基于社区划分的影响力最大化算法

王 双,李 斌,刘学军,胡 平
WANG Shuang, LI Bin, LIU Xuejun, HU Ping

南京工业大学 电子与信息工程学院,南京 211816
College of Electronic and Information Engineering, Nanjing Tech University, Nanjing 211816, China

WANG Shuang, LI Bin, LIU Xuejun, et al. Division of community-based influence maximization algorithm. Computer Engineering and Applications, 2016, 52(19):42-47.

Abstract: Influence maximization is a significant research direction in social networks. Its main purpose is to get the most influential users to make the range of influence diffusion maximizing. With the coming of big data, the traditional greedy algorithm can not overcome the time problem of influence maximization effectively because of high time complexity for large-scale social networks. This paper proposes the community division to solve influence maximization. Large-scale social networks are divided into smaller community modules using influence probability. Thus, the isolation between sub-communities is also eliminated at greatest degree considering the boundary nodes between communities. To improve the efficiency further, this paper considers an independent influence path as an influence evaluation unit in each community. At the same time, the most influential nodes are found utilizing parallel processing at every community. Finally, the paper verifies the feasibility and efficiency of the proposed algorithm by simulation experiment, which can adapt to the large-scale social networks better.

Key words: social network; influence maximization; community division; influence diffuse

摘 要:影响力最大化问题是社会网络中的重要研究方向,其主要目的是获取社会网络中最有影响力的用户使通过这些用户获得影响传播范围的最大化。随着大数据时代的来临,传统的贪心算法因为复杂度高而不能有效解决大规模社会网络下影响力最大化的时间问题。提出一种基于社区划分的影响力最大化算法,利用影响概率将大规模社会网络分成较小的社区模块,并考虑社区边界节点之间的联系,从而最大程度缩小因社区划分造成的社区间的孤立。为进一步提高算法效率,在每个社区中以影响路径作为影响评估单元,同时对每个社区并行处理以便更高效地获取有影响力的节点。通过仿真实验验证了算法的可行性和高效性,其可以较好地适应大规模社会网络环境。

关键词:社会网络;影响力最大化;社区划分;影响传播
文献标志码:A **中图分类号:**TP393 doi:10.3778/j.issn.1002-8331.1411-0171

1 引言

近年来,互联网和Web2.0技术的发展为信息生成和传播带来了巨大变化,推动了社会网络发展,如国外的大型社交网站Facebook、Twitter以及国内的微博、人人、QQ等。各种社会网络服务不断出现,已渗透到生活的各方面,成为人类沟通交流、共享信息和传播信息的重要媒介和平台。社会网络在信息、观点、创新的传播扩散过程中发挥了基使础性作用,而随着用户数目的持续增长,社会网络的规模 快速扩大,其核心作用越来越显著,引起了广泛的积极研究。其中影响最大化问题作为社会网络分析领域的关键问题之一,其目的是在社会网络所有节点中选择最具影响力的节点作为初始活跃节点,使得经过影响传播,社会网络中最终被影响的节点个数最多,由此可以广泛应用在广告投放[1]、水质监

基金项目:国家公益性科研专项(No.201310162);连云港科技支撑计划项目(No.SH1110)。

作者简介:王双(1989—),男,硕士研究生,研究领域为社会网络、数据挖掘,E-mail:shuang.you.good@163.com;李斌(1979—),男,博士,讲师,研究领域为数据挖掘、传感器网络;刘学军(1971—),男,博士,副教授,研究领域为数据库、数据挖掘、传感器网络;胡平(1962—),男,博士,教授,研究领域为远程教育、计算机智能。

收稿日期:2014-11-15 **修回日期:**2015-01-03 **文章编号:**1002-8331(2016)19-0042-06
CNKI **网络优先出版:**2015-05-29, http://www.cnki.net/kcms/detail/11.2127.TP.20150529.1606.019.html

万方数据

测、疫情监控等重要场景中。例如基于在水质监测和疫情监控中，需要定位在哪些地点进行水质监控和疫情监控，才能使监控范围最大化，及时发现水质污染和疫情爆发。影响力最大化问题的求解直接影响到市场营销、水质监测等应用策略的制定和部署，对系统的有效性、可扩展性等方面起着重要作用。

