基于社区度的边界节点影响力最大化算法*

王 双,李 斌,刘学军,胡 平
(南京工业大学 电子与信息工程学院,江苏 南京 211816)

摘 要:通常跨社区的信息传播更具有现实意义,而且大范围的信息传播往往也是跨社区的。为此提出一种基于社区度的边界节点影响力最大化算法,利用社会网络中的社区结构对社区中与其他社区有连接边的边界点进行研究,从而缩小选择初始节点的范围,降低时间复杂度。同时为更准确地评估边界节点的影响力,综合节点度、节点所直接相连社区数以及相应社区的规模作为社区度来衡量节点在信息传播中的重要性。最后通过实验验证了本算法相比其他算法具有更大的影响传播范围和更低的时间复杂度。

关键词:影响力最大化;社区度;边界节点;社会网络
中图分类号:TP311 **文献标识码**:A **文章编号**:0258-7998(2015)05-0145-04
DOI:10.16157/j.issn.0258-7998.2015.05.036

An influence maximization algorithm of boundary nodes based on degree of community

Wang Shuang, Li Bin, Liu Xuejun, Hu Ping
(College of Electronic and Information Engineering, Nanjing Tech University, Nanjing 211816, China)

Abstract: Information spread is more practical significance between communities, and a wide range of information spread is also cross-community. In this respect, the paper presents an influence maximization algorithm based on degree of community for boundary node, utilizes the community structure of social network to research boundary nodes which transfer information outwardly to other communities, thereby shrinking the range of initial nodes to reduce the computational complexity. At the same time, in order to assess the influence of the nodes accurately, the integrated node degree, the number of community nodes connected directly and the size of the communities are used as community degrees to measure the importance of the nodes in the dissemination of information among the community. Finally, the proposed algorithm has a greater impact on the spread of range and lower time complexity when compared with others through experiments to validate the algorithm.

Key words: influence maximization; community degree; boundary node; social network

0 引言

近年来,随着互联网和 Web 技术的不断革新,影响最大化问题作为社会网络分析领域的重要问题得到了快速发展,并已引起越来越多的学者关注。Li 等[1]研究了基于位置感知的影响力最大化问题,通过用户影响力的上界选择种子节点并消除不重要的节点,减少了初始种子节点的选择范围。Kim 等[2]基于 IC 模型将独立影响路径作为影响评估单元,但是算法未考虑不同节点影响力的相关性。Zhao 等[3]提出基于网络社区结构的节点影响力度量策略,但是算法未考虑节点度在信息传播中的重要性。Jung 等[4]提出了基于 IC 扩展模型的 IRIE 算法。Guo 等[5]提出基于个性化的影响力最大化算法,对给定目标用户,寻找对该目标用户影响力最大的节点。Cao 等[6]提出动态规划算法(OASNET)解决影响力最大化,但此算法没有考虑社区间的连通性。Tian 等[7]提出结合启发算法和贪心算法的影响力最大化算法 HPG,但算法在启发阶段仅以节点的度计算潜在影响力,没有充分考虑节点的其他属性。与上述研究不同,本文将社区间的边界节点作为初始种子节点集的候选集,以减少社区内大量非边界点的计算时间,提高运行效率。同时,传统以

* 基金项目:国家公益性科研专项(201310162);连云港科技支撑计划项目(SH1110)

节点度的倒数衡量节点间影响力忽略了邻居节点对节点影响的差异，基于此本文综合考虑边界节点的度、所连社区数、所连社区规模均值3个因素衡量节点对邻居的影响力传播的重要性，以更准确衡量节点影响力。

1 基于社区度的边界节点影响力最大化算法

本文提出的基于社区度的边界节点影响力最大化算法(CDEA)建立在具有社区结构的社会网络基础上，利用HPG算法框架实施。CDEA算法将社区连接边的两端节点作为边界节点，从边界节点集中选择初始传播种子节点，通过LT模型在社会网络中传播，使初始种子节点产生的影响范围最大。CDEA算法从边界节点集中选择初始种子节点是由于在具有社区结构的社会网络中，跨社区的信息最大化传播往往更有现实意义，而边界节点是社区间信息交流的大使，跨社区的信息传播必然会经过社区边界节点，因此可以先不考虑社区内大量的非边界节点，只考虑少量的边界节点，可极大降低算法运行时间。同时CDEA算法用边界节点的度和与边界节点直接相连的社区数以及社区规模均值综合衡量节点的潜在影响力，提高计算节点在贪心阶段被快速激活的可能性。

