基于有重叠社区划分的社会网络影响最大化方法研究

胡庆成 张 勇 邢春晓

(清华大学计算机科学与技术系 北京 100084) (清华大学信息技术研究院 北京 100084)

摘 要 社会网络中影响最大化问题是指在特定传播模型下,对于给定的值,寻找具有最大影响范围的节点集,这是一个组合优化问题,Kempe等人已经证明该问题是NP-hard问题,其研究在理论和现实应用中都具有重大意义。文中提出一种新的影响最大化算法——有重叠社区划分的影响最大化算法(*K*-clique Heuristic算法),该算法的思路是在现实社会网络中跨越多个社交圈子的节点的传播领域越广,其交叉性更强、传播范围更广、影响力更大。所提算法与已有典型算法有相近的运行结果,且有更好的现实应用性和可解释性,为这项具有挑战性的研究提供了新的思路和方法。

关键词 社会网络,影响最大化,信息传播,贪心算法,社区划分

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.005

K-clique Heuristic Algorithm for Influence Maximization in Social Network

HU Qing-cheng ZHANG Yong XING Chun-xiao

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

(Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

Abstract Influence maximization is the problem of obtaining a set of nodes with specified size in social network to maximize their aggregate influence under certain influence diffusion model, and it can yield significant benefit both in theory and real life. Influence maximization has been proved to be NP-hard by Kempe D et al. This paper proposed a new algorithm for influence maximization named K-clique Heuristic. The basic idea of the algorithm is that the nodes in social network spans multiple social circles. If these nodes are more widely spread in field and range, they have greater intersectionality and influence. The experimental results show that the proposed model is effective, and it may also shed light on the profound problem of influence maximization in social network.

Keywords Social network, Influence maximization, Information diffusion, Greedy algorithm, K-clique

1 引言

随着在线社交网络的蓬勃发展,以交友、信息分享等为目的的社交网络成为传播信息、表达观点、分享信息的理想平台,同时为影响最大化研究提供了真实的经验数据支撑,这种复杂的社会网络关系对信息传播和扩散有着至关重要的作用。影响最大化问题(Influence Maximization)通过分析人们相互之间的影响模式和影响力传播方式,既能从社会学角度加深对人们社会行为的理解,又能促进政治、经济和文化活动等领域的交流与传播,在理论和现实应用中具有重大意义。例如,其能有效地控制疾病的传播、流言的散布和计算机病毒的扩散,还可以传播新产品、新思想、新技术,以推进社会化进程。

影响最大化问题是在给定预算的前提下,选择 S 个初始传播种子节点,使得其最终的传播范围最大化。文献[1-2]给出了复杂网络中最有影响力节点的相关研究进展及各种方法的分析综述。Domingos[3]和Richardson[4]首先将影响最大化问题归纳为一个算法问题,主要是在社交网络中找出最有影响力的成员并为他们提供免费的样品,希望通过他们向网络中的其他成员推荐,从而达到营销的目的。这种通过口口相传(Word of Mouth)的影响力传播方式,能达到以最小的费用将新产品在整个网络中最大范围地进行推广的商业营销目标。Kempe等人(简称KKT算法)[5]形式化地表示了该问题,首次证明了复杂网络上的影响最大化问题的求解是一个NP-hard问题;并给出了与最优解的比为 $1-1/e\approx63\%$ 的近似贪心算法(Greedy Algorithm,GA),但其主要缺点是时间复

到稿日期:2017-03-11 返修日期:2017-06-02 本文受国家自然科学基金(91646202),教育部在线教育研究中心在线教育研究基金(2017YB142),千人计划、清华大学自主科研计划基金,863计划:心血管疾病大数据平台的构建和应用研究(SS2015AA020102)资助。

胡庆成(1977-),男,博士,主要研究方向为信息传播、大数据分析、复杂网络等,E-mail:huqingcheng@tsinghua.edu.cn(通信作者);**张 勇**(1973-),男,博士,副教授,主要研究方向为大数据管理与分析、数据科学;**邢春晓**(1967-),男,教授,博士生导师,主要研究方向为大数据、知识工程和数据科学。