随着社会网络技术的发展，当前的社会网络规模越来越庞大，具体体现为节点数量多，节点之间关系复杂。同时社会网络动态性越来越强，节点数目以及节点间关联关系的变化频繁，随机性强，难以预测。社会网络的这些特点直接导致挖掘网络中最有影响力节点时计算量大，运行时间长。随着社会网络规模的不断扩大以及大数据时代的来临，迫切需要研究大规模社会网络下解决影响力最大化问题更高效的方法。基于此，提出基于新型社区划分的影响力最大化算法，通过影响传播概率划分社区，在社区划分的基础上通过独立影响路径在社区中寻找k个最有影响力的节点，同时通过并行处理实现社区中挖掘k个节点的并行性，提高算法的高效性，有效改善大规模社会网络中挖掘最有影响力节点的效率。

2　相关研究

社会网络中影响力最大化问题就是如何选择 k 个种子节点进行传播，从而使最终影响传播范围最大。近几年此问题吸引了众多学者的研究。Li[2]等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值选择种子节点，并消除不重要的节点。Luo 等[3]在非统一的网络中基于幂律法则的影响力分布提出具有高 PageRank 的启发式算法，此算法通过在小部分仅仅包含很高 PageRank 的集合中寻找种子节点。Chen 等[4]将社区结构和度中心性信息整合，提出 CDH-Kcut 和 CDH-SHRINK 两个算法，并验证了算法在 HDM 网络影响传播范围的最大化。Zhu 等[5]基于影响传递性和有限的传播距离提出了基于半定规划的算法，该算法即使在限制种子节点的情况下也可以获得比较高的影响传播范围。Goyal 等[6]从基于数据的角度研究影响力最大化，提出了信用分布模型，通过可用的传播轨迹表示影响力是如何在网络流动的，并用该模型评估期望的影响传播范围，同时还研究了用户影响能力的等级。Chen 等[7]人给出了两个有效算法，明显减少了有影响的候选节点，虽然同样没有克服忽略社区间联系的缺点，但在保证影响效果的前提下，使算法效率比贪婪算法提高了一个数量级。文献[8]基于 IC 模型将独立影响路径作为影响评估单元，并通过增加简单的 OpenMp 程序以实现影响力最大化的可扩展性和并行性。陈浩[9]利用线性阈值提出了基于节点激活阈值的启发式算法，综合考虑了节点的激活阈值和节点间的影响力，将影响力最大化问题求解分为启发阶段和贪心阶段。Galstyan 等[10]第一次提出基于社区结构的影响力最大化问题的解决方案，但该方法局限于两个连接稀疏的社区网络，而实际中的网络一般包含很多社区网络。Wang 等[11]提出一种基于社区发现求解影响力最大化问题的 CGA 算法，首先选择了一个效率较高的社区发现算法，然后对这个社区发现算法进行改进，提出利用组合熵的方法来整合较小的社区，从而使社区更加均匀。该算法性能与贪心算法比有很大数量级的提高，缺点是时间复杂度高和由于网络分割使边损失造成的效率降低。Cao 等[12]把具有社区性质的社会网络中的影响力最大化问题当作资源分配问题解决，并给出了动态规划算法（OASNET）解决最佳资源分配的问题，但此算法没有考虑社区间的连通性。郭进时等[13]提出了基于社区结构的影响力最大化算法，并将影响传播范围和影响传播时延作为节点影响力的衡量标准，但是文章并没有对社区结构作详细描述，只是基于已经划分好的社区结构。冀进朝等[14]利用已有的社区挖掘算法划分社区结构，用迭代选择跨越社区最多的 k 个节点作为影响的初始传播点，在小型网络和中等规模网络数据集上用实验表明算法优势，但是对于大规模社会网络效果并不明显。

与以上研究不同，本文先基于影响概率对社区划分，然后基于改进的 InfG[13]传播模型在每个社区中利用影响路径寻找最有影响力的节点，同时考虑到社区间连接每个社区的边界节点，并按照比例从中寻找部分最有影响力的节点。最后将社区中最有影响力的节点和边界节点集中最有影响力的节点整合，缩小局部影响最大化和全局影响最大化的差距。

3　基于社区划分的影响力最大化算法

3.1　新型社区划分算法

研究社会网络首先需要为社会网络建模，以模拟其实际的行为。本文采用有向带权图 $G(V, E, W)$ 描述社会网络，V 为社会网络中所有用户的集合，V 集合中的节点表示用户，E 为用户间关系集合即为边集合，边表示相连的两个用户间有关联，W 为用户间紧密程度的集合，在图中用权值 W 表示。本文中的节点和用户是一个概念。详细的描述如定义 1。

