节点v的拓扑势定义为$\varphi(v)$, 计算如式 (6):

$$\varphi(v)=\sum_{j=1}^{n}\left(m(j)\times \mathrm{e}^{-\left(\frac{d_{vj}}{\delta}\right)^2}\right) \tag{6}$$

其中, $m(j)$表示节点的质量, d_{vj}表示节点v与节点j之间的距离. δ是优化因子, 用于控制节点之间相互影响的距离, 其计算如式 (7):

$$\delta_{opt}=\arg\min H(\delta) \tag{7}$$

其中, $H(\delta)=-\sum_{i=1}^{n}\frac{\varphi_v(\delta)}{Z}\log\frac{\varphi_v(\delta)}{Z}$表示势熵, $Z=\sum_{i=1}^{n}\varphi(v)$表示标准化因子, $\varphi_v(\delta)$表示优化因子大小为δ节点v的拓扑势.

在式 (6) 中, 质量m是描述节点在网络中所具有的固有特性, 在网络中质量的概念有着许多的物理意义, 例如质量可以描述一个节点在网络中的重要性, 也可以描述一个节点在网络中所具有的信息含量的大小. 目前基于拓扑势的大部分研究都忽略了节点之间质量的差异, 并且将节点之间的质量都设置为统一大小的

值, 这会造成节点拓扑势计算准确性的损失, 并且会降低重叠社区发现算法的准确性. 由于在节点结构熵的计算过程中考虑了节点的局部结构, 而且计算节点的结构熵是对节点的局部影响传播能力的一种度量方式, 因此本文在 CMIM 算法中计算了异质信息网络中每一个节点的结构熵, 并将所计算出的结构熵值作为节点的质量, 如式 (8):

$$m(v) = ME_v \tag{8}$$

在计算了每一个目标类型节点的拓扑势之后, 根据节点处于网络拓扑结构中的位置遍历整个网络, 找出其中具有高势值的极值点. 将这些极值点作为社区的中心点, 并且以极值点为中心, 基于网络的拓扑结构向周围扩散, 将小于社区中心点拓扑势值的节点纳入社区, 最终形成整个网络的重叠社区结构. 由于在异质信息网络中相同类型的节点之间不一定存在边使之直接相连, 故在进行异质信息网络重叠社区结构划分时, 采用元路径提取目标类型节点之间的相对位置关系进行社区的划分. 基于节点拓扑势的算法伪代码如算法 1.

算法 1. 基于节点拓扑势的社区划分算法

1) 输入: 异质信息网络$G=(V,E,T,R)$, 种子节点类型T_m, 元路径p
2) 输出: 社区结构$\{C_1,C_2,\cdots,C_n\}$
3) 初始化社区中心点集合$CentralityN$=[]以及社区结构字典C={};
4) 遍历网络G并计算节点的拓扑势;
5) 根据种子集节点类型和元路径p找寻网络中目标类型节点的拓扑势极大值点加入$CentralityN$, 并初始化社区C_v;
6) 基于元路径p将拓扑势小于社区C_v中边缘节点的目标类型节点加入C_v, 直至社区大小不再变化;
7) 返回最终的社区结构.

在将给定的异质信息网络G划分为n个社区$\{C_1, C_2,\cdots,C_n\}$之后, 本文基于社区计算节点所属社区的权重以度量节点的全局影响. 对于同时属于多个社区的枢纽节点, 其全局影响会随着所在社区数量的增加而增大, 减少了“富人俱乐部现象”所带来的损失. 社区权重计算如式 (9), 节点全局影响力的计算如式 (10).

$$W_{C_i} = \frac{N_{C_i}}{N} \tag{9}$$

其中, N_{C_i}表示社区C_i的规模大小, N表示社区结构总体规模大小.

$$GI_v = \sum_{C_i}^{C} W_{C_i} \tag{10}$$

其中, GI_v表示节点v的全局影响力.

