

竞争环境中基于主题偏好的利己信息影响力 最大化算法

曹玖新^{1),2)} 闵绘宇¹⁾ 王浩然³⁾ 马卓²⁾ 刘波¹⁾

¹⁾(东南大学计算机科学与工程学院 南京 211189)

²⁾(东南大学网络空间安全学院 南京 211189)

³⁾(东南大学软件学院 南京 211189)

摘 要 社会计算领域的影响力传播问题一般仅研究单条信息的最大化传播,但在实际网络环境中,更多的是存在多条具有竞争关系的信息在网络中进行传播,且在传播过程中不同信息会相互影响.竞争环境中的影响力最大化问题即是为某一条竞争信息选择种子节点集合,使之最终影响的节点数目最多.该文针对多条相似信息的竞争传播问题,考虑传播的同步性,建立基于竞争的线性阈值扩展模型;基于用户交互的主题偏好计算不同类别信息下节点间的影响概率,并结合扩展的传播模型和信息扩散的特点,提出基于节点子图的影响力计算方法.以两条竞争信息传播为例,考虑利己信息对竞争信息传播策略的掌握情况,设计两种解决方案:(1)在已知竞争信息A的种子节点选择策略时,提出节点避让的利己信息影响最大化算法NA(Node Avoidance);(2)在未知竞争信息A的种子节点选择策略时,提出使用种子节点选择策略集为信息A选择种子集合,并提出策略无关的利己信息影响最大化算法I3SL(Independent Seed Selection Strategy of Leader).该文在新浪实证数据集上进行了对比实验,结果表明:(1)NA算法在竞争环境中能使利己信息传播范围最广,平均优于其他算法26.2%,且表现出较好的通用性和稳定性;(2)I3SL算法在未知竞争信息的选择策略时也能够保证利己信息的传播范围,平均优于其他算法18.23%;(3)该文所提两种算法在运行时间上较其他启发式算法表现较弱,但两种算法在运行时间和传播范围两方面均取得了较好的平衡.

关键词 竞争环境;影响最大化;线性阈值模型;主题偏好;节点子图

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2019.01495

Self-Interest Influence Maximization Algorithm Based on Subject Preference in Competitive Environment

CAO Jiu-Xin^{1),2)} MIN Hui-Yu¹⁾ WANG Hao-Ran³⁾ MA Zhuo²⁾ LIU Bo¹⁾

¹⁾(School of Computer Science and Engineering, Southeast University, Nanjing 211189)

²⁾(School of Cyber Science and Engineering, Southeast University, Nanjing 211189)

³⁾(College of Software Engineering, Southeast University, Nanjing 211189)

Abstract Traditional influence maximization generally involved the situation of one message spread in propagation model. However, in the actual network, multiple competitive messages are spread synchronously, and influence each other alternately. How to choose the initial set of seeds for each competitive message so as to get the max number of influenced nodes in a competitive environment is an NP-hard problem. To solve the problem happened during simultaneous

收稿日期:2017-12-19;在线出版日期:2018-11-22. 本课题得到国家自然科学基金项目(61772133,61472081)、江苏省重点研发项目(BE2018706)、江苏省计算机网络技术重点实验室、江苏省网络与信息安全重点实验室(BM2003201)、东南大学计算机网络和信息集成教育部重点实验室(93k-9)资助. 曹玖新,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为社会计算、服务计算、网络安全. E-mail: jx.cao@seu.edu.cn. 闵绘宇,博士研究生,主要研究方向为社会计算、复杂网络、Multi-Agent 系统. 王浩然,硕士研究生,主要研究方向为社会网络. 马卓,博士研究生,主要研究方向为复杂网络、社会计算. 刘波,博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为普适计算、社会计算.

competitive dissemination, our work considers the synchronization of competitive information dissemination, as well as the law that user's influences of different information are not the same. Few researchers analyze the content of competitive information on the influence of competition. But in fact, users' topic preferences and interactive preferences will be mined through their interactive content. So we allow for the topic preference of interaction between users, then take the advantage of topic model to get the probability between users in the different kinds of information, respectively. Based on that, we define the competition-based linear threshold model to fit the characteristic of information dissemination in this simultaneous competitive dissemination, and followed by the method proposed to calculate the node's influence based on node subgraph. To take an example from the dissemination of two competitive information, we take the mastery of competitors' dissemination strategies into account, and provide two solutions: (1) The Node Avoidance Influence Maximization Algorithm (NA) suppose the competitors' dissemination strategies are known to each other, which would avoid selecting nodes that can maximize the competitors' information diffusion. (2) The Independent Seed Selection Strategy of Leader Influence Maximization Algorithm (I3SL) we proposed provides a novel solution for the case that the competitors' dissemination strategies are unknown. Experiments are carried out to verify the effectiveness of algorithms on real Weibo data set, and the results under the competition based on linear threshold model show that: (1) In the competitive environment, NA algorithm is superior to the traditional heuristic algorithms in the scope of influence, averaging 26.2% better than other algorithms, with a good performance in universality and stability; (2) The I3SL algorithm can also guarantee the scope of influence when the selection strategy of competitive information is unknown, averaging 18.23% better than other algorithms. (3) As for the running time, our two algorithms are inferior to the heuristic algorithm, but they still have a better balance between efficiency and scope of influence.

Keywords competitive environment; influence maximization; linear threshold model; topic preference; subgraph of nodes

1 引言

近年来,社交网络中信息传播范围广、速度快、用户和信息数量庞大等特点,使其逐渐衍生成为巨大的产品营销平台,并基于用户交互模式形成了新的营销方式——病毒式营销。病毒式营销以口头传播为基础,其关键是寻找社交网络中影响力较大的用户集合作为信息传播的种子节点集,并通过传播模型影响其他节点,最终使得该集合能够最大化影响其他用户,这便是影响力最大化问题的原型。

Domingos 等人^[1]最先从算法角度抽象影响力最大化问题。Kempe 等人^[2]首次将该问题定义为一个离散最优化问题,并证明该问题是一个 NP-hard 问题,同时提出一种贪心爬山算法得到该问题的近似解。但贪心算法的时间复杂度较高,无法高效解决大规模社交网络下的问题。对此,研究者相继

提出了多种启发式算法^[3-11]以提高运行效率。曹玖新等人^[3]基于 k -核概念提出核覆盖算法(Core Covering Algorithm, CCA)。李阅志等人^[4]结合现有启发式算法提出基于 k -核过滤的影响最大化算法。马茜等人^[5]考虑实际情况中节点可能受各种影响而无法激活,提出一种选择替补节点的算法。

然而上述研究均仅考虑单条信息在网络中的传播,而在现实网络中,随处可见多条竞争信息同时在社交网络中传播的情形。面对网络中鱼龙混杂的广告信息,用户往往会综合考虑发布信息的商家口碑、好友推荐、个人消费能力和偏好等,最终选择唯一的产品进行消费。例如苹果和小米相继推出新款手机,阿迪和安踏同出新款单品,美团和糯米同出折扣机票等等。在上述环境中,如何借助社交网络为利己广告获得更大的宣传效果是至关重要的。因此,对竞争环境中利己信息影响力最大化问题的研究更具现实意义。

2007 年 Bharathi 等人^[12]首次提出存在竞争的影响力最大化问题,近年来对该问题的研究仍持续升温^[13-27],主要表现为研究者对该问题进行了拓展和细化,提出竞争环境中利己信息影响力最大化问题^[13-16]、舆论谣言传播问题^[17-19]、竞争信息影响力阻碍最大化问题^[20-22]、合作影响力最大化问题^[23-24]等。

当复杂多样的竞争信息散布在社交网络,很容易导致用户接收到大量冗余信息进而产生反感心理,产品商的广告收益也因此而大打折扣。文献^[28]表明,信息主题偏好的不同会影响其在同一网络中的影响范围。因此,在信息传播模型中充分考虑用户主题偏好,会更贴合实际社交网络中信息传播模式,且能较为精确地分析用户间相互影响的程度。

针对上述问题,本文的目标是从利己信息的角度研究其在竞争环境中的影响力最大化问题。其贡献包含如下三个方面:(1)考虑传播的同步性,设计基于传统线性阈值传播模型扩展模型。通过对用户文本及好友间的交互信息进行分析,挖掘好友间在不同主题下的影响概率,结合扩展的线性阈值模型和信息扩散的特点,提出基于节点子图的影响力计算方法;(2)当已知竞争信息的广告投放策略时,利己信息的影响力最大化问题便演变为:已知竞争信息的种子节点集合,如何寻找利己信息的种子节点集合使之影响范围最广。对此,提出节点避让的利己信息影响最大化算法 NA(Node Avoidance);(3)当未知竞争信息的广告投放策略时,竞争环境中的影响力最大化问题将不再具有单调子模特性^[24]。基于竞争信息种子节点选择策略集合,本文提出策略无关的利己信息影响最大化算法 I3SL(Independent Seed Selection Strategy of Leader)。

本文第 2 节对竞争环境中的影响力最大化问题的研究现状进行总结与分析;第 3 节给出本文的问题定义及扩展的传播模型;第 4 节描述基于主题偏好的节点影响概率建模方法;第 5 节给出基于子图的节点影响力计算方法;第 6 节根据对竞争信息节点选择策略的掌握情况,分别提出 NA 算法和 I3SL 算法;第 7 节对实验数据集、实验设计及评价标准做描述;第 8 节对实验结果进行对比与分析;最后对本文进行总结。

2 相关工作

研究者从 2003 年首次研究单条信息的影响力

最大化问题^[1],到 2007 年首次研究竞争环境中的影响力最大化问题^[12],再到 2015—2017 年更加细化地从多种角度研究多条竞争信息的影响力最大化问题,例如从社区角度^[14]、从舆情角度^[18,19]、从广告成本角度^[20]等,竞争环境中影响力最大化问题的研究持续升温。