定义 1（社会网络形式化建模）社会网络用图 $G=(V, E, W)$ 建模，其中 $V=\{v_1, v_2, \cdots, v_n\}$ $(n=|V|)$ 为社会网络中所有个体或节点的集，$E=\{e_1, e_2, \cdots, e_n\}$ $(n=|E|)$ 为社会网络 G 中连接节点与节点之间的边的集，$W=\{w_1, w_2, \cdots, w_n\}$ $(n=|W|)$ 为社会网络 G 中每一条边的权值集合。设网络中的社区个数为 M，则有社区集合为 $C=(C_1, C_2, \cdots, C_M)$。

定义 2（影响概率）用影响概率表示一个节点对其邻居节点的影响可能性。

万方数据

$$P_{v_i \to v_j} = \frac{w_{ij} + T_{ij}}{w_{\max} + T_{\max}} \tag{1}$$

其中，T_{ij} 为节点 v_i 对 v_j 的影响次数，此参数可以通过信息传播日志获得。$T_{\max}$ 是社会网络 G 中节点间的最大影响次数，即 $T_{\max} = \max(T_{ij})((i,j) \in E)$。$w_{ij}$ 是社会网络 G 中节点 v_i 到 v_j 的边权值，$w_{\max}$ 是整个网络中的最大边权值。

定义 3（社区标签）根据影响概率，如果 $P_{v_i \to v_j} \leqslant \theta$（$\theta$ 为已设好的阀值），则说明节点 v_i 没有足够的能力影响其邻居 v_j，因此为节点 v_i、v_j 分配不同的社区标签 $C(v_i)$、$C(v_j)$，且 $C(v_i) \neq C(v_j)$；否则，如果 $P_{v_i \to v_j} > \theta$，则 $C(v_i) = C(v_j)$。

$$\begin{cases} C(v_i) - C(v_j) \neq 0, P_{v_i \to v_j} \leqslant \theta \\ C(v_i) - C(v_j) = 0, P_{v_i \to v_j} > \theta \end{cases} \tag{2}$$

定义 4（节点社区属性）节点社区属性主要用于标记一个节点是否已经包含于某个社区，初始状态时所有节点的社区属性为 0，即所有节点都不属于某一个社区。一旦节点 v 被分配了社区标签 $C(v)$，则将其社区属性设为 1，表示节点已经存在于某一个社区中。

$$\begin{cases} v(C) = 1, C(v) = \text{True} \\ v(C) = 0, C(v) = \text{False} \end{cases} \tag{3}$$

定义 5（边界节点标签）为网络 G 中的相邻节点分配边界节点标签，初始值为 0，表示节点不是社区间的边界节点，当 $P_{v_i \to v_j} \leqslant \theta$ 时相邻的节点 v_i、v_j 具有不同社区标签，此时 v_i、v_j 为两个不同社区的边界节点，并将其边界标签设置为 1；否则将边界标签设为 0，表示节点不是边界节点。

$$\begin{cases} B(v) = 1, P_{v_i \to v_j} \leqslant \theta \\ B(v) = 0, P_{v_i \to v_j} > \theta \end{cases} \tag{4}$$

根据两个相邻节点 v_i、v_j 间的影响概率，如果影响概率 $P_{v_i \to v_j}$ 小于等于 θ，则为相邻的两个节点 v_i、v_j 分配不同的社区标签 $C(v_i)$、$C(v_j)$，否则分配相同的社区标签（即 $C(v_i) = C(v_j)$）。与此同时，当邻居节点 v_i、v_j 间的影响概率 $P_{v_i \to v_j}$ 小于等于 θ 时，认为相邻节点 v_i、v_j 为两个不同社区的边界节点，为其分配边界节点标签 $B(v_i)$、$B(v_j)$，且 $B(v_i) = B(v_j) = 1$。然后分别判断相邻节点 v_i、v_j 邻居的 $P_{v_i \to v_j}$ 和节点社区属性 $v(C)$，如果 $P < \theta$ 且 $v(C) = 0$，则为其分配一个不同的社区标签并将边界节点设为 1，否则将其加入已有的社区并设置社区属性 $v(C)$ 为 1。如此重复遍历相邻节点 v_i、v_j 的所有邻居，直到社会网络中所有节点的社区属性 $v(C) = 1$ 停止遍历，然后将节点社区标签 $C(v)$ 相等的节点社区统一为 C_k ($k \in [1, M]$)，最后形成社区集合 $C = (C_1, C_2, \cdots, C_M)$。社会网络图 G 中如果节点 v 的社区标签 $C(v) = 0$，说明此节点还没有社区标签即此节点还没有被遍历到，继续遍历，如果 G 中 $C(v) \neq 0 (\forall v \in V)$ 则说明所有的节点都已经被遍历过，社区划分过程结束。本文提出的这种社区划分方式可以有效地解决以往社区划分算法中的社区重叠问题，充分考虑了影响力的传播，更有利于影响力最大化问题的解决，也可以通过阀值 θ 控制每个社区的相对大小，θ 越大则社区越多，θ 越小则社会网络中的社区个数越小，即社区划分越碎。