3.3 节点的最终影响度量

在计算了异质信息网络的中目标类型节点的局部影响与全局影响之后, 本文构造评判矩阵$\boldsymbol{A}$计算节点最终影响的大小.

$$\boldsymbol{A} = \begin{bmatrix} LI_{v_1} & GI_{v_1} \\ \vdots & \vdots \\ LI_{v_n} & GI_{v_n} \end{bmatrix} \tag{11}$$

为了减少计量单位的不统一而造成的误差, 首先对$\boldsymbol{A}$中的指标进行标准化处理, 由式 (12) 得到标准化处理后的矩阵$\boldsymbol{A}'$.

$$\boldsymbol{A}' = \begin{bmatrix} LI'_{v_1} & GI'_{v_1} \\ \vdots & \vdots \\ LI'_{v_n} & GI'_{v_n} \end{bmatrix} \tag{12}$$

其中, $LI'_{v_i} = \dfrac{LI_{v_i}}{\sum_{j}^{n} LI_{v_j}}$, $GI'_{v_i} = \dfrac{GI_{v_i}}{\sum_{j}^{n} GI_{v_j}}$.

在得到了标准化的矩阵$\boldsymbol{A}'$后, CMIM 算法考虑了局部影响和全局影响对节点最终影响扩散范围的影响占比, 设置了参数α, β对节点的LI'、GI'值进行约束, 节点的最终影响计算见式 (13).

$$UI_{v_i} = \alpha LI'_{v_i} + \beta GI'_{v_i} \tag{13}$$

其中, UI_{v_i}指节点v_i的最终影响. CMIM 算法伪代码如算法 2.

算法 2. CMIM 算法

1) 输入: 异质信息网络$G=(V,E,T,R)$, 种子节点类型T_m, 社区结构$\{C_1,C_2,\cdots,C_n\}$
2) 输出: 种子集合S
3) 初始化种子集合S;
4) 遍历网络G并计算目标类型节点的局部影响和全局影响;
5) 初始化评判矩阵$\boldsymbol{A}$并进行标准化处理得到$\boldsymbol{A}'$;
6) 根据$\boldsymbol{A}'$计算节点的最终影响力, 并选择最大的前k个节点加入到S中;
7) 返回种子集合S.

4 实验结果与分析

4.1 实验准备

数据集: 本文将在两个真实的异质信息网络数据集 (DBLP 数据集和 YELP 数据集)[19] 上来验证 CMIM 算法的性能. 其中 DBLP 数据集中包含了 4 种对象类型和 3 种关系类型, YELP 数据集中包含了 4 种对象类

型和 4 种关系类型.

对比算法: 为了验证本文所提算法的有效性, 本文采用以下方法作为 CMIM 算法的对比算法: (a) VoteRank 算法[24] 为网络中每个节点都设置了一个投票能力, 通过迭代选取网络中票数最多的节点作为种子节点, 在每次迭代开始时被选中的节点不再具有投票权, 并衰减其邻居的投票能力以减少富人俱乐部的影响. (b) MPIE 算法[10] 通过不同的元路径从异质信息网络中提取多个同质子图, 然后利用链接熵和相互熵来衡量不同同质子图中节点影响, 最终整合为节点的最终影响. (c) Entropy 算法[9] 是异质信息网络中一种关于熵的启发式算法, 通过限制相同类型节点间的最短路径从异质信息网络中抽取一个同质子图, 基于子图计算节点的熵值, 取熵值高的节点为种子节点. (d) MSIM 算法[19] 通过为节点构造元结构, 并基于元结构计算节点的路径熵和结构熵以确定节点的影响.

度量标准: 在 IM 问题的研究中, 影响范围 (influence spread) 是一种被广泛应用于评价算法性能好坏的评价指标, 定义为在信息扩散过程结束之后, 种子集合成功激活网络中节点的数量. 扩散范围 (influence spread) 值越大, 说明算法效果越好.

扩散模型: 本文采用线性阈值模型 (linear threshold, LT) 作为扩散模型. LT 模型规定了网络中的节点都存在一个激活阈值, 若非激活节点的已激活邻居对其的影响总和达到激活阈值, 则该节点被激活.