2.1 竞争环境中的影响力最大化问题

竞争环境中的影响力最大化问题一般可分为两种解决方式:其一从利己角度出发,寻找影响最大的种子节点集;其二从抑制竞争者的角度,当多个竞争信息传播时,使竞争对手的影响力最小化。Bharathi 等人^[12]最早提出并证明该问题可以用贪心算法来找到近似最优解。Bozorgi 等人^[14]利用社区划分算法划分网络,并计算每个节点在所属社区中的影响力,提出 CI2 算法解决该问题,并提出基于 LT 模型扩展的 DCM 传播模型。Lim 等人^[19]证明当两个竞争公司在社交网络中同时散布谣言时,使用级联中心性可以明确表征纯粹策略纳什均衡,并基于线性阈值模型模拟谣言传播。Zhu 等人^[20]提出研究竞争环境中最小化种子节点的投放成本问题,并同时考虑抑制竞争对手的影响力传播问题。Ribeiro 等人^[22]研究了竞争信息延迟传播中,新信息如何通过竞争关系影响已被其他信息激活的用户。Zhang 等人^[24]研究了多条信息传播的情况,将信息间的关系分为相互独立、依赖、竞争三种类型,并提出动态阈值传播模型,根据信息策略定制的同步和异步性,分别对利己信息的影响力最大化问题进行了研究。

同时,也有较多文献对竞争环境中的传播模型进行研究。Tzoumas 等人^[29]证明了竞争环境中的影响力最大化问题的存在性和可计算性等,并以线性阈值模型为基础对该问题进行研究,实验发现当网络中存在多条信息同时传播时,信息之间会存在相互竞争,并且可以通过进一步延伸和扩展传统的线性阈值模型来解决该问题。Borodin 等人^[15]以多种形式扩展线性阈值模型来研究基于竞争的影响力传播问题,分别提出:基于竞争的比重线性阈值模型、基于竞争的阈值分离模型、基于竞争的强制阈值模型等,并证明该问题在线性阈值扩展模型中是 NP-hard 问题。Beutel 等人^[26]研究了多条信息之间的关系,提出利用两种信息间的相关系数研究它们在同步传播过程中相互作用的程度。有研究者基于传统单条信息的传播模型进行了扩展,提出了多活动独立级联模型(Multi-Campaign Independent Cascade Model)^[17]、抑制传播模型(Diffusion-Containment

Model)^[21]、存在竞争的线性阈值模型(Competitive Liner Threshold Model, CLT)^[27]等. 上述文献针对不同的问题情境对传播模型进行了合理扩展,但对节点影响概率的计算考虑较为简单,大都从网络或社区结构出发,较少考虑用户交互的内容和语义信息.

通过对上述文献研究问题、传播模型、算法及数据集等方面的分析发现,文献[19]、[20]、[26]与本文研究目标存在较大出入,文献[14]所采用的传播机制与本文差异较大,而文献[24]中所用算法为贪心策略,仅适用小规模数据集.

2.2 基于主题模型的影响力最大化研究

社交网络中用户对某一主题的偏好程度往往会决定用户接受该主题信息的程度,但现有研究竞争环境中的影响力最大化问题较少考虑主题因素对信息传播的影响. Barbieri 等人^[30]证明用户主题偏好可以影响用户的社交地位、权威和信任,因此提出基于主题的影响力级联模型(Topic based Independent Cascade, TIC),在此基础上进一步提出一种基于相似性的方法来研究基于主题的影响力最大化问题. Chen 等人^[31]提出线上线下分离策略来解决基于主题的影响力传播问题,其思想是在离线环节为每个主题筛选出备用节点集,然后在在线环节从备用节点集中选取种子节点,结果显示基于主题的信息传播模型的影响质量和影响效果均优于未考虑主题因素的信息传播模型.

综上所述,在现有研究中,同时考虑竞争条件和基于用户主题偏好的影响力最大化问题的研究工作甚少.然而在现实社会中,面对社交网络中普遍存在的具有竞争关系的信息,用户好友偏好的不同、用户主题偏好的不同、信息发布时间先后的不同等均会影响用户接受该信息的程度.例如,假设有两家餐厅 A、B 同在做线上推广,其中 A 是老牌餐饮店,口碑较高,以招牌川菜、环境优美吸引食客,而 B 为餐饮新店,以创新川味、新品折扣吸引食客,可以推测对于无辣不欢但介意价格的用户,选择 B 的可能性较大,但对既爱吃辣又介意就餐环境的用户,则会有较大可能选择 A. 在影响力最大化问题中,同时考虑竞争条件和用户主题偏好,一方面更贴合现实社交网络环境,另一方面能更好的模拟实际信息传播的过程.

3 问题定义及传播模型

本节在线性阈值扩展模型下对问题进行了定义和描述. 本文采用的数学符号及含义如表 1 所示.

表 1 符号说明

符号	含义
$G(V, E)$	社交网络有向图
θ_v^B	传播模型中节点 v 变成 B 激活态的阈值
φ_v	子图 $G^B(v)$ 生成时设置的阈值
$G^B(v)$	在信息 B 下, 节点 v 生成的子图
$p_{uv}^{'B}$	在信息 B 下, 节点 v 被邻居节点 u 影响的概率
p_{uv}^B	在信息 B 下, 节点 v 被邻居节点 u 影响归一化后的概率
$p_v^B(u)$	在信息 B 下, 在节点子图 $G^B(v)$ 中节点 u 通过所有可达路径影响根节点 v 的概率
$\Delta Inf^B(u, v)$	在信息 B 传播过程中, 子图 $G^B(v)$ 中其他节点 u 对根节点 v 的影响力增量
$Inf^B(v)$	在信息 B 下, 节点 v 在 $G(V, E)$ 上的影响力
$N^{in}(v)$	节点 v 所关注的节点集合, 即节点 v 的入度
$N^{out}(v)$	关注节点 v 的节点集合, 即节点 v 的出度
$\alpha p^B(v)$	节点 v 被信息 B 激活的概率

3.1 问题定义

“物竞天择, 适者生存”的竞争法则在社会领域被世人膜拜, 同样, 社交网络中具有利益关系的商家间也正积极寻找其在网络社会中的生存之道. 由于用户偏好、商家口碑等的不同, 对社交网络中的用户而言, 商家间存在着不对称的竞争环境. 如何在地位不对称的社交网络中获得利己信息的影响最大化, 是本文的研究目标. 由于多条竞争信息同时传播问题可以简化为两条竞争信息同步传播问题, 因此本文以两条竞争信息传播问题为例, 研究利己信息在竞争环境中的影响力最大化问题. 在综合考虑竞争信息同步传播过程中竞争商家间对彼此广告投放策略掌握程度的不同, 分别提出两种情况下的问题解决方法: (1) 当已知竞争商家的广告投放策略时, 可将问题定义为: 已知竞争信息 A 的广告投放策略, 即已知信息 A 的种子节点集合, 为利己信息 B 寻找种子节点集合, 使之在网络中的影响范围最大; (2) 当未知竞争商家的广告投放策略时, 竞争环境中的影响力最大化问题不具有单调子模特性, 难以确定该问题的最优解^[24]. 可将问题定义为: 给定一个种子节点选择策略集合, 寻找一种与竞争信息 A 的种子节点选择策略无关的利己信息影响最大化算法, 使之能确保利己信息 B 的影响范围最大.

根据研究问题, 定义本文传播模型如下: 假设存在两条具有竞争关系的信息同时在网络中传播, 记为竞争信息 A 和利己信息 B. 社交网络中的每个节点有三种状态: A 激活态、B 激活态和不活跃态. 处于不活跃态的节点可能仅会被信息 A 或信息 B 影响, 又或者同时受两者影响. 然而由于两条信息具有竞争关系, 因此节点只可能会被一条信息激活, 即

节点的激活态只能是 A 激活态或 B 激活态, 且该过程不可逆。

3.2 基于竞争的线性阈值扩展模型

社交网络影响力问题的研究需要借助传播模型来模拟信息在网络中的传播情况, 最基本且广泛使用的传播模型有独立级联模型^[32]和线性阈值模型^[33], 模型将节点分为两种状态: 活跃和不活跃。不活跃态节点可被其他活跃态邻居节点激活变成活跃态, 但反之不可。线性阈值模型的核心思想是通过累积效应来改变节点状态。阈值 θ 用于表示节点被激活的难易程度, 阈值越大表明节点越不容易被激活, 且节点 v 所受其他邻居节点的影响力之和不超过 1。

本文基于线性阈值模型进行扩展, 考虑到信息 A 和 B 虽具有竞争关系, 但二者的主题分布未必完全一致。且考虑社交网络中, 用户偏好、商家口碑等的不同导致商家所处的竞争环境处于不对等状态, 在信息 A 和 B 的传播中, 节点被各信息影响的概率是不同的, 分别记为 p_{uv}^A 和 p_{uv}^B , 且有 $\sum_{u \in N^{in}(v)} p_{uv}^A \leq 1$ 。

同理可解释 p_{uv}^B 。由于社交网络中用户互动程度不同, 即使两用户间具有双向边关系, 其彼此间的影响概率也不一定相同。假定节点 v 的两个独立阈值为 θ_v^A 和 θ_v^B , 其值均在 $[0, 1]$ 范围内, 分别表示接受信息 A 和信息 B 的阈值。在每个时间步, 信息 A 和信息 B 独立地同步传播, 与传统的线性阈值模型传播过程相似, 竞争环境中的线性阈值模型传播过程描述如下:

(1) 分别给定信息 A 和信息 B 的种子节点集合 S_A 和 S_B , 在时间步 t , 不活跃节点 v 会受所有活跃邻居节点的影响。

(2) 在时间步 t , 若 $\sum_{u \in N^{in}(v)} p_{uv}^A \geq \theta_v^A$ 且 $\sum_{u \in N^{in}(v)} p_{uv}^B < \theta_v^B$, 在时间步 $t+1$, 节点 v 处于 A 激活态;

(3) 在时间步 t , 若 $\sum_{u \in N^{in}(v)} p_{uv}^A < \theta_v^A$ 且 $\sum_{u \in N^{in}(v)} p_{uv}^B \geq \theta_v^B$, 节点 v 在时间步 $t+1$ 处于 B 激活状态, 否则处于不激活态;

(4) 在时间步 t , 若 $\sum_{u \in N^{in}(v)} p_{uv}^A \geq \theta_v^A$ 且 $\sum_{u \in N^{in}(v)} p_{uv}^B \geq \theta_v^B$, 节点 v 在时间步 $t+1$ 处选择被影响力累计总和较大的信息影响, 即 $\sum_{u \in N^{in}(v)} p_{uv}^A > \sum_{u \in N^{in}(v)} p_{uv}^B$? A 激活态: B 激活态;

(5) 当信息 A 或 B 均不可能再激活所有不活跃态的节点时, 传播过程结束。

4 基于主题偏好的节点影响概率建模

在实际社交网络中, 用户关注关系、用户交互强弱以及用户主题偏好等的不同, 均会造成用户间相互影响的程度不同, 即节点间的影响概率不同。本节通过求解节点的主题分布来计算节点间的影响概率。

4.1 节点间交互内容的主题分布求解

社交网络中, 用户可以根据个人兴趣偏好或线下关系选择关注不同的用户。因此, 用户与不同邻居好友间的交互内容各不相同, 即不同邻居间的影响是不同的。文献^[34]对社交网络新浪微博实证数据的分析表明, 用户个体偏好是经过长期的交互作用而逐渐形成的在一段时期内较为稳定的个性特征, 且用户的交互偏好较为固定, 用户更偏向与自己关注的好友进行交互。由于用户个体偏好变化与相邻用户间交互内容的变化是相互作用的, 即随着用户个体偏好的动态变化, 相邻用户间的交互内容也会随之变化, 反之亦然。考虑到信息传播的时效性及信息的生存周期性, 挖掘用户间在某一时段内交互内容的主题分布, 既可以反映出用户在该时段内的个体偏好, 也能够信息传播时效内更加精确地计算用户在不同主题上受各个邻居好友的影响分布。

为此, 本文使用 LDA 主题模型^[35]来发现节点 v 对节点 u 在某时段内交互内容的主题分布, 并将此分布作为有向边 $\langle u, v \rangle$ 的影响概率分布。如图 1 所示, 两个节点之间的关注关系对应一条有向边, 其上的交互内容对应一篇文档, 如果网络中具有关注关系的节点间均有交互行为, 则 n 条关注关系有向边对应 n 篇文档。由于 LDA 模型是以语料库为基准生成主题的, 不同的语料库会生成不同的主题, 因此为保证网络中所有文档的主题编号和个数的一致

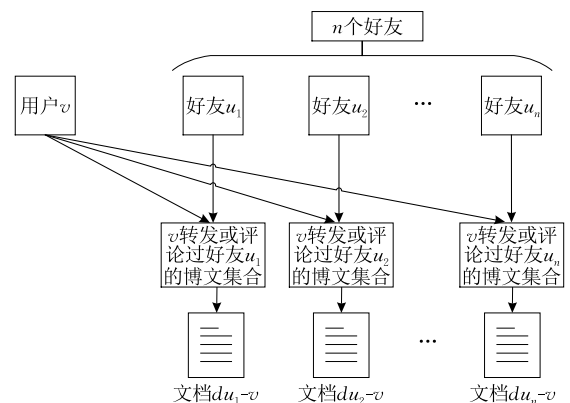


图 1 文档生成模型

性,需将网络中所有文档作为一个语料库输入至 LDA 模型中.同时,为保证传播信息 A 和 B 也服从相同主题编号和个数下的主题分布,可将信息 A 和 B 作为生成的 LDA 模型的输入,即可得到信息 A 和 B 在相同主题编号下的分布.

4.2 基于主题偏好的节点间影响概率计算

通过 LDA 模型可以得到节点间基于主题的影响概率分布,以信息 B 传播为例,信息 B 通过好友 u 影响节点 v 的概率 $p'_{uv} (p'_{uv} \leq 1)$ 的计算公式(1)为

$$p'_{uv} = \sum_{j=1}^k p(\lambda_B | z_j) p(u, v | z_j) \quad (1)$$

其中, $p(\lambda_B | z_j)$ 表示信息 B 为主题 z_j 的概率; $p(u, v | z_j)$ 表示节点 u 在主题 z_j 下对节点 v 的影响概率. p'_{uv} 为信息 B 的主题向量与节点间交互内容的主题向量的向量内积.为保证 v 的邻居节点对其的影响概率累和小于等于 1,对 p'_{uv} 做归一化处理.邻居节点 u 通过有向边 $\langle u, v \rangle$ 影响节点 v 的概率 p_{uv}^B 的计算公式(2)为

$$p_{uv}^B = \frac{p'_{uv}}{|N^{\text{in}}(v)|} \quad (2)$$

虽然信息间具有竞争关系,但由于传播内容的不同,会导致具有竞争关系的信息的主题分布不同,进而导致节点间传播概率的不同,且互为关注关系的好友间的影响概率也不同.

5 基于子图的节点影响力计算

5.1 节点子图生成

社交网络中的三度影响力理论指出信息在网络传播过程中存在衰减,并发现相距三度之内的节点间属强连接关系,强连接关系可以使节点间相互影响产生行为,而距离三度之外的节点间为弱连接关系,信息在弱连接中基本不产生影响^[36-37].结合三度影响理论,在计算节点 v 所受影响力时,可忽略计算影响作用较小的其他节点的影响力,只需计算影响作用大于某一阈值 φ_v 的其他节点对节点 v 的影响.本文使用节点子图方法为网络中的每个节点确定其受影响的节点集合.

定义节点 v 的子图 $G(v)$ 为: $\forall v \in V$, 若 $p_{u,v} \geq \varphi_v$, 其中 $u \in N^{\text{in}}(v)$, 则将节点 u 和边 $\langle u, v \rangle$ 加入到子图 $G(v)$ 中;采用广度优先搜索算法遍历 $G(v)$ 图中新加入的节点 u , 计算 u 的入度邻居 w 通过 u 影响 v 的概率 $p_v(w)$, 其中 $w \in N^{\text{in}}(u)$, 若 $p_v(w) \geq \varphi_v$, 则将节点 w 及边 $\langle w, u \rangle$ 加入到的 $G(v)$ 中.

子图 $G(v)$ 中节点 v 为根节点,根节点只会受到子图内的其他节点影响,因此使用节点子图计算节点的影响力可以提高影响最大化算法的运行效率.由于节点间在不同信息下有不同的影响概率 $p_{u,v}$, 因此对于同一节点 v , 在不同信息下有不同的子图.

算法 1. 节点子图 $G^B(v)$ 生成算法.

输入: 社交网络 $G(V, E)$, 节点 v , 节点 v 的阈值 φ_v^B .

$G(V, E)$ 中任一有向边 $\langle u, v \rangle$ 的影响概率 p_{uv}^B

输出: 节点 v 的子图 $G^B(v)$

1. Initialize: $X = \emptyset$; $Y = \emptyset$; $p_v^B(u) = 0$; $p_v^B(v) = 1$
2. IF $\max p_v^B(u)_{u \in V \setminus X} \geq \varphi_v^B$ DO
3. $x = \arg \max p_v^B(u)_{u \in V \setminus X}$
4. $Y = Y \cup \{(x, w) | w \in X\}$
5. $X = X \cup \{x\}$
6. FOREACH $u \in N^{\text{in}}(x)$ DO
7. $p_v^B(u) += p_{ux}^B \cdot p_v^B(x)$
8. END FOR
9. END IF
10. RETURN $D = (X, Y)$ as $G^B(v)$

其中, X 表示节点子图 $G^B(v)$ 中所有节点的集合, Y 为节点子图 $G^B(v)$ 中所有的边集合.若节点子图 $G^B(v)$ 中, 有 $m_x = |X|$ 个节点, 社交网络 $G(V, E)$ 中共 m 个节点, 则为每个节点生成子图的时间复杂度为 $O(m + m_x \log m)$.