基于影响概率的社区划分算法 CBIP（Community-Based Influence Probability）详细描述：

输入：社会网络 $G = (V, E, W)$，阀值 θ。

输出：$C = (C_1, C_2, \cdots, C_N)$。

1. for $v_i \in V$ do
2. 计算社会网络 G 中节点 v_i 与其邻居节点 v_j 间的影响概率 $P_{v_i \to v_j}$
3. 为节点 v_i 分配社区属性 $v_i(C)$ 和边界节点标签 $B(v_i)$，初始值均为 0
4. end for
5. 随机函数 RANK() 在 G 中选择两个相邻的节点 v_i、v_j
6. if $(P_{v_i \to v_j} < \theta)$
7. $v_i \leftarrow C(v_i)$，$v_j \leftarrow C(v_j)$，且 $C(v_i) \neq C(v_j)$
8. $B(v_i) = B(v_j) = 1$，$v_i(C) = v_j(C) = 1$
9. 对 v_i、v_j 分别执行（9）~（17）
10. for v 的邻居 $nei(v)$
11. for $u \in nei(v)$
12. if $(u(C) = 0)$
13. if $(P_{v \to u} > \theta \quad C(u) = C(v))$
14. else 对 $\forall v \in V$，$v(C)$ 是否为 1，如果不是对节点 u 从第 6 步执行
15. else break
16. end for
17. end for
18. else
19. $v_i(C) = v_j(C) = 1 v_i \leftarrow C(v_i), v_j \leftarrow C(v_j)$, 且 $C(v_i) = C(v_j)$，
20. 对 $nei(v_i)$、$nei(v_j)$ 执行第 9~17 步
21. 将 $C(v)$ 相同的节点整合到社区 C_k ($k \in M$)，形成社区集 $(C_1, C_2, C_3, \cdots, C_M)$

CBIP 社区划分算法，针对大规模社会网络划分，可在较短时间内按照要求将整个社会网络划分成具有 M 个社区的子网络，同时通过边界节点保留社区间的联系，以确保各个社区不是孤立的。一个简单的例子如图 1，虚线框表示一个社区，节点间的数字表示影响概率，红色节点表示边界节点，边界节点是与其他社区联系的重要纽带，其中社区间的边际节点影响概率比较小，而社区内的节点彼此间的影响概率较大。