1.1 节点社区度

节点社区度是指边界节点的度、与边界节点直接连接的社区数、直接相连的社区规模均值三者叠加。社区度既考虑节点度，也结合节点在社会网络中的位置和连通性，可以较好地评估节点对影响力传播的重要性。

社会网络中与多个社区有连接边的节点称为边界节点。节点 u 的社区表示为 $C_i^u(i\in[1,M],u\in V)$。如果 u 的邻居集 $\mathrm{Nei}(u)$ 中有部分节点不属于 C_i^u，则认为这些节点为边界节点，用 $B(u)(u\in C_i^u,\exists_{v\in\mathrm{Nei}(u)}\notin C_i^u)$ 表示。如果 $B(u)=1$ 则代表节点 u 是边界节点，如果 $B(u)=0$ 则代表节点 u 不是边界节点。

社区规模主要用于描述节点所连社区的平均节点数，规模越大，影响越大，也更有利于信息的传播，而规模小的社区在信息传播过程中的重要性相对要小。用社区中包含的节点数表示 u 所在社区 C_i^u 的规模 $\mathrm{SC}(C_i^u)$。

$$\mathrm{SC}(C_i^u)=\left|C_i^u\right| \qquad (i\in[1,M]) \tag{1}$$

定义1（社区度）社区度主要用于衡量边界节点在影响力传播中的重要性。社区度是节点在社区中邻居数、与节点直接相连的社区数以及所连社区规模均值的叠加和。节点在社区中的邻居数越多，表明节点在社区中越重要，对其他节点产生影响的可能更大。节点直接相连的社区数越多，说明节点与其他社区产生交流的机会越大，对信息的广泛传播具有重要意义。而所连社区规模的大小将直接影响信息能否通过所连社区继续向下一个社区扩散，所连社区规模越大，则此社区在整个社会网络中对信息的快速传播作用越大。因此采用三者的叠加和可以突出节点在信息传播中的重要性。社区度 $\mathrm{CDeg}(u)$ 定义如下：

$$\mathrm{CDeg}(u)=\begin{cases}\mathrm{IDeg}(u)+\mathrm{ODeg}(u)+\dfrac{1}{e^{\mathrm{AvgN}(u)^{-1}}} & B(u)=1\\ \mathrm{IDeg}(u) & B(u)=0\end{cases} \tag{2}$$

其中 $\mathrm{IDeg}(u)=\sum_{v,u\in C_i^u,v\in\mathrm{Nei}(u)} n_{uv}$ 表示节点 u 在社区 C_i^u 中拥有的邻居数，$n_{uv}=1$ 表示节点 u 和 v 是邻居，否则 $n_{uv}=0$。

$\mathrm{ODeg}(u)=\sum_{v\notin C_i^u,v\in\mathrm{Nei}(u)} n_{uv}$ 表示与节点 u 直接相连的社区数。

$\mathrm{AvgN}(u)=\dfrac{\sum_{j\neq i,v\in\mathrm{Nei}(u),v\notin C_i^u}\mathrm{SC}(C_j^v)}{\mathrm{ODeg}(u)}$ 表示节点 u 的邻居社区的社区规模均值，C_j^v 表示节点 v 的社区，且与 u 所在社区相邻。

社区规模均值可缩小，因为邻居社区数少而邻居社区节点数多或邻居社区多而邻居社区节点数少所造成的差异能更好地平衡社区数和每个社区规模间的关系，因此综合考虑与节点直接相连的邻居社区以及相应社区规模均值，可更准确描述社区度，提高节点获取潜在影响力节点的精度。

1.2 节点影响概率

本文综合边的权重和节点间的交流频度计算节点间的影响概率，更有效地度量节点间相互影响的概率。影响概率即为节点激活邻居的可能性，当节点的所有已激活邻居对其影响概率和大于等于特定的阈值 θ，则节点被激活。本文定义节点 u 对邻居节点 v 的影响概率为 $b_{uv}=\dfrac{w_{uv}t_{uv}}{\sum_{u'\in\mathrm{Nei}(v)}(w_{u'v}t_{u'v})}$，其中 t_{uv} 为社会网络 G 中节点 u 和 v 信息交流频度，w_{uv} 为节点 u 到 v 的边权重值，$\mathrm{Nei}(v)$ 表示节点 v 的邻居节点集。

1.3 CDEA算法框架

社区度描述了边界节点在整个网络中的拓扑结构和重要性，与传统单一采用节点度描述节点与邻居的关联度相比，可更好地衡量节点潜在影响力。本文对HPG算法改进，基于社区度提出新的混合算法CDEA。CDEA算法分为启发阶段和贪心阶段。