5.2 节点影响力计算

在网络 $G(V, E)$ 中, 节点 u 对信息 B 的影响力计算公式(3)如下:

$$Inf^B(u) = \sum_{v \in \text{set}(u)} p_v^B(u) \cdot (1 - p_v^A(u)) \quad (3)$$

其中, $\text{set}(u)$ 表示包含节点 u 的所有子图的根节点集合.对于节点子图 $G^B(v)$ 中的除根节点外其他节点 u 对根节点 v 的影响概率 $p_v^B(u)$ 的计算如公式(4)所示.

$$p_v^B(u) = \sum_{x \in N^{\text{out}}(u)} p_{ux}^B \cdot p_v^B(x) \quad (4)$$

在传播模型中, 节点 u 仅在被信息 B 激活后才有机会影响其他节点.而其被激活的概率 $ap^B(u)$ 受其关注的邻居节点影响.在信息 B 传播过程中, 节点 u 被信息 B 激活的概率 $ap^B(u)$ 的更新公式(5)为

$$ap^B(u) = \begin{cases} \sum_{x \in N^{\text{in}}(u)} ap^B(x) \cdot p_{xu}^B, & u \in X \setminus S_B \\ 1, & u \in X \cap S_B \end{cases} \quad (5)$$

当节点 u 被选为信息 B 的种子节点时, 即 $ap(u) = 1$ 时, 节点 u 变为 B 激活态后, 只会影响其所在的各个子图中其他节点, 其中节点 u 所关注的邻居节点对其的影响力将变为 0, 而关注 u 的邻居节点的激活概率将会增加.由此, 在节点子图 $G^B(v)$

中,定义节点 u 对 v 的影响力增量计算公式(6)为

$$\Delta Inf^B(u, v) = p_v^B(u) \cdot a \cdot p^B(u) \quad (6)$$

5.3 基于子图的节点影响力计算实例

通过 LDA 主题模型可将社交网络建模成一个可以感知主题的有向网络拓扑图,如图 2 所示. 图中节点 u_1 与 u_6 之间为双向边,向量 $(0.3, 0.6, 0.1)$ 和 $(0.2, 0.1, 0.7)$ 分别表示 u_1 对 u_6 在三个主题下的影响概率和 u_6 对 u_1 在三个主题下的影响概率. 信息 B 下,节点 u_1 影响 u_6 的概率为 $(0.2, 0.1, 0.7) \cdot (0.3, 0.6, 0.1)^T = 0.19$, u_6 对 u_1 的影响概率为 $(0.2, 0.1, 0.7) \cdot (0.2, 0.1, 0.7)^T = 0.54$. 假设节点 u_1 和 u_4 在 t 时刻处于 B 激活态,则此刻节点 u_6 受到邻居节点的影响力总和为 0.197.

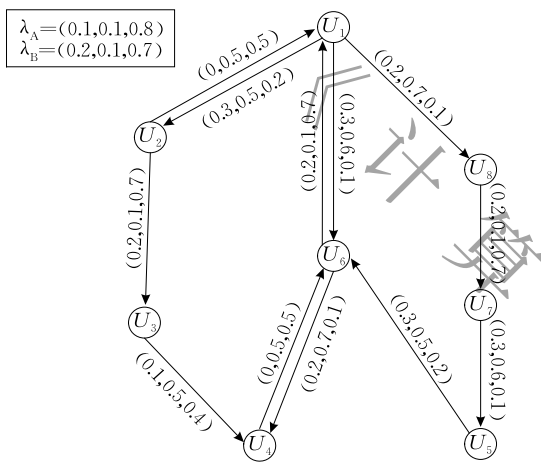


图 2 网络拓扑图示例

由于信息 A 和 B 的主题分布不同,其生成的节点子图也不同. 如图 3 所示,同一节点 v 在两条信息下生成不同的子图, $G^A(v)$ 表示在信息 A 下,节点 v 仅会受到 a, b, c, d 四个节点的影响; $G^B(v)$ 表示在信息 B 下,节点 v 仅会受到 a, u, c, d, e 五个节点的影响.

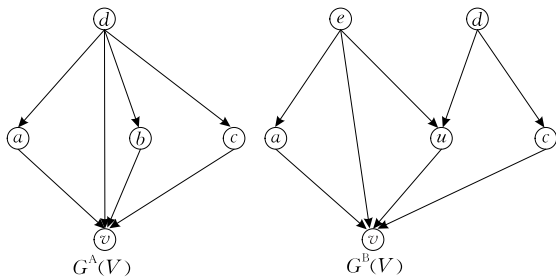


图 3 信息 A 与信息 B 下的节点子图示例

在计算信息 B 下点 e 对 v 的影响力时,只需累加计算从 $e \rightarrow a \rightarrow v, e \rightarrow v, e \rightarrow f \rightarrow v$ 这三条路径下点 e 对 v 的影响概率,即

$$p_v^B(e) = p_{ea}^B p_v^B(a) + p_{ev}^B + p_{ef}^B p_v^B(f).$$

6 竞争环境中基于主题偏好的影响力最大化算法

根据第 3 节问题定义,可将竞争环境中利己信息的影响力最大化问题依据利己信息对竞争信息种子节点选择策略的掌握情况分别求解. 当利己信息掌握了竞争信息种子节点选择策略,本文考虑到实际情况中竞争商家与种子用户之间涉及合同等利益因素,种子用户通常只能发布一个竞争商家的信息. 因此为避免竞争信息间选择同一节点作为种子节点的情况,且考虑竞争因素,提出节点避让的利己信息影响最大化算法 NA(Node Avoidance). 当未知竞争信息的广告投放策略时,本文假定竞争信息 A 种子节点选择策略集合为已知,设计一种竞争信息 A 的种子节点获取方法,以模拟竞争信息传播环境,提出策略无关的利己信息影响最大化算法 I3SL (Independent Seed Selection Strategy of Leader).

6.1 节点避让的利己信息影响最大化算法

节点避让的利己信息影响最大化算法(NA)的算法思想为:已知竞争信息 A 的种子节点集合,信息 B 在选择种子节点时,考虑信息 A 对节点的影响,并避免选择信息 A 的种子节点,仅在剩余节点中选择信息 B 的影响最大化种子节点集合. NA 算法伪代码如算法 2 所示.

算法 2. 节点避让的利己信息影响最大化算法.

输入: 社交网络 $G(V, E)$, 信息 B 的种子节点集合大小 k , 信息 A 的种子节点集合 S_A

输出: 信息 B 的种子节点集合 S_B

1. Initialize: $Inf^B(v) = 0$; $S_B = \emptyset$
2. FOREACH v in V DO
3. Generate $G^A(v)$ use Algorithm1
4. Compute $p_v^A(u)$ use Formula 5.2
5. END FOR
6. FOREACH v in $V \setminus S_A \cup S_B$ DO
7. Generate $G^B(v)$ use Algorithm1
8. Compute $p_v^B(u)$ use Formula 5.2
9. END FOR
10. FOREACH x in $V \setminus S_A \cup S_B$ DO
11. Compute $Inf^B(x)$ use Formula 5.1
12. END FOR
13. FOR $i = 1$ to k DO
14. $s = \arg \max Inf^B(x)$; (x in $V \setminus S_B \cup S_A$)
15. FOREACH v in $set(s) \setminus S_B \cup S_A$
16. //对于已激活节点 s 关注的邻居节点 u 将不再通过节点 s 影响 v , 从而导致节点 u 对 v 的影响概率减小, 节点 u 的影响力减小

```

17.  $\Delta p_v^B(s) = -p_v^B(s); u \text{ in } S_B \Delta p_v^B(u) = 0$ 
18. FOREACH  $u \in N^{\text{in}}(s) \text{ in } G^B(v)$  DO
19.   Compute  $\Delta p_v^B(u)$  use Formula 5. 2
20.    $p_v^B(u) += \Delta p_v^B(u)$ 
21.   Compute  $\Delta Inf^B(u, v)$  use Formula 5. 4
22.    $Inf^B(u) += (\Delta p_v^B(u) - \Delta Inf^B(u, v)) \cdot$ 
       $(1 - p_v^A(u))$ 
23. END FOR
24. //对于关注已激活节点  $s$  的邻居节点  $u$ , 其被激活
    的概率增大.
25.  $\Delta a p_v^B(s) = 1 - a p_v^B(s); u \text{ in } S_B \Delta a p_v^B(u) = 0$ 
26. FOREACH  $u \in N^{\text{out}}(s) \text{ in } G^B(v)$  DO
27.   Compute  $\Delta a p_v^B(u)$  use Formula 5. 3
28.    $a p_v^B(u) += \Delta a p_v^B(u)$ 
29.   Compute  $\Delta Inf^B(u, v)$  use Formula 5. 4
30. //由于节点  $s$  通过节点  $u$  能够影响根节点  $v$ , 导致  $u$ 
    对  $v$  的影响力增加, 因此节点  $u$  自身的影响力应减
    掉该部分的影响力.
31.  $Inf^B(u) -= \Delta Inf^B(u, v) (1 - p_v^A(u))$ 
32. END FOR
33. END FOR
34.  $S_B = S_B \cup \{s\}$ 
35. END FOR
36. RETURN  $S_B$ 

```

在 NA 算法中, 假设在信息 A 和信息 B 下, 每个子图生成时间为 t_φ , 子图平均节点的个数为 m_φ 个, $set(u)$ 的平均规模为 l_φ , 则在节点个数为 n 的网络 $G(V, E)$ 中, 寻找 k 个影响力最大的种子节点集合, NA 算法的平均时间复杂度为 $O(2nt_\varphi + kl_\varphi m_\varphi \log n)$.

6.2 策略无关的利己信息影响最大化算法

由于在未知竞争者策略的情况下, 竞争环境中的影响力最大化问题将不具有单调子模特性, 且无法确定最优解的边界. 在现有研究中, 一般随机选择一种策略作为竞争信息的种子节点选择策略, 其实质等同于本文已知竞争信息种子节点选择策略的情况. 为保证算法的通用性, 本文设计与竞争信息种子节点选择策略无关的利己信息影响最大化算法 (I3SL). I3SL 算法思想如下:

(1) 给定竞争信息种子节点选择策略集合 $M = \{m_1, m_2, \dots, m_n\}$ 和竞争信息种子节点集合大小 k . 对每一个策略 $m_i (m_i \in M)$, 基于子模特性, 计算其影响最大化种子节点集合 $S^i = \{s_1^i, s_2^i, \dots, s_k^i\}$, 集合 S^i 中某一节点 v 的影响力记为 $Inf_{s_v}^i$.

(2) 合并 n 个种子集合, 获得候选种子节点全集 $S^T = \{s_1^1, s_2^1, \dots, s_k^1, s_1^2, \dots, s_{j-1}^j, s_j^j, \dots, s_{k-1}^{n-1}, s_k^{n-1}, s_1^n, \dots, s_k^n\}$, 其中 $0 < j \leq k, 0 < i \leq n$;

(3) 若 $\exists s_{v_1}^i, s_{v_2}^j \in S^T$, 且 $s_{v_1}^i = s_{v_2}^j$, 则合并两节点,

并更新节点 v 的影响力 $Inf_{s_v} = Inf_{s_{v_1}}^i + Inf_{s_{v_2}}^j$; 否则节点 v 的影响力 $Inf_{s_v} = Inf_{s_v}^i$.