杂度高、速度慢；同时，还给出了独立级联（Independent Cascade）模型和线性阈值（Linear Threshold）模型。Chen 等人[6]提出了两种改进的贪心算法 NewGreedy 和 MixGreedy，此外，还提出了一种改进的度数最大算法 DegreeDiscount。NewGreedy算法从原始网络中去掉对传播没有影响的边，得到一个小网络，然后在小网络中进行影响力传播，其优点是不需要每次都从整个网络进行考虑。DegreeDiscount 算法[6]是对探索式算法的一种优化策略的改进，在实验结果与贪心算法相近的情况下，运行效率有了很大提升。Kimura 等人[7]提出了一种通过分解极大强连通子图来寻找影响最大化的算法。基于此，文献[8]给出了一种基于用户间最大影响路径的方法，但其通过最短路径传播的假设限制性太强。Wang 等人[9]发现影响力的传播大多发生在社区之间，由此提出了一种贪心策略结合动态规划的算法来选取初始用户，较大程度地提升了算法的执行效率。Galstyan 等人[10]利用收益递减策略来研究市场收益最大化。Li 等人[11]提出了积极、消极影响最大化 PRIM 模型。Hu 等人[12]提出了随机节点最大度邻居方法（RMDN），其在具有较好的传播效果的同时，时间效率也相对较高。

现实社会中，信息总是在各类社交圈中传播，人们总是由于成长阶段、工作学习、兴趣爱好、生活地域等原因形成了一个又一个关系紧密的社交圈。这与我们社交网络形成的网络关系基本相符，如我们会有同学圈、同事圈、娱乐伙伴圈和学术交流圈等社交圈；我们认识的这些人之间可能相互认识，也可能不认识。从自己的角度来看，我们是联系这些人的 Ego 节点[13]。由此认为，最具有影响力的节点是能最大化联系整个社交网络并跨越多个圈子的人，类似于整个复杂网络中的结构洞（Structural Holes）[14-16]，这些圈子中的有重叠社区划分是研究信息传播的关键基础。基于此，本文提出了一种新的影响最大化模型——有重叠社区划分影响最大化算法（K-clique Heuristic 算法），该算法的基本思想符合现实生活中一个人会跨越多个领域的情况，传播范围更广，影响力更大。所提算法与已有典型算法拥有相近的运行结果，且有更好的现实应用性和可解释性。

本文第 2 节介绍了影响最大化研究的背景知识及实验传播模型；第 3 节介绍了本文的算法模型；第 4 节在 2 个实际社会网络中对各种算法进行了实验和分析比较；最后总结全文并给出未来的工作计划。

2　相关研究

在社会网络中，影响最大化问题可以帮助我们有效地控制疾病的传播、流言的散布和计算机病毒的扩散，还可以传播新产品、新技术、新思想，加快推进社会化进程。由于影响最大化问题是 NP-hard 的，因此设计新的算法模型一直是研究的重要内容。

为了表述方便，表 1 列出了文中需要的重要变量参数。

表 1　重要变量参数的说明

Table 1　Description of significant parameters

变量参数	描述	变量参数	描述
n	网络节点个数	m	网络边的个数
S	初始传播种子节点集合	T	最终被传播节点集合
s	被选中的节点数	$k_{\min}$	节点最小度数
p	传播概率	$k_{\max}$	节点最大度数
R	算法迭代循环次数	$P(k)$	节点度数是自然数 k 的概率

2.1　影响最大化问题的定义和贪心算法

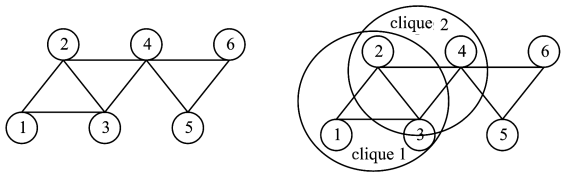
问题定义：给定网络 $G=(V,E)$ 和常数 $s\leqslant|V|$，V 为社会网络中的个体节点，E 代表个体之间的关系。如果初始传播的种子节点集合为 S，传播过程结束后预期的激活节点集合为 $T=\delta(S)$，找出节点集合 $S\subseteq V$ 且 $|S|=s$，使得传播范围 $\delta(S)$ 最大。

算法 1　GeneralGreedy(G,s)

1. initialize S=Ø and R=10000
2. for i=1 to s do
3.　for each vertex v ∈ V\S do
4.　　$s_v=0$
5.　　for i=1 to R do
6.　　　$s_v+=|\delta(S\cup\{v\})|$
7.　　end for
8.　　$s_v=s_v/R$
9.　end for
10. $S=S\cup\{\underset{v\in V\setminus S}{\operatorname{argmax}}\{s_v\}\}$
11. end for
12. output S