基于LT传播模型的影响累积特性在启发阶段对边界节点启发式寻找最有影响力的 k' 个节点作为种子节点。这些节点不是局部影响力最大的节点，但是其潜在影响力在后续信息传播激活过程中将会被充分挖掘，最终获得比KK算法更大的影响范围。令 $\inf(u)$ 为节点 u 对所有未被激活邻居节点的影响力之和，则：

$$\inf(u)=\sum_{v\in\mathrm{Nei}(u),v\in A(u)} b_{uv} \tag{3}$$

基于此，对节点 u，其潜在影响力PI为：

$$\mathrm{PI}(u)=\mathrm{CDeg}(u)+(1-e^{-\inf(u)}) \tag{4}$$

这里 $CDeg(u)$ 表示节点 u 的社区度，$Nei(u)$ 表示节点 u 的邻居节点集合，$A(u)$ 表示节点 u 的邻居中未被激活的节点集。PI 综合考虑了节点的社区度和对邻居中未激活节点的影响范围。启发阶段从未激活的节点中选择潜在影响力最大的节点，并将其加入初始集合，直到出现 k_1 个已经被激活的节点。贪心阶段基于 LT 信息传播模型，应用 KK 算法不断计算边际影响收益，在局部最优的情况下获取 $k-k_1$ 个最有影响力的节点。

CDEA 算法具体步骤如下：

输入：社会网络 $G=(V,E,W)=\{C_1,C_2,C_3,\cdots,C_M\}$，阈值 θ，启发因子 c，初始集合大小 k。

输出：大小为 k 的目标节点集 S，最终激活节点数 k_0，启发阶段激活节点数 k_1，贪心阶段激活节点数 k_2。

(1)集合 $S_0=\phi$，$k_1=k-\lceil ck \rceil$，$k_2=\lceil ck \rceil$；

(2)FOR $i=0$ TO k_1；

(3)基于社区度概念从未激活的节点中选择 PI 最大的节点 v；

(4)$S_{i+1}=S_i \cup \{v\}$；

(5)基于集合 S_{i+1} 激活未激活的节点，并更新节点的 PI；

(6)END FOR；

(7)FOR $i=0$ TO k_2；

(8)计算每一个尚未激活节点被激活后的影响增量；

(9)选择影响力增量最大的节点 v；

(10)令 $S_{i+k_1+1}=S_{i+k_1} \cup \{v\}$；

(11)更新被节点 v 影响的节点状态；

(12)END FOR；

(13)输出目标节点集 S。

1.4 CDEA 算法复杂度分析

启发阶段，由于静态社会网络中式(2)中节点社区度是固定的，并且只需要计算社区间的边界节点的社区度，而非整个网络中所有节点，且 $inf(u)$ 是基于前一个初始种子节点并更新整个网络后确定，基本不耗时间，因此时间复杂度为 $O(C)$。同时启发阶段不但获取了大量具有潜在影响力的节点，而且也激活了大量节点。贪心阶段，由于有大量节点已被激活，未激活节点比初始状态下需要激活节点数减少了大部分，即可看作 KK 算法运行在小规模数据集，因此算法复杂度比 KK 小。

KK、HPG 以及 CDEA 算法不同阶段的时间复杂度如表 1 所示。其中 Q_1、Q_2 分别表示启发阶段 CDEA 算法和 HPG 算法每个种子节点的平均激活节点数。T_1、T_2、T_3 分别表示贪心阶段 CDEA 算法、HPG 算法、KK 算法每个种子的平均激活范围。N、M、M' 分别表示社会网络 G 的节点数、边数、社区边界节点数。$M'<<M<<N$，因此可推断 CDEA 算法比 LPG 算法、KK 算法时间复杂度低很多。

表 1 KK、HPG 以及 CDEA 算法不同阶段的时间复杂度

	启发阶段	贪心阶段
CPG 算法	$O(M'/N\times Q_1)\times k_1$	$O(N\times T_1)\times k_2$
HPG 算法	$O(M'/N\times Q_2)\times k_1$	$O(N\times T_2)\times k_2$
KK 算法	$O(N\times T_3)\times k$	