(4) 按照影响力大小为候选种子节点排序, 选择 top- k 个影响力最大的候选种子节点作为竞争信息的种子节点集合 $S = \{s_1, s_2, \dots, s_k\}$;

(5) 最后利用本文提出的 NA 算法, 计算利己信息的种子节点集合.

算法 3. 策略无关的利己信息影响最大化算法.

输入: 社交网络 $G(V, E)$, 信息 A 的种子节点集合大小 k_A , 信息 B 的种子集合大小 k_B , 信息 A 的种子节点选择策略集 S_S

输出: 信息 B 的种子节点集合 S_B

```

1. Initialize:  $Inf^A(v) = 0; Inf^B(v) = 0; S_A = \emptyset;$ 
    $S' = \emptyset; S_B = \emptyset$ 
2. //使用策略集合为信息 A 找寻种子节点候选集合
3. FOREACH strategy in M
4.   FOREACH v in G(V, E)
5.     Compute  $Inf(v)$  use strategy
6.   END FOR
7.   FOR  $i = 1$  to  $k_A$  DO
8.      $s = \arg \max Inf(v); (v \text{ in } V \setminus S')$ 
9.     IF  $s \text{ in } S'$ 
10.       $Inf^A(s) += Inf(s)$ 
11.     ELSE
12.       $Inf^A(s) = Inf(s)$ 
13.     END IF
14.      $S' = S' \cup \{s\}$ 
15.   END FOR
16. END FOR
17. FOR  $i = 1$  to  $k_A$  DO
18.    $s = \arg \max Inf^A(s); (v \text{ in } V \setminus S_A)$ 
19.    $S_A = S_A \cup \{s\}$ 
20. END FOR
21. Select  $S_B$  using NA-Algorithm
22. RETURN  $S_B$ 

```

在 I3SL 算法中, 其时间复杂度是由 NA 算法的运行时间和信息 A 种子节点选择策略集计算节点的影响力共同决定.

7 实验设计

7.1 数据集及实验环境

本文选择爬取国内主流社交网络——新浪微博的数据作为实证数据集对算法的有效性进行验证. 采用滚雪球的方式通过爬虫程序爬取新浪微博 2012 年 3 月至 2012 年 8 月的部分数据, 其中包括 19535931 个用户, 10785921 条博文及博文评论 501117 条. 通过

对新浪微博用户活跃度分析以及最大连通图的计算,最终获得的最大连通子图 $G(V,E)$ 包含 44 514 个较为活跃的节点,以及他们之间 162 398 条关注边关系,微博共 220 137 条,评论共 359 937 条.并通过话题关键字统计分析发现,原创、转发或评论过有关“iPhone 4S”的博文共计 10 315 条,原创、转发或评论过有关“小米手机”的博文共计 5663 条,因此本文将 iPhone 4S 的广告推荐设为信息 A,将小米手机的广告推荐设为利己信息 B 进行实验.通过对爬取数据的预处理,得到一个由活跃节点组成的网络拓扑图.对节点的双向关注关系度数进行分析,如图 4 所示,关注关系边服从幂律分布特征,说明按照本文的采样方法采集到的局部网络能够保持原网络的度特征.

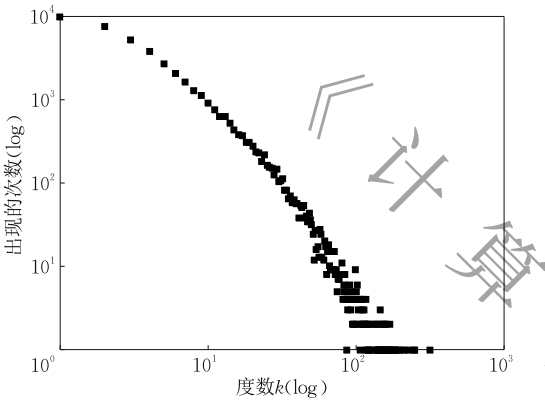


图 4 双向关注关系度数分布

本文在一台配置为 Intel(R) Core(TM) i5 处理器(4 核)、8GB 内存惠普 PC 机上进行了实验,操作系统为 Windows 7 旗舰版(64 位),编程语言为 C++ 和 python.

7.2 实验设计及评价指标

本文使用主题模型的方法计算节点间影响概率.在概率主题模型中,常用困惑度(perplexity)来评估主题模型的优劣,困惑度越小表明模型越好,且困惑度一般随着潜在主题数量的增加而呈现递减的规律^[35].

困惑度的计算公式(7)如下:

$$perplexity(M) = \exp \left\{ -\frac{1}{|N|} \sum_{w \in D} \log(P(w)) \right\} \quad (7)$$

其中, N 表示文档中所有单词的个数, $P(w)$ 表示每个单词 w 在文档集合 D 中出现的概率.通过计算不同主题数目的困惑度,便可以确定主题模型的最佳主题个数.

从第 2 节相关工作分析中可知,针对多信息同步传播环境下节点影响最大化问题,现有工作均对经典传播模型进行了合理的扩展,但与本文在研究目标、所采用的基本机制等方面存在较大差异,难以

进行具体对比分析,因此本文实验选取目前较为经典的 4 种启发式算法进行对比,以验证本文所提两种算法的有效性,对比算法如表 2 所示.由于贪心算法在较大规模的社交网络中运算时间过长,鉴于本文的数据规模较大,因此对比算法中不包含贪心算法.

表 2 实验对比算法

算法	算法描述
PageRank	衡量网页重要程度的算法,阻尼因子=0.15
Degree	利用节点度数找寻种子节点
CCA	利用核数性质和影响半径找寻种子节点
DegreeDiscount	利用度折扣思想计算种子节点的邻居节点的度数
NA	在竞争环境中通过避让竞争节点使利己信息影响范围最大
I3SL	在未知竞争信息节点选择策略时提出的策略无关的利己信息影响最大化算法

实验分别选取 1~50 个种子节点个数,并取 20 000 次蒙特卡洛模拟传播过程的实验结果的平均值作为最终的影响范围.种子节点数目为 i 时,算法 A_1 和 A_2 在的影响范围上的差异计算公式(8)为

$$Diff(A_1, A_2, i) = \frac{\sigma(A_1, i) - \sigma(A_2, i)}{\sigma(A_2, i)} \quad (8)$$

其中, $\sigma(A_1, i)$ 表示种子数目为 i 时算法 A_1 的影响范围.

为更好地反应各算法的优劣,使用各算法在种子节点数目为 1~50 时的影响范围的平均差异来衡量各算法之间的差异,平均差异的计算公式(9)为

$$Diff_{avg}(A_1, A_2) = \frac{1}{k} \sum_{i=1}^k Diff(A_1, A_2, i) \quad (9)$$

8 实验结果分析

影响力最大化算法常用运行时间和影响范围两个评价指标评估算法优劣.其中运行时间表示算法找到种子集合所花费的时间;影响范围表示通过选择的种子集合在传播模型中进行扩散传播,最终影响到的节点数目.

8.1 主题困惑度分析

本文使用 ICTCLAS 对博文进行分词处理,根据 4.2 节介绍的文档生成模型,使用 LDA 主题模型对所有文档进行主题发现.参考文献[35]中确定主题个数的方法,本文也使用主题困惑度随主题个数变化趋势来确定主题个数.如图 5 所示,主题模型困惑度随着主题个数的增多而减小,但当主题个数大于 50 时,模型困惑度降低趋势逐渐平缓,考虑到主题模型运算时间会随主题个数的增多而大幅增加,因此本文选择的主题个数为 50.

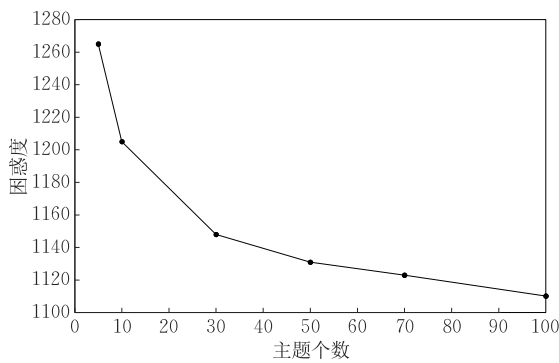


图 5 主题困惑度随主题个数变化趋势

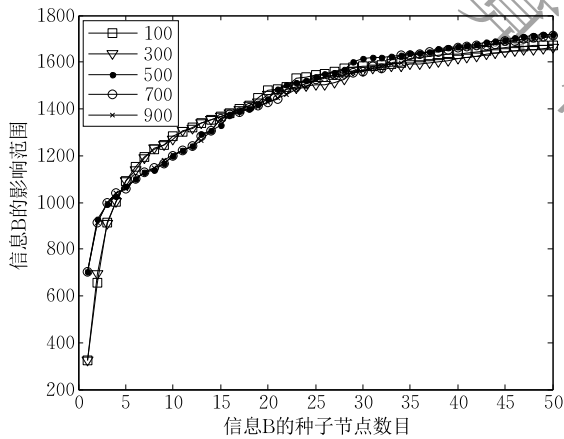
8.2 结果分析

NA 与 I3SL 算法在生成节点子图时会受阈值 φ 的影响, 阈值 φ 值越小, 子图的规模越大, 当 φ 值趋于无穷小时, 子图便是全局网络拓扑图 $G(V, E)$. 本文参考文献[11]确定两条信息在扩展的线性阈值模型中的阈值 θ , 参考文献[27]中阈值 φ 的取值标准, 分别选取 $1/\varphi = 100, 300, 500, 700, 900$ 进行对比实验. 本文将对信息 A 种子节点选择策略已知

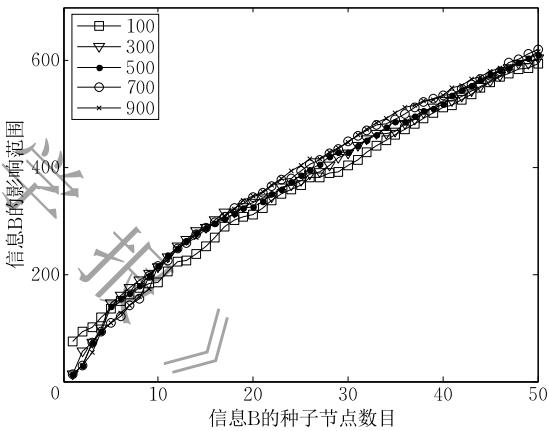
和未知的情况分别进行比较. 即在对 NA 算法的分析中, 本文将竞争信息 A 采用 PageRank、Degree 或 Random 策略作为已知信息, 并在信息 A 选择不同策略下分别对算法进行了对比分析; 在对 I3SL 算法的分析中, 本文使用 PageRank、Degree 或 Random 三种种子节点选择策略作为种子节点选择策略集合, 模拟选择信息 A 的种子集合, 并与启发式算法进行了分析比较.