4.2 实验结果

4.2.1 有效性验证

为了验证 CMIM 算法的性能, 本文在 DBLP、YELP 数据集上进行了信息扩散模拟实验, 分别在 DBLP、YELP 数据集中选择k个作者节点和k个用户节点作为种子集. 其中, k分别取 (10, 20, 30, 40, 50). 结果如图 2 所示.

通过分析图 2 可以得出, 不同算法在不同数据集上表现出不同的性能, 且随着种子集合规模的增长, 扩散范围也会随之增大.

较于 MSIM 算法, CMIM 算法在两个数据上性能都有所提升, 这说明适当的考虑社区结构有利于种子集最终扩散范围的提升. VoteRank 算法通过迭代衰减已被选入种子集合的节点及其邻居的影响来减小网络中“富人俱乐部现象”带来的扩散损失. 由于在 DBLP 数据集中作者节点之间没有直接相连的边, 故 VoteRank 算法性能表现较差, 正也因为 YELP 数据集中存在着用户关系的同质子图, 其算法性能在 YELP 数据上有所提升. 这说明在异质信息网络中通过衰减种子节点及其邻居影响的方法来减少“富人俱乐部现象”造成损失的方法已不适用. 且从结果中可以看出较于 VoteRank 算法, 本文所提出的 CMIM 算法在性能上有着明显的提升. 这表明在异质信息网络中基于重叠社区的思想来减少“富人俱乐部现象”带来扩散损失的方法是有效的. 较于 MSIM、MPIE、Entropy 算法, CMIM 算法在两个数据集上的扩散范围均有所提升, 这说明考虑节点在整个网络中全局影响能够更有效地度量节点的影响力. 相较于 DBLP 数据集, CMIM 算法在 YELP 数据集上的提升并不明显, 这也说明了基于重叠社区的影响力度量方法比较依赖于数据集的结构特征. 但从 DBLP、YELP 数据集上的扩散结果显示 CMIM 算法最终的影响扩散范围总体较优, 这证明在异质信息网络中结合节点的局部影响和全局影响更有利于精确的度量节点的影响力.

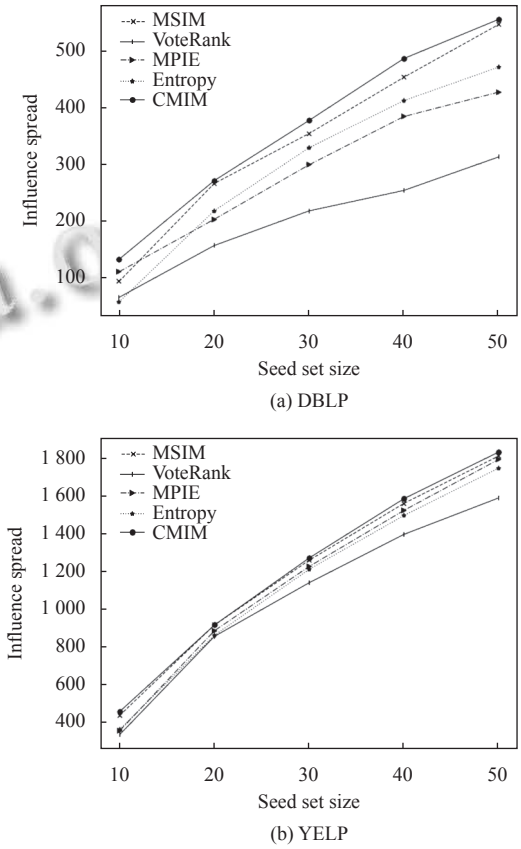

图 2 算法扩散范围比较

4.2.2 时间效率分析

为了对比 CMIM 算法的时间效率, 本文在 DBLP 和 YELP 数据集上统计了 CMIM 算法与对比算法在选取 50 个种子节点的时间花销, 结果如表 1 所示, 其中 CMIM-C 算法是统计 CMIM 算法在进行好重叠社区结构划分之后选择种子节点集合的运行时间.