2 实验

本文实验中线性阈值模型中的每个节点阈值 θ 取固定值 0.5，启发因子 c 用于平衡不同阶段的步数，其中启发阶段为 $k-\lceil ck \rceil$ 步，贪心阶段为步 $\lceil ck \rceil$，当 $c=1$ 时表明此时为纯 KK 贪心算法。实验使用的数据集来自公开数据[8]，社区密度是每个社区平均所含节点数。数据集统计信息如表 2 所示。

表 2 数据集信息

序号	数据集名称	节点数	边数	平均度	社区数	社区密度
1	com-DBLP	317 080	1 049 866	6.622	13 477	23.53
2	com-Youtube	1 134 890	2 987 624	5.265	8 385	135.347
3	com-Orkut	3 072 441	117 185 083	76.281	6 288 363	0.488

第一个数据集是计算机类英文文献数据中的论文作者合作网络，边代表两个作者共同发表过论文，边上的权重值表示作者间的合作次数。第二个数据集是视频分享网站 Youtube 中的用户视频分享网，边代表用户间为彼此分享过视频，边上的权重值代表用户共享视频的次数。第三个数据集是 Google 公司推出的免费在线社交网站 Orkut 的朋友关系网，边代表用户间是朋友关系，边上权重值表示用户交流次数。

为了充分说明本文提出的基于社区度影响力最大化算法，实验在 3 个数据集上，从不同 k 值下的影响范围和算法运行时间两方面将本文提出的 CDEA 算法与 KK 算法、HPG 算法进行对比，验证算法在大规模社会网络下的准确性和高效性。

2.1 影响范围对比

首先考察 Youtube 数据集中 CDEA 算法在不同启发因子 c 下影响范围的变化，实验结果如图 1 所示。由图可知，除 $c=0$ 外，其他 c 值影响范围基本都比 $c=1$ 大，且 $c=0.5$ 时影响范围最大。由于 $c=1$ 时，CDEA 算法只有贪心阶段没有启发阶段，因此影响范围比其他 c 值的影响范围小。$c=0$ 表明此时 CDEA 算法只有启发阶段没有贪心阶段，所有初始节点都是静态选择 PI 最大的节点，没有传播过程参与，因此影响范围最小。实验结果表明 $c=$

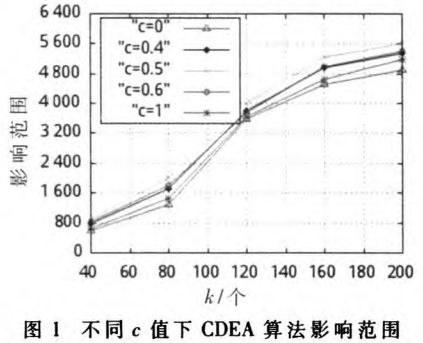

图 1 不同 c 值下 CDEA 算法影响范围

0.5 时 CDEA 算法影响范围最大，因此下面的实验中设 $c=0.5$。

其次考察 3 个数据集上 CDEA 算法影响范围的变化，实验结果如图 2 所示。由图可知相同 k 值下，Youtube 数据集上 CDEA 算法的影响范围最大，Orkut 数据集中影响范围最小，这说明本文提出的算法更适用于社区密度比较大的社会网络。随着初始种子节点 k 逐渐变大，CDEA 算法在 3 个数据集上影响范围随之扩大，且当 k 在[80，160]之间时影响范围的变化速率比较大，k 值超过 160 后影响范围变化速率逐渐减小，这是因为随着 k 的增大，新增加的种子节点能激活的节点不断减少，其影响范围也在降低。

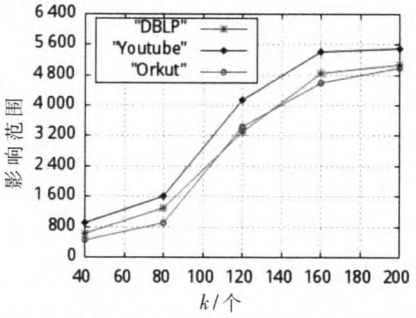

图 2 不同数据集上 CDEA 算法的影响范围

最后考察 Youtube 数据集中 KK 算法、HPG 算法、CDEA 算法在不同 k 值下影响范围的变化，实验结果如图 3 所示。由图可知，KK 算法的影响范围呈线性，HPG 和 CDEA 算法呈曲线上升，且 CDEA 算法在 k 值大于 120 时比 HPG 算法影响范围大，这是因为 CDEA 算法综合考虑了节点度、社区度以及社区规模均值 3 个因素，使影响传播的范围在大规模社会网络中更广。