8.2.1 NA 算法对比分析

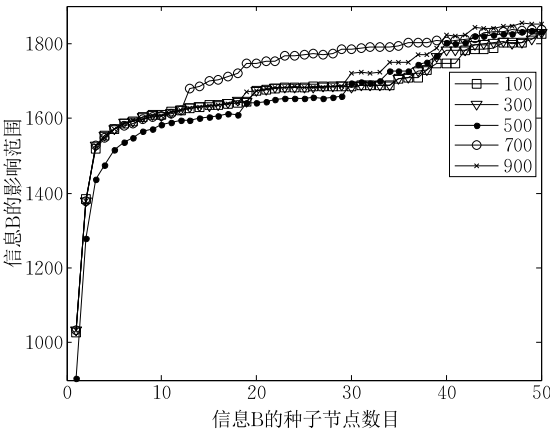
根据本文第 5 节对算法 1 的分析, 节点子图的生成过程会受阈值 φ 的影响. 阈值 φ 设置越小, 生成的节点子图的规模就越大, 算法时间复杂度越高. 因此, 实验中首先需要确定利己信息传播时, 节点受其影响转变成激活态的阈值. 如图 6 所示, (a)、(b)、(c) 分别表示采用 PageRank 算法、Degree 算法、Random 算法为信息 A 选取种子集合时, 利己信息 B 在 NA 算法下的影响效果. 综合分析种子节点及其影响范围, 确定在图 6(a) 中 $1/\varphi = 300$, 在图 6(b) 中 $1/\varphi = 500$, 在图 6(c) 中 $1/\varphi = 100$.



(a) 信息A选择PageRank作为种子选择策略



(b) 信息A选择Degree作为种子选择策略



(c) 信息A选择Random作为种子选择策略

图 6 在信息 A 种子选择策略已知情况下信息 B 选择 NA 算法中 $1/\varphi$ 取不同值时的影响效果

图 7 所示为竞争环境中,采用 PageRank 算法、Degree 算法、Random 算法为信息 A 选取种子集合时,信息 B 的影响效果.可清晰地看出,随着节点数的增加,NA 算法的优势越来越明显.

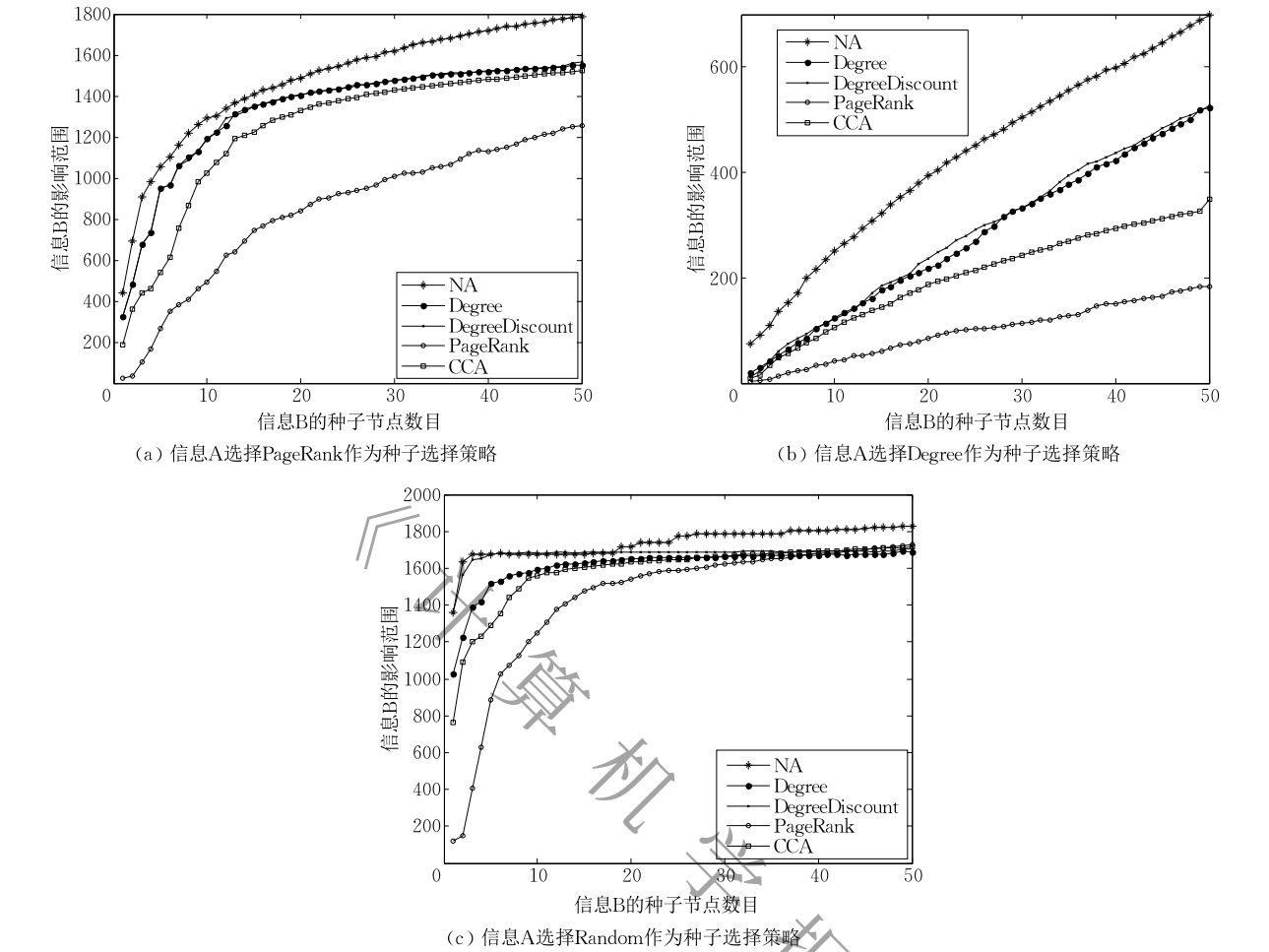


图 7 在信息 A 种子选择策略已知情况下信息 B 在不同策略下的影响效果对比

利用公式(9)可以计算各对比算法与 NA 算法在影响范围方面的差异.当采用 PageRank、Degree、Random 算法分别为信息 A 选取种子集合时,信息 B 使用各对比算法与 NA 算法在影响范围方面的差异如表 3 所示.由于 NA 算法考虑了竞争信息 A 对节点的影响,因此 NA 算法显著优于其他对比算法.

信息 B 使用的 算法名称	信息 A 使用的算法名称		
	PageRank 下的 差异百分比/%	Degree 下的 差异百分比/%	Random 下的 差异百分比/%
PageRank	-45.03	-77.2	-16.87
Degree	-12.30	-35.0	-6.66
CCA	-16.50	-53.0	-8.20
DegreeDiscount	-12.10	-33.7	-3.03

当信息 A 与 B 同时在基于竞争的线性阈值扩展模型中进行传播影响时,信息 A 的影响范围在竞争环境中会有所抑制.分析信息 B 的种子节点的选择对竞争信息 A 的抑制作用,如图 8 所示,由于 NA

在选取种子节点时主要关注的是利己信息 B 的影响最大化,而较少考虑所选节点对信息 A 的抑制作用,因此算法在抑制信息 A 影响效果方面优势不明显.使用 Random 算法为信息 A 选择初始种子节点集合时,如图 8(c)所示,信息 A 的影响范围最小.且随着信息 B 种子节点数目的增加,信息 A 的影响范围在各对比算法下均趋于平稳,而在 NA 算法中仍有下降趋势,但 NA 算法对信息 A 的影响范围的抑制作用最小,这是由于信息 A 种子节点集合的影响范围较小,导致信息 A 对信息 B 的竞争作用减小.