表 1 算法运行时间对比 (s)

算法	DBLP	YELP
CMIM	35.45	111.64
CMIM-C	0.63	0.78
MSIM	1.41	23.42
VoteRank	4.37	4.46
Entropy	8.17	107.04
MPIE	64 800.43	113 400.52

从表 1 中可以分析出, 随着数据集的复杂程度的增加, 所有算法的运行时间均有所增加. 其中, VoteRank 算法运行时间最为稳定, 但结合图 2 中的结果可以得出它的性能在异质信息网络中相对较差. 而 CMIM、MSIM、Entropy 和 MPIE 算法的在两个数据集上增长幅度较大, 且这 4 类算法都属于异质信息网络中的 IM 算法. 这表明从异质信息中提取异质信息依赖网络的网络模式和规模大小, 且时间花销较大. MPIE 算法和 Entropy 算法的时间花销主要集中在构造同质子图. 从 CMIM 算法和 CMIM-C 算法的运行时间对比可以看出, CMIM 算法的运行时间主要花费在重叠社区构造的过程上, 节点的全局影响力计算的时间花费较小. 结合图 2 所示的算法对比结果, 可以得出结论: CMIM 算法是在 MSIM 算法的基础上, 以更多的时间花销来换取更大的影响扩散范围, 但总体的时间花销在可接受的范围之内.

4.2.3 节点质量的影响

在描述 CMIM 算法的过程中, 选择采用节点的结构熵值 (ME) 作为节点的质量. 本文通过求节点的度 (Degree)、信息熵值 (I-Entropy)、PR 值以及将节点质量统一置为 1 (Unified-1) 等多种方式作为对比, 来比较在不同方法下求得的节点的质量对最终种子集合的扩散范围的影响. 在实验过程中保留节点局部影响力和全局影响力的计算方式, 比较最终选择出的k个种子节点的信息扩散范围, 其中k取 (20, 50). 实验结果如图 3 所示.

从图 3 中可以看出不同的质量度量方式会对扩散结果产生不同的影响, 当选择用节点的结构熵值作为节点质量的度量方式时, 最终扩散范围均优于其他度量方式. 这表明了使用结构熵作为节点质量的度量方式能够更有利于最终种子集合的影响扩散.

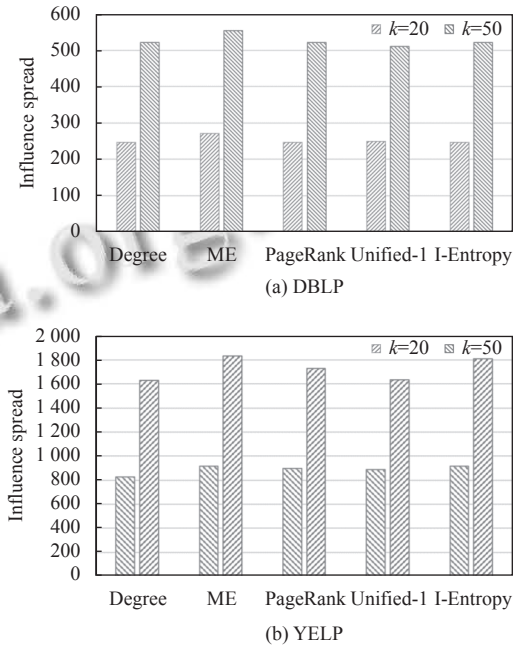

图 3 不同节点质量度量方式的影响

4.2.4 参数权重的影响

在 CMIM 算法计算节点的最终影响力时, 设置了参数α, β对节点的局部影响力和全局影响力进行了约束. 因此本文测试了多组α和β的线性组合对种子集和最终扩散范围的影响, 其实验结果如图 4 所示, 其中横坐标表示α和β的大小, 纵坐标表示大小为 50 时种子集合的扩散范围.