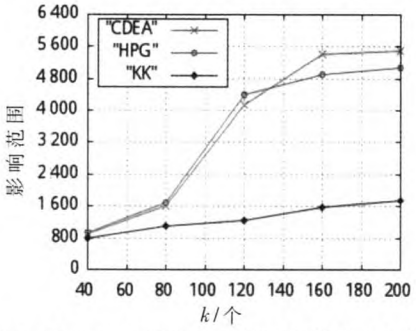

图 3 Youtube 数据集上不同算法的影响范围

2.2 运行时间对比

首先对比 3 个数据集上 CDEA 算法寻找 k 个种子节点需要的运行时间，实验结果如图 4 所示。由图可知，算法在 Youtube 数据集上运行时间最小，在 Orkut 数据集上运行时间最大，这是由于 CDEA 算法充分利用了节点的社区度属性，而 Youtube 数据集社区密度大，因此运行时间相对比较小。当 k 值不断变大时，CDEA 算法在不同数据集中运行时间的增长率比较小，因为该算法充分考虑了社会网络中社区的边界节点，更适合大规模社会网络。

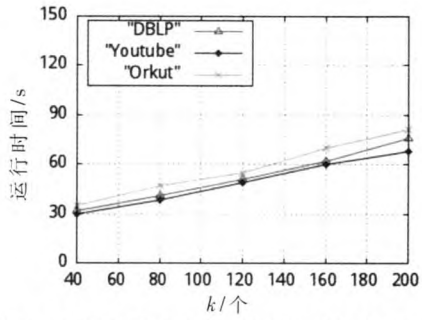

图 4 不同数据集上 CDEA 算法的运行时间

最后在 Youtube 数据集中比较 CDEA、HPG、KK 算法的运行时间。实验结果如图 5 所示。由图可知，随着 k 值的不断增长，KK 算法的运行时间不断变长，而 CDEA 和 HPG 算法的运行时间相对比较稳定，呈线性增长，且当 k 不断变大时，CDEA 算法的运行时间低于 HPG 算法。这是因为 CDEA 算法充分考虑了社区边界节点，尤其是在大规模社会网络中，极大地减少了寻找初始种子节点的时间。

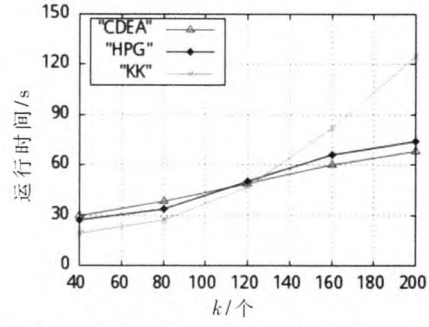

图 5 Youtube 数据集上 CDEA、HPG、KK 算法运行时间

3 总结

本文提出一种基于社区度的边界节点影响力最大化算法 CDEA，与其他关于影响力最大化问题研究不同的是：CDEA 算法利用社会网络的社区结构，将社区中的边界节点作为最有影响力节点的候选集，在两层算法模型框架下，启发阶段根据网络社区从边界节点中选择具有潜在影响力的节点集，贪心阶段在启发阶段基础上应用 KK 算法计算。实验表明，本文提出的算法既有效地降低了时间复杂度，又能较好地应用于大规模社会网络。接下来的工作将对算法作进一步改进，改善评估节点影响力的因素，提高算法的精度。

参考文献

[1] LI G，CHEN S，FENG J，et al.Efficient location-aware influence maximization[C].Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data.Snowbird，Utah，2014：1621.

[2] KIM J，KIM S K，YU H.Scalable and parallelizable processing of influence maximization for large-scale social networks[C].Proceedings of the 29th IEEE International Conference on Data Engineering.Brisbane，Australia，2013：

(下转第 151 页)

由于个体的差异，在进行基于 SSVEP-BCI 系统设计之前，需要对操控者进行刺激频率的选择。由于在频率域上每个操控者可用的频率并不是太多，为了实现多任务的操控目的，在实验中采用多频序列编码范式，其利用频率在时间尺度上的置换完成对 SSVEP-BCI 系统刺激模块的编码，是一种周期性的直接编码方案[6]。