2.2　K-clique 有重叠社区划分算法

最早的有重叠社区划分是由 Palla 等人于 2005 年在 *Nature* 中提出的 K-clique 社区划分算法 CMP[17]。他们定义社区内的任意节点应与多数节点连接，即一个社区可以表示为多个子图的并集且要求子图的内部节点之间必须有联系。在数学上定义一个完全连接的全连接图为 K-clique，其中 K 为节点的数目，邻接 K-clique 是指两个 K-clique 至少共享 $K-1$ 个节点。图 1 给出了一个简单的 K-clique 社区网络划分图例。



(a)K-clique 网络关系例图　　(b)K-clique 网络关系例图中 $K=3$ 的部分社区划分

图 1　K-clique 社区网络划分图例

Fig. 1　K-clique community division in complex networks

2.3　影响力的传播模型

社会网络中每个节点有两种状态，即激活状态和未激活状态：若一个节点已经接受了信息，则其为激活节点，否则为非激活节点。激活节点对于未激活节点存在影响，即一个节点的邻居激活节点越多，则该节点被激活的可能性就越大；新激活节点又会影响其他处于未激活状态的邻居节点。在网络环境中，最主要的交互活动是信息的发布、分享和扩散，因此

影响力在社会网络中的作用过程和信息的扩散过程有内在的紧密联系和十分相似的机制。传播模型在影响力传播问题的研究过程中发挥着非常重要的作用,独立级联模型和线性阈值模型是信息传播过程进行建模的重要方法。

2.3.1 独立级联模型

独立级联模型即 IC 模型(Independent Cascade Model)[5],是基于相互粒子系统设计的一个信息扩散模型,可以描述为:在复杂网络 $G=(V,E)$ 中,对于 V 的每一个顶点 u 和它的邻居节点 v,存在一条连边 $e(u,v)$,p_{uv} 表示在传播过程中 u 对邻居节点 v 的影响力概率,p_{uv} 的取值是独立的。如果 u 在 t,u 时刻是激活状态,并且其邻居 v 是未激活的,那么 u 将尝试以概率 p_{uv} 去激活 v。如果激活过程成功,那么在 $t+1$ 时刻 v 就成了激活状态。但是无论成功与否,u 不能再试图去激活 v。如果 v 在 t 时刻有多个邻居都处于激活状态,则它们尝试激活 v 的顺序是任意的。系统从初始态开始传播,直到没有新的节点可以被激活为止。

2.3.2 线性阈值模型

线性阈值模型即 LT 模型(Liner Threshold Model)[5],是诸多阈值模型的核心,可以描述为:在复杂网络 $G=(V,E)$ 中,定义 $N(u)$ 为节点 u 的邻居节点集合。被激活的节点 u 对邻居节点 v 存在影响 b_{uv},一个节点 u 的所有邻居节点对 u 的影响力总和小于或等于 1,即:

$$\sum_{v\in N(u)} b_{uv}\leqslant 1$$

每个节点 u 有一个特定阈值 $\theta_u\in[0,1]$,如果 $\sum_{v\in N(u)} b_{uv}\geqslant\theta_u$,则 u 被激活。LT 模型中,当一个激活节点 u 尝试激活它的未激活邻居 v 但没有成功时,节点 u 对节点 v 的影响力 b_{uv} 被积累起来,这种积累对后面其他邻居节点激活 v 是有贡献的,直到节点 v 被激活或传播过程结束。这就是 LT 模型的"影响积累"特性,因此 LT 模型与 IC 模型是不同的。

3 算法模型

社区结构的重叠性普遍存在于各种社交网络中,网络中的节点可以同时属于多个派系,通常认为最具有影响力的节点是能最大化联系整个社交网络的节点,特别是能跨越多个社交圈子的节点,这些节点能在不同领域中传播信息,类似于整个社交网络中的结构洞。我们所提出的 K-clique Heuristic 算法的基本思路是:首先对整个社交网络进行 K-clique 划分,然后找出在所有有重叠划分的社区中出现次数最多的 Top-K 个节点集合,其中 K 是控制社区内部紧密程度的参数。所提算法的具体描述如算法 2 所示。