从图 4 中可以看出不同α和β的线性组合对最终扩散产生了不同的影响, 这说明在度量节点影响力时, 考虑全局影响和局部影响的不同比重会对最终扩散范围产生不同的影响. 而当α和β的值分别为 0.6, 0.4 时, 在 DBLP 数据集中达到了最优. 当α和β的值分别为 0.7, 0.3 时, 在 YELP 数据集中达到了最优, 这说明了相同的权重组合在不同的数据集中的性能表现不同. 而且从这两组参数可以看出较于全局结构, 从局部结构对节点的影响力进行度量更为重要. 因此, 在进行有效性分析实验时, 在 DBLP 数据集中选用的参数组合为 0.6, 0.4, 在 YELP 数据集中选用的参数组合为 0.7, 0.3.

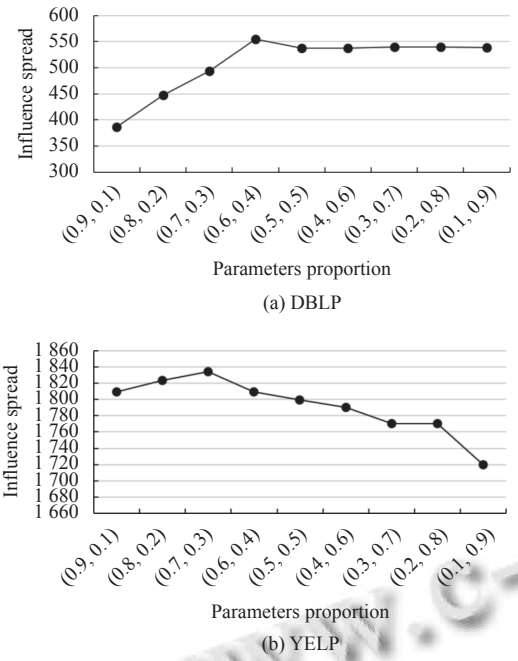

图 4 参数权重的影响

5 结束语

针对现有的异质信息网络中的 IM 算法未考虑网络的中“富人俱乐部现象”或节点的全局影响, 导致最终节点影响力的度量往往不够精确的不足. 本文充分考虑了节点在异质信息网络中的位置关系, 分析了节点在整个网络拓扑结构中的局部与全局影响, 提出了异质信息网络中基于社区与结构熵的影响力最大化算法. 在真实数据集合上进行了大量实验证明了所提算法的有效性, 并分析了所提算法中相关参数的影响. 在未来的研究工作中, 可以在本文的基础上考虑实现对异质信息网络中含有不同对象类型的节点集合算法的研究, 尝试以深度学习、注意力机制等方法实现相关参数的自动化配置与构造.

参考文献

1 Domingos P, Richardson M. Mining the network value of customers. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2001. 57–66.

2 Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC: ACM, 2003. 137–146.

3 单菁, 申德荣, 寇月, 等. 基于重叠社区搜索的传播热点选择方法. 软件学报, 2017, 28(2): 326–340. [doi: 10.13328/j.cnki.jos.005117]

4 Wang ZX, Sun CC, Xi JK, *et al*. Influence maximization in social graphs based on community structure and node coverage gain. Future Generation Computer Systems, 2021, 118: 327–338. [doi: 10.1016/j.future.2021.01.025]

5 Liu XY, Wu SY, Liu C, *et al*. Social network node influence maximization method combined with degree discount and local node optimization. Social Network Analysis and Mining, 2021, 11(1): 31. [doi: 10.1007/s13278-021-00733-3]

6 Qiu LQ, Yang ZQ, Zhu SW, *et al*. LTHS: A heuristic algorithm based on local two-hop search strategy for influence maximization in social networks. Journal of Intelligent & Fuzzy Systems, 2021, 41(2): 3161–3172.