采用 PageRank、Degree、Random 算法分别为信息 A 选取种子集合时,信息 B 使用各算法的运行时间如表 4 所示.由于 Degree、CCA、DegreeDiscount 算法只考虑了网络的拓扑结构,因此运算时间均小于 1s,而 PageRank、NA 算法需要花费时间计算节点间的影响概率,且 NA 算法还需针对每条信息为每个节点生成不同的节点子图,因此运行时间较长,

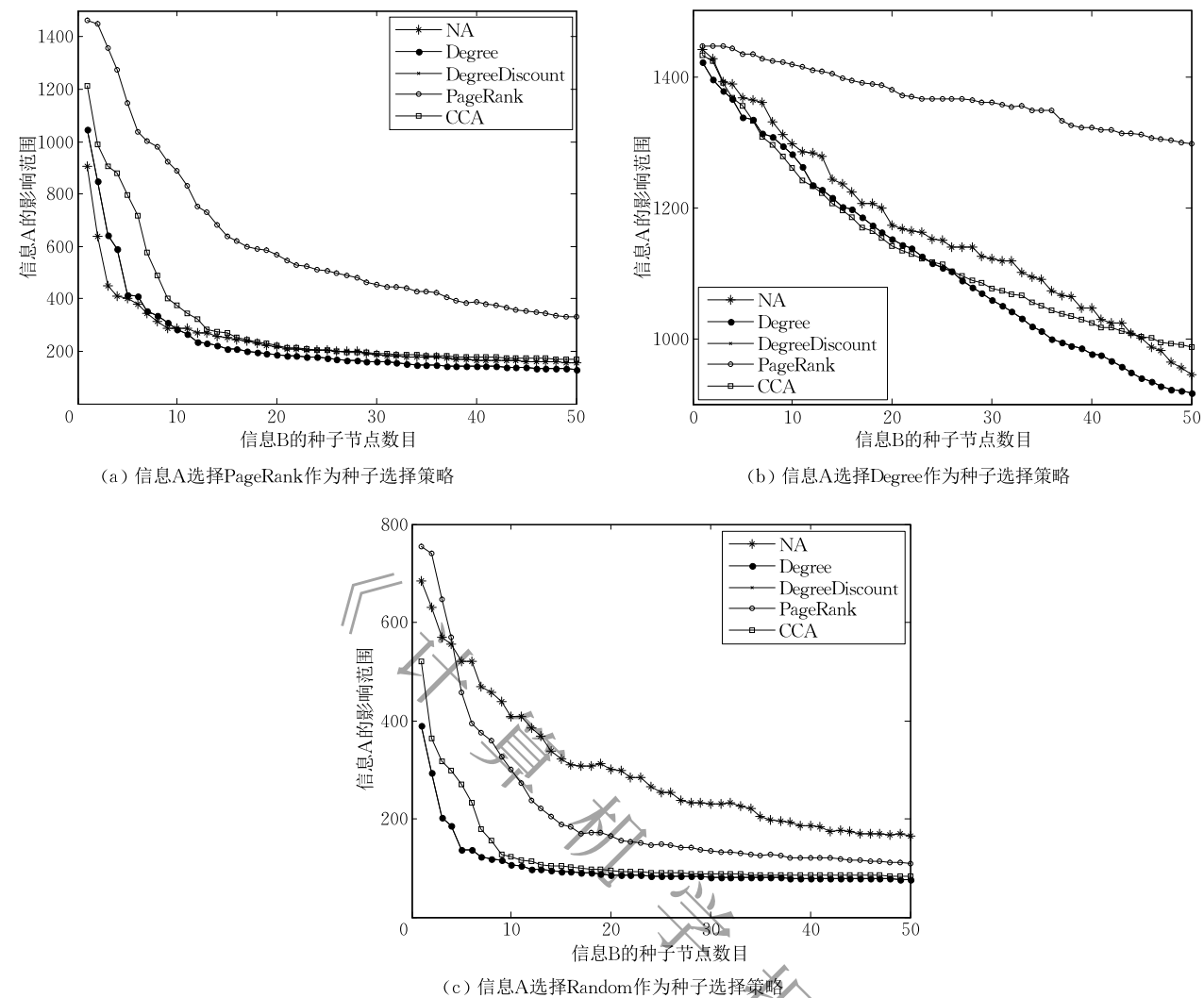


图 8 在信息 A 种子选择策略已知情况下信息 B 选择不同策略对信息 A 影响范围的抑制效果对比

但都在 1 min 左右. 其中 NA 算法在 Degree 情况下的运行时间明显小于在其他两种情况下的运行时间,是由于使用 Degree 算法为信息 A 选取的种子集合为度最大的节点集合,NA 算法在生成信息 B 的节点子图时便不考虑大度节点,大量缩减了节点子图的生成时间. 而 NA 算法在 Random 情况下的运行时间最长,是由于阈值设置较大($1/\varphi=100$),导致生成的节点子图规模较大,生成时间较长,并最终导致节点影响概率花费时间也较长.

表 4 各算法运行时间

信息 B 使用的 算法名称	信息 A 使用的算法名称		
	PageRank 下的 运行时间/s	Degree 下的 运行时间/s	Random 下的 运行时间/s
PageRank	51.386	51.570	51.340
Degree	0.234	0.250	0.250
CCA	0.750	0.830	0.702
DegreeDiscount	0.031	0.015	0.015
NA	69.050	25.160	98.760

8.2.2 I3SL 算法对比分析

当信息 A 的种子策略已知时,由表 3 所示,各算法的影响范围按从大到小的顺序依次为:NA、DegreeDiscount、Degree、CCA、PageRank. 在未知其策略的情况下,为证明本文所提算法能够适合各种竞争策略,现采用 PageRank、Degree、Random 三种种子节点选择策略作为策略集合为竞争信息 A 选取种子节点. 利己信息 B 在 I3SL 算法下的影响效果如图 9 所示,当 $1/\varphi=700$ 时,I3SL 算法的影响范围具有较好的优势,故设 $1/\varphi=700$.

图 10 所示 I3SL 算法与其他算法在竞争环境中的影响结果,即便未知竞争信息 A 的种子节点选择策略,I3SL 算法依旧明显优于其他对比算法.

表 5 所示为各算法影响范围的百分比差异. 发现,在未知信息 A 种子节点选择策略时,使用 I3SL 算法的结果表现最优.

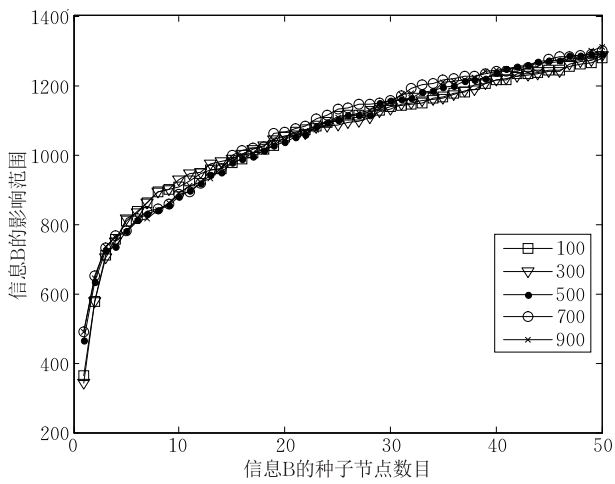


图 9 在信息 A 种子选择策略未知情况下信息 B 选择 I3SL 算法中 $1/\varphi$ 取不同值时影响效果对比

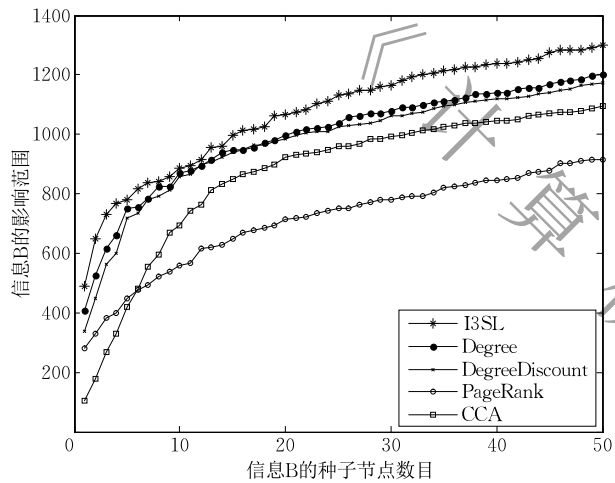


图 10 在信息 A 种子选择策略未知情况下信息 B 在不同策略下的影响效果对比

表 5 各算法与 I3SL 算法在影响范围方面的对比

信息 B 使用的算法名称	差异百分比/%
PageRank	-34.55
Degree	-7.40
CCA	-21.46
DegreeDiscount	-9.70

分析信息 B 的种子节点的选择对竞争信息 A 的抑制作用,如图 11 所示.对比信息 A 策略已知的情况,可知信息 A 使用策略集合选择的种子集合的影响范围略低于 PageRank 和 Degree 情况下的影响范围.

表 6 所示为各算法的运算时间.由于 I3SL 算法需要先模拟信息 A 的种子集合之后,再确定信息 B 的种子集合,因此算法的运行时间较长.

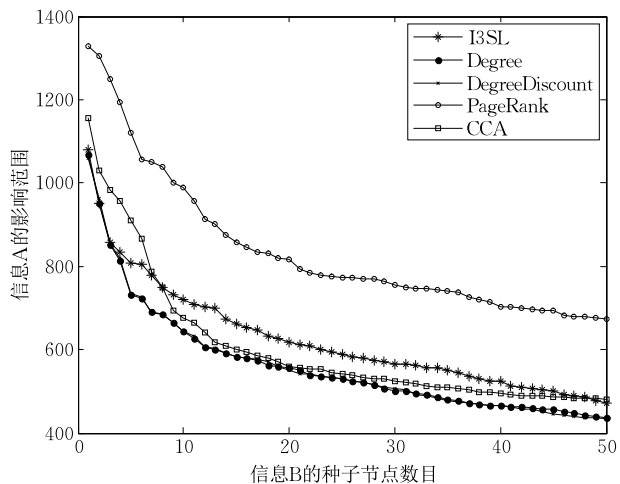


图 11 在信息 A 种子选择策略未知情况下信息 B 选择不同策略对信息 A 影响范围的抑制效果对比

表 6 各算法运行时间

信息 B 使用的算法名称	运行时间/s
PageRank	53.624
Degree	0.280
CCA	0.760
DegreeDiscount	0.017
I3SL	89.960

9 总 结

本文主要研究竞争环境中利己信息的影响力最大化问题.考虑节点间交互内容的主题偏好特点,设计基于主题模型的节点间影响概率计算方法,并基于传统线性阈值模型设计竞争环境中的传播模型,最后结合传播模型的特点及信息扩散的特征,提出基于节点子图的影响力计算方法,以两条竞争信息传播为例,提出两种利己信息影响最大化算法:节点避让的利己信息影响最大化算法 NA(Node Avoidance)和策略无关的影响最大化算法 I3SL(Independent Seed Selection Strategy of Leader).基于本文提出的线性阈值扩展模型和实证数据集上进行实验,结果显示:针对竞争信息的不同选择策略,本文所提的两种算法在影响范围方面与其他算法相比均表现最优.