算法 2 K-clique Heuristic(G,s)

```
1. initialize S=Ø,dict=[]
2. KC=KCliqueCommunity(G,k)
3. for each vertex v ∈ V\S do
4.    for each clique kc_i∈KC do
5.       if v ∈kc_i then
6.          dict[v]+=1
7.       end if
8.    end for
9. end for
10. S=argmaxTop(dict[v])[1:s]
11. Return S
```

4 实验与结果分析

实验运行的硬件环境如下:处理器 Intel® Core™ i5 CPU M430 @2.27GHz,内存(RAM)4GB。

4.1 实验数据

鉴于不同类型的社会网络通常具有相似的网络结构特征,选取两个实际社会网络进行实验分析比较。表 2 列出了各个网络的属性特征:1)Blogs 网络数据[18],即 MSN 博客空间中交流的关系网络;2)Netscience 合著关系网络[19],选择了其中的最大连通子图作为 379 个作者的关系网络。

表 2 现实社会网络的属性情况

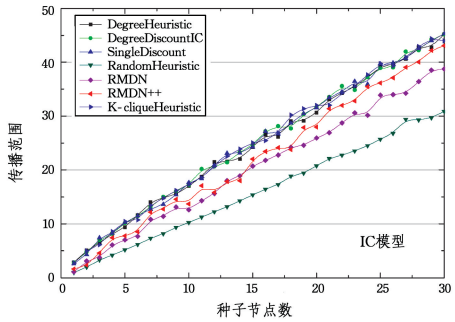
Table 2 Attributes of real social newtork

Network Name	n	m	$\langle k\rangle$	$k_{\max}$	$k_{\min}$	d
Blogs	982	6803	3.42	189	1	6.227
Netscience	379	914	4.82	34	1	6.061

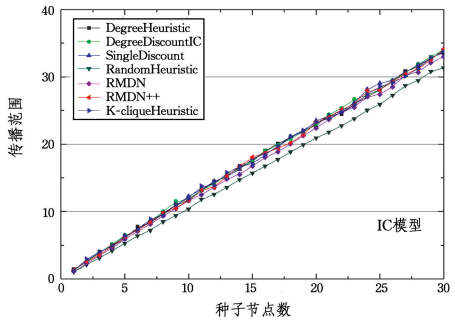
其中,n 是网络中的节点数,m 为边数,$\langle k\rangle$ 表示网络中的平均度数据,$k_{\max}$ 为节点中的最大度数,$k_{\min}$ 为节点中的最小度数,d 为节点之间最短路径的平均数。

4.2 实验效果

我们应用 IC 模型和 LT 模型对本文所提出的 K-clique Heuristic 算法与已有的典型算法在 2 个真实社会网络中的传播范围进行了分析比较。为了保持实验的易读性,在实验模拟传播过程中,每次选取传播种子节点 $s(1\leqslant s\leqslant 30)$ 作为传播源的种子集合,传播概率 $p=0.01$(如果节点的传播能力很强,很难区分单个个体的重要性),传播范围取 10000 次迭代的均值,实验结果如图 2、图 3 所示,其中横轴为根据各算法中传播影响力最大的节点进行排序所得的集合大小,纵轴为所选传播种子节点的相应传播范围大小。且复杂度分析统一取 $s=30$,时间取迭代运行 10000 次的平均运行时间。



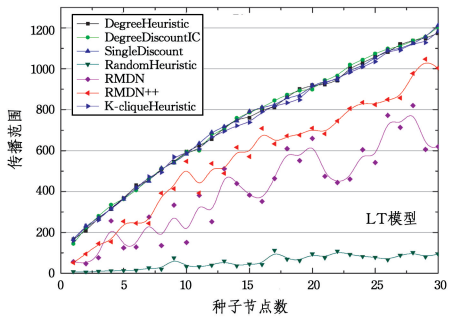
(a)Blogs 网络



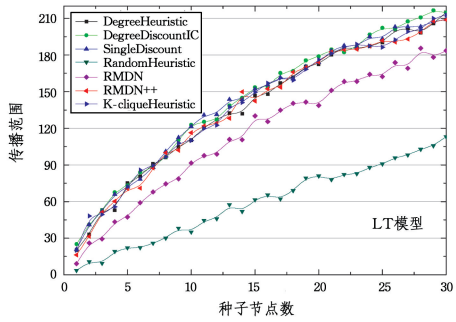
(b)Netscience 网络

图 2 IC 模型下不同算法的运行结果

Fig. 2 Performance of different algorithms under IC model



(a)Blogs 网络



(b)Netscience 网络

图 3　LT 模型下不同算法的运行结果

Fig. 3　Performance of different algorithms under LT model

从图 2 和图 3 可以看出，所提出的 K-clique Heuristic 算法在 IC 模型与 LT 模型中的运行效果与其他算法相似，说明我们的模型同样可以应用于其他类型的社会网络，而且算法的适应性较好。与最初的设想即最具有影响力的节点是那些能最大化联系整个社交网络社区的节点基本一致。