7 Nie TY, Guo Z, Zhao K, *et al*. Using mapping entropy to identify node centrality in complex networks. Physica A: Statistical Mechanics and Its Applications, 2016, 453: 290–297. [doi: 10.1016/j.physa.2016.02.009]

8 Antelmi A, Cordasco G, Spagnuolo C, *et al*. Social influence maximization in hypergraphs. Entropy, 2021, 23(7): 796. [doi: 10.3390/e23070796]

9 Li CT, Lin SD, Shan MK. Influence propagation and maximization for heterogeneous social networks. Proceedings of the 21st International Conference on World Wide Web. Lyon: ACM, 2012. 559–560.

10 Yang YD, Zhou LH, Jin Z, *et al*. Meta path-based information entropy for modeling social influence in heterogeneous information networks. Proceedings of the 20th IEEE International Conference on Mobile Data Management (MDM). Hong Kong: IEEE, 2019. 557–562.

11 Wu GH, Gao XF, Yan G, *et al*. Parallel greedy algorithm to multiple influence maximization in social network. ACM Transactions on Knowledge Discovery from Data, 2021, 15(3): 43.

12 Biswas TK, Abbasi A, Chakrabortty RK. An MCDM integrated adaptive simulated annealing approach for influence maximization in social networks. Information Sciences, 2021, 556: 27–48. [doi: 10.1016/j.ins.2020.12.048]

13 Xie XQ, Li JH, Sheng Y, *et al*. Competitive influence maximization considering inactive nodes and community homophily. Knowledge-based Systems, 2021, 233: 107497. [doi: 10.1016/j.knosys.2021.107497]

14 Tong J, Shi LL, Liu L, *et al*. A novel influence maximization algorithm for a competitive environment based on social

media data analytics. Big Data Mining and Analytics, 2022, 5(2): 130–139. [doi: 10.26599/BDMA.2021.9020024]

15 Kuhnle A, Alim MA, Li X, *et al*. Multiplex influence maximization in online social networks with heterogeneous diffusion models. IEEE Transactions on Computational Social Systems, 2018, 5(2): 418–429. [doi: 10.1109/TCSS.2018.2813262]

16 Keikha MM, Rahgozar M, Asadpour M, *et al*. Influence maximization across heterogeneous interconnected networks based on deep learning. Expert Systems with Applications, 2020, 140: 112905. [doi: 10.1016/j.eswa.2019.112905]

17 Li Y, Li LL, Liu YJ, *et al*. MAHE-IM: Multiple aggregation of heterogeneous relation embedding for influence maximization on heterogeneous information network. Expert Systems with Applications, 2022, 202: 117289. [doi: 10.1016/j.eswa.2022.117289]

18 Deng XH, Long F, Li B, *et al*. An influence model based on heterogeneous online social network for influence maximization. IEEE Transactions on Network Science and Engineering, 2020, 7(2): 737–749. [doi: 10.1109/TNSE.2019.2920371]

19 徐智敏, 周丽华. 异质信息网络中基于元结构的影响力最大化. 计算机仿真, 2021.

20 刘佳玮, 石川, 杨成, 等. 基于异质信息网络的推荐系统研究综述. 信息安全学报, 2021, 6(5): 1–16. [doi: 10.19363/J.cnki.cn10-1380/tn.2021.09.01]

21 Sun YZ, Han JW. Mining heterogeneous information networks: A structural analysis approach. ACM SIGKDD Explorations Newsletter, 2012, 14(2): 20–28.

22 Sun YZ, Han JW, Yan XF, *et al*. PathSim: Meta path-based top-K similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment, 2011, 4(11): 992–1003. [doi: 10.14778/3402707.3402736]

23 Liu YS, Ye XF, Yu CY, *et al*. TPSC: A module detection method based on topology potential and spectral clustering in weighted networks and its application in gene co-expression module discovery. BMC Bioinformatics, 2021, 22(S4): 111. [doi: 10.1186/s12859-021-03964-5]

24 Zhang JX, Chen DB, Dong Q, *et al*. Identifying a set of influential spreaders in complex networks. Scientific Reports, 2016, 6: 27823. [doi: 10.1038/srep27823]

(校对责编: 孙君艳)