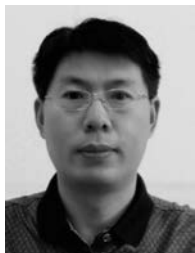
未来的工作将在以下几个方面进行.首先在对问题分析中考虑时间因素,研究网络拓扑结构的动态性及传播时延对多信息同步传播结果的影响.其次,对本文所提算法进行优化,使其找寻的种子节点能够满足自身信息的影响最大化,同时能够最大化抑制竞争信息的影响范围.

致 谢 感谢审稿专家和编辑在论文评审过程中提出的宝贵意见和建议!

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, USA, 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, USA, 2003: 137-146
- [3] Cao Jiu-Xin, Dong Dan, Xu Shun, et al. A k -core based algorithm for influence maximization in social networks. Chinese Journal of Computers, 2015, 38(2): 238-248(in Chinese)
(曹玖新, 董丹, 徐顺等. 一种基于 k -核的社会网络影响最大化算法. 计算机学报, 2015, 38(2): 238-248)
- [4] Li Yue-Zhi, Zhu Yuan-Yuan, Zhong Ming, et al. k -core filtered influence maximization algorithms in social networks. Journal of Computer Applications, 2018, 38(2): 464-470(in Chinese)
(李闯志, 祝园园, 钟鸣. 基于 k -核过滤的社交网络影响最大化算法. 计算机应用, 2018, 38(2): 464-470)
- [5] Ma Qian, Ma Jun. Discovering the substitutes for the seeds in influence maximization problems. Chinese Journal of Computers, 2017, 40(3): 674-686(in Chinese)
(马茜, 马军. 在影响力最大化问题中寻找种子节点的替补节点. 计算机学报, 2017, 40(3): 674-686)
- [6] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-208
- [7] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1029-1038
- [8] Wang Rui, Liu Yong, Zhu Jing-Hua, et al. Social network information diffusion model based on user's influence and interesting. Journal on Communications, 2017, 38(s2): 113-121(in Chinese)
(王瑞, 刘勇, 朱敬华等. 基于用户影响与兴趣的社交网信息传播模型. 通信学报, 2017, 38(s2): 113-121)
- [9] Xie Sheng-Nan, Liu Yong, Zhu Jing-Hua, et al. Research on topic-based local influence maximization algorithm in social network. Journal of Frontiers of Computer Science & Technology, 2016, 10(5): 646-656(in Chinese)
(谢胜男, 刘勇, 朱敬华等. 社会网中基于主题的局部影响最大化算法研究. 计算机科学与探索, 2016, 10(5): 646-656)
- [10] Wang Q, Jin Y, Lin Z, et al. Influence maximization in social networks under an independent cascade-based model. Physica A: Statistical Mechanics and Its Applications, 2016, 444: 20-34
- [11] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model//Proceedings of the IEEE International Conference on Data Mining. Sydney, Australia, 2010: 88-97
- [12] Bharathi S, Kempe D, Salek M. Competitive influence maximization in social networks//Proceedings of the International Workshop on Web and Internet Economics. Berlin, Germany, 2007: 306-311
- [13] Grossman D A, Frieder O. Information Retrieval: Algorithms and Heuristics. Berlin, Germany: Springer, 2012
- [14] Bozorgi A, Samet S, Kwisthout J, et al. Community-based influence maximization in social networks under a competitive linear threshold model. Knowledge-Based Systems, 2017, 134: 149-158
- [15] Borodin A, Filmus Y, Oren J. Threshold models for competitive influence in social networks//Proceedings of the International Workshop on Internet and Network Economics. Berlin, Germany, 2010: 539-550
- [16] Carnes T, Nagarajan C, Wild S M, et al. Maximizing influence in a competitive social network: A follower's perspective//Proceedings of the International Conference on Electronic Commerce. Minneapolis, USA, 2007: 351-360
- [17] Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks. Games & Economic Behavior, 2010, 70(2): 194-227
- [18] Zhou W, Jia W, Haghighi M, et al. A sword with two edges. Propagation studies on both positive and negative information in online social networks. IEEE Transactions on Computers, 2015, 64(3): 640-653
- [19] Lim Y, Ozdaglar A, Teytelboym A. Competitive rumor spread in social networks. ACM SIGMETRICS Performance Evaluation Review, 2017, 44(3): 7-14
- [20] Zhu Y, Li D, Zhang Z. Minimum cost seed set for competitive social influence//Proceedings of the IEEE INFOCOM 2016 the IEEE International Conference on Computer Communications. San Francisco, USA, 2016: 1-9
- [21] Liu W, Yue K, Wu H, et al. Containment of competitive influence spread in social networks. Knowledge-Based Systems, 2016, 109(C): 266-275
- [22] Ribeiro B, Faloutsos C. Modeling website popularity competition in the attention-activity marketplace//Proceedings of the 8th ACM International Conference on Web Search and Data Mining. Shanghai, China, 2015: 389-398
- [23] Narayanam R, Nanavati A. Viral marketing for product cross-sell through social networks. Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer, 2012: 581-596

- [24] Zhang J, Wang S, Zhan Q, et al. Intertwined viral marketing in social networks//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway, USA, 2016: 239-246
- [25] Wu H, Liu W, Yue K, et al. Selecting Seeds for Competitive Influence Spread Maximization in Social Networks. Social Computing. Singapore: Springer, 2016
- [26] Beutel A, Prakash B A, Rosenfeld R, et al. Interacting viruses in networks: Can both survive?//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 426-434
- [27] He X, Song G, Chen W, et al. Influence blocking maximization in social networks under the competitive linear threshold model//Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, California, USA, 2012: 463-474
- [28] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-815
- [29] Tzoumas V, Amanatidis C, Markakis E. A game-theoretic analysis of a competitive diffusion process over social networks. Internet and Network Economics. Berlin, Germany: Springer, 2012: 1-14
- [30] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowledge and Information Systems, 2013, 37(3): 555-584
- [31] Chen W, Lin T, Yang C. Efficient topic-aware influence maximization using preprocessing. The Computing Research Repository, 2014, 9197: 1-13
- [32] Watts D J. A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences, 2002, 99(9): 5766-5771
- [33] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978, 83(6): 1420-1443
- [34] Diao S M, Liu Y, Zeng Q A. Empirical analysis of relationship-based user reposting behavior on microblog network. International Journal of Interdisciplinary Telecommunications and Networking, 2015, 7(3): 1-12
- [35] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022
- [36] Christakis N A, Fowler J H. Social contagion theory: Examining dynamic social networks and human behavior. Statistics in Medicine, 2013, 32(4): 556
- [37] Arney C. Connected: The surprising power of our social networks and how they shape our lives — How your friends' friends' friends affect everything you feel, think, and do. Mathematics and Computer Education, 2014, 48(2): 201



CAO Jiu-Xin, Ph.D., professor.

His research interests include social computing, web services and cyber security.

MIN Hui-Yu, Ph.D. candidate. Her research interests include complex network, social computing, multi-agent

system.

WANG Hao-Ran, M.S. candidate. His research interest is social computing.

MA Zhuo, Ph.D. candidate. Her research interests include complex network, social computing, privacy protection.

LIU Bo, Ph.D., associate professor. Her research interests include pervasive computing and social computing.

Background

In recent years, social network analysis has obtained a great attention with the huge number of users and the popularization of the social network sites such as Sina Weibo, FaceBook and Twitter. In the field of social network research, there are following research directions: theoretical analysis of social network structure, information propagation in social networks, community structure analysis and so on. The problem we study in this paper is influence maximization problem in the competitive environment which is about information propagation. It is the problem of obtaining a set of nodes with specified size in a social network to maximize their aggregate influence for the target information under the

competitive linear threshold model in the competitive environment. The problem is very meaningful and it has many practical applications such as marketing by identifying influential users, online advertising and so on.

The traditional influence maximization problem was proposed as a discrete optimization problem in 2003. The optimization problem is proved to be NP-hard and the balance of the running time and spread range is the main bottleneck for large social networks. So far, there are two kinds of algorithms to solve the problem. One is the greedy algorithms which have larger spread range and spend more running time. Another is the heuristic algorithms which run faster but have

lower spread range. The competitive influence maximization problem was proposed in 2007. Researchers then explored this problem in extended independent cascade model and linear threshold model. And it has been proved that in this two extended propagation models, the competitive influence maximization problem is still a NP-hard problem.

To solve the problem happened during simultaneous competitive dissemination, our work considers the synchronization of competitive information dissemination, as well as the law that user's influences of different information are not the same. We consider the mastery of competitors' dissemination strategies into account, and provide two algorithms. The Experiments results show that the two algorithms we

propose have the better performance in the scope of influence, also have a better balance between efficiency and scope of influence.

This work is supported by the National Natural Science Foundation of China under Grants No. 61772133, No. 61472081, the Jiangsu Provincial Key Project No. BE2018706, the Key Laboratory of Computer Network Technology of Jiangsu Province, the Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201, the Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No. 93k-9.

《计算机学报》