多频序列编码原理如图 3 所示。

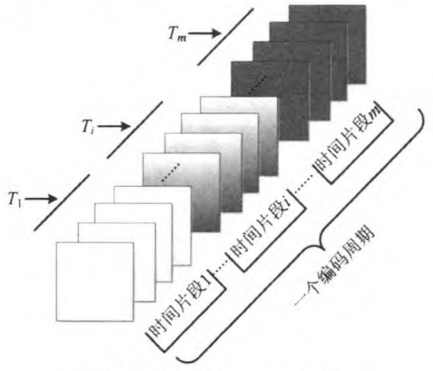

图 3 多频序列编码原理

2.1.2 MOVEP

MOVEP 的刺激形式如图 4 所示，4 号位的一条线从方框右边向左边快速移动，在本系统中，线从右到左的移动时间是 250 ms。若被试注意线起始出现的时刻，那么在后顶部位的电极处便可记录到 MOVEP 特征信号。

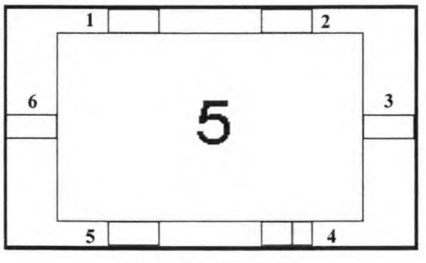

图 4 MOVEP 刺激

由于被试间的 N200 和 P200 的潜伏期存在差异，为了优化系统的性能，根据双样本 t 检验和 ANOVA 方法为每个被试选取最优的时间窗。在优化后的时间窗内，按一定的降采样率提取特征点。

MOVEP 时域信号特征如图 5 所示。

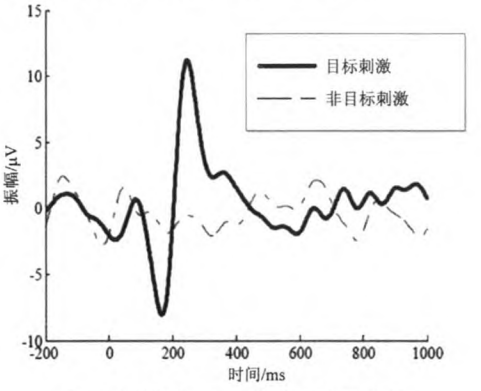

图 5 MOVEP 时域信号特征示意图

图 5 是从 P3 电极处采集的 EEG 信号经过叠加平均后的信号特征图，其中实线代表目标刺激的 EEG 信号，虚线代表非目标刺激的 EEG 信号。

2.2 实验结果

通过实验验证 SSVEP-BCI 系统目前达到的技术指标为：(1)能实现对屏幕上同时出现的 6 种目标进行区分选择；(2)控制精确度大于 80%；(3)锁定时间约 500 ms。

MOVEP-BCI 系统目前达到的技术指标为：(1)能实现对屏幕上同时出现的 6 种任务进行区分选择；(2)控制精确度大于 80%；(3)任务选择时间约 500 ms。

3 小结

本文使用 SSVEP 来快速选择打击目标，使用 MOVEP 进行任务分类，使用异步 MI 作为开关，实现自如的 EEG 与传统控制方式的转换，取代单纯靠目视来锁定目标的传统头盔。经过实验验证，该系统通过和机载火控雷达及专业任务系统的紧密配合，在对超高速、小型目标的快速锁定及脑控任务选择上具有较好的效果。

(下转第 155 页)

(上接第 148 页)

266-277.

[3] 赵之滢，于海，朱志良，等.基于网络社团结构的节点传播影响力分析[J].计算机学报，2014，37(4)：753-766.

[4] JUNG K，HEO W，CHEN W.IRIE：Scalable and robust influence maximization in social networks[C].Proceedings of the 12th International Conference on Data Mining.Brussels，Belgium，2012：918-923.

[5] GUO J，ZHANG P，ZHOU C，et al.Personalized influence maximization on social networks[C].Proceedings of the 22nd ACM International Conference on Information & Knowledge Management.San Francisco，USA，2013：199-208.

[6] CAO T，WU X，WANG S，et al.OASNET：an optimal allocation approach to influence maximization in modular social networks[C].Proceedings of the 2010 ACM Symposium on Applied Computing.ACM，2010：1088-1094.

[7] 田家堂，王轶彤，冯小军.一种新型的社会网络影响力最大化算法[J].计算机学报，2011，34(10)：1956-1965.

[8] YANG J，LESKOVEC J.Defining and evaluating network communities based on ground-truth[C].Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics.ACM，2012：3.

(收稿日期：2015-03-02)

作者简介：

王双(1989-)，男，硕士，主要研究方向：社会网络、数据挖掘。

李斌(1979-)，男，博士，讲师，主要研究方向：数据挖掘、传感器网络。

刘学军(1971-)，男，博士，教授，主要研究方向：数据库、数据挖掘、传感器网络。