结束语　在现实社会中，影响最大化问题有助于扩大新知识、新产品的有效传播范围，同时也可以有效地预测、分析和控制疾病的传播、流言的散布及计算机病毒的扩散。在给定的有限预算前提下，在社会网络中找出影响最大化传播种子集合一直是研究的热点与难点。为此，我们提出了一种新的影响最大化模型——有重叠社区划分影响最大化算法(K-clique Heuristic 算法)，该算法的思路是在现实社会网络中跨越多个社交圈子的节点能把信息传播到不同领域，这些节点的传播范围更广，传播方式也更具多样性，影响力更大。所提算法与已有典型算法有相近的运行结果，且有更好的现实应用性和可解释性，为这项具有挑战性的问题提供了新的思路和方法。

参 考 文 献

[1] REN X L, LV L Y. Review of ranking nodes in complex networks[J]. Chinese Science Bulletin, 2014, 59(13), 1175-1197. (in Chinese)
任晓龙，吕琳媛. 网络重要节点排序方法综述[J]. 科学通报，2014，59(13)：1175-1197.

[2] LIU J G, REN Z M, GUO Q, et al. Node importance ranking of complex network[J]. Chinese Journal of Physics, 2013, 62(17): 178901. (in Chinese)
刘建国，任卓明，郭强，等. 复杂网络中节点重要性排序的研究进展[J]. 物理学报，2013，62(17)：178901.

[3] DOMINGOS P, RICHARDSON M. Mining the network value of customers[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2001: 57-66.

[4] RICHARDSON M, DOMINGOS P. Mining knowledge-sharing sites for viral marketing[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002: 61-70.

[5] KEMPE D, KLEINBERG J, TARDOS E. Maximizing the spread of influence through as ocial network[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003: 137-146.

[6] CHEN W, WANG Y, YANG S. E cient influence maximization in social networks[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 199-208.

[7] KIMURA M, SAITO K. Tractable models for information diffusion in social networks[M]// Knowledge Discovery in Databases (PKDD 2006). Springer, 2006: 259-271.

[8] CHEN W, YUAN Y, ZHANG L. Scalable influence maximization in social networks under the linear threshold model[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 2010: 88-97.

[9] WANG Y, CONG G, SONG G, et al. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 1039-1048.

[10] GALSTYAN A, MUSOYAN V, COHEN P. Maximizing influence propagation in networks with community structure[J]. Physical Review E, 2009, 79(5): 056102.

[11] LI D, XU Z M, CHAKRABORTY N, et al. Polarity related influence maximization in signed social networks[J]. PloS one, 2014, 9(7): e102199.

[12] HU Q, ZHANG Y, XU X, et al. RMDN: New Approach to Maximize Influence Spread[C]// 2015 IEEE 39th Annual Computer Software and Applications Conference. IEEE, 2015: 702-711.

[13] MARSDEN P V. Egocentric and sociocentric measures of network centrality[J]. Social Networks, 2002, 24(4): 407-422.

[14] WALKER G, KOGUT B, SHAN W. Social capital, structural holes and the formation of an industry network[J]. Organization Science, 1997, 8(2): 109-125.

[15] AHUJA G. Collaboration networks, structural holes, and innovation: A longitudinal study[J]. Administrative Science Quarterly, 2000, 45(3): 425-455.

[16] BURT R S. Structural holes: The social structure of competition [M]. Cambridge: Harvard University Press, 2009.

[17] PALLA G, DERE′NYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[18] XIE N. Social network analysis of blogs[D]. United Kingdom: University of Bristol, 2006.

[19] BOCCALETTI S, LATORA V, MORENO Y, et al. Complex networks: Structure and dynamics[J]. Physics Reports, 2006, 424(4): 175-308.