万方数据

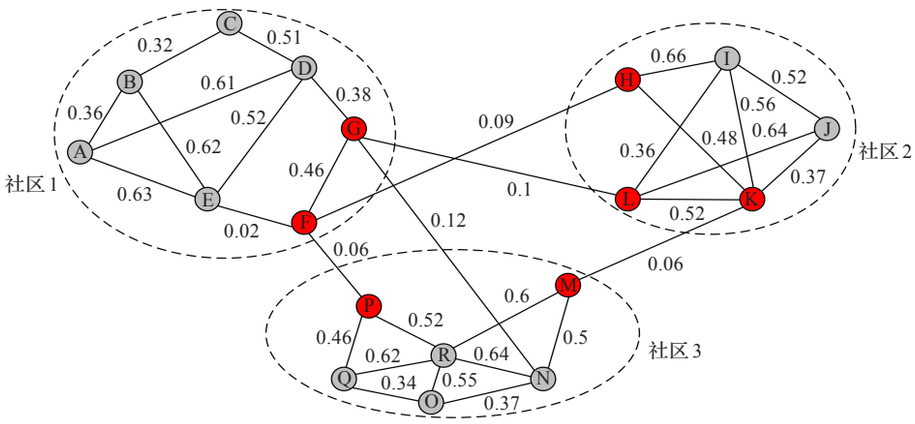

图1　具有三个社区的社会网络划分结构

3.2　基于社区的影响力最大化算法

3.2.1　CBI传播模型

CBI(Community-Based Influence)传播模型是基于InfG[13]的改进传播模型,相比InfG传播模型,CBI简化了社区内节点和边界节点的选取比例。众所周知,社会网络中社区间的连接具有稀疏性,即社区间的边界节点之间联系相对较少,但是缺少这些边界节点间的联系,社区间的信息将无法传播,所以CBI传播模型不仅考虑了社区内节点的影响同时也考虑了连接社区的边界节点的影响,因此在整个社会网络中寻找 k 节点时, $k-k'$ 节点从边界节点集 B_C 中选取, k' 节点从社区内选择,其中 $k-k'$ 节点的选取采用IPA贪心算法。

定义6(有效节点集) 基于CBIP社区划分算法,已知网络 $G=(C_1, C_2, \cdots, C_M)$ 具有 M 社区,每个社区中的最有影响力节点集为 F_{C_w},连接社区间的边界节点集为 $B_{C_w}, w\in[1, M]$。定义网络的有效节点集 $Valid$ 为网络 G 的所有最有影响力的节点集与边界节点的并集:

$$Valid=\bigcup_{w=1}^{M}(F_{C_w}\cup B_{C_w}) \tag{5}$$

每个社区中的有影响节点集 F_{C_w} 对社区内的节点影响比较大,而边界节点集 $B_{C_w}(w\in[1, M])$ 对社区内的节点影响比较小,但它是社区间的纽带,不容忽视,虽然影响小,但是对整个社会网络的影响传播起着至关重要的作用。影响力最大化问题就是如何选择 k 节点集使影响力传播的范围最大,针对此模型,则是在 $Valid$ 中选择 k 节点集,使经过这 k 节点产生的影响最大。

设 k 为初始影响节点集合 S 大小, k'_w 为社区 C_w 中最有影响力的节点数,即 $\left|F_{C_w}\right|=k'_w$, b 为边界节点集合 B_C 中的最有影响力的节点数。每个社区中最有影响力的节点数和边界影响力节点数满足:

$$\begin{cases}\left|C_1\right|:\left|C_2\right|:\cdots:\left|C_M\right|=k'_1:k'_2:\cdots:k'_M\\ k'_1+k'_2+\cdots+k'_M=k'=k-b\end{cases} \tag{6}$$

3.2.2　基于CBI传播模型的CMA算法

传统的影响力最大化问题,为了寻找含有 k 个种子节点的初始集 S,需要计算整个社会网络中所有节点的边际影响传播收益,虽然利用影响最大化目标函数的子模特性,大大降低了每轮节点边际影响收益的计算个数,避免了计算所有节点的边际影响传播收益,提高了计算效率。但是对于在13万节点的数据集上寻找最有影响力的50个节点仍需要花费数小时,难以满足当今的大规模社会网络。本文提出CMA(Community Maximization Algorithm)算法,以影响概率作为基本衡量因素,将一条影响路径作为影响传播范围的评估单元,为了有效控制节点影响路径的数目,通过选择节点边中影响概率最大的作为有效路径,这样只要计算头节点的影响路径,则完成了此条路径所有节点的影响路径。为了保证影响路径的传播有效性并控制长度的无限增长,当影响概率小于一定概率时中断影响路径的探寻。

定义7(影响传播范围) 将节点的影响传播路径作为节点影响传播范围的评估单元,用节点的影响传播路径表示影响传播范围,则节点 v 的影响传播范围 $R(v)$ 定义为:

$$\begin{cases}ip(v)=<v_1=v, v_2, \cdots, v_t=u>, t\geqslant 2\\ P_{v\to g}=\max(P_{v\to v'})>3\theta, v', g\in nei(v)且g\in ip(v)\\ R(v)=comp(ip(v))\end{cases} \tag{7}$$

$ip(v)$ 表示从节点 v 出发的影响路径,包含 t 个节点, $ip(v)$ 必须满足以下几个条件:(1) $B(v)=0$,即节点 v 为非边界节点;(2)路径中相邻的 $ip(v)$ 节点(如 $<v_2, v_3>$)是邻居;(3)相邻节点间的影响概率为节点与其邻居节点间的最大影响概率,即 $P_{v\to g}=\max(P_{v\to v'})(v', g\in neighbor(v))$ 且 $P_{v\to g}>1.5\theta$;(4) $ip(v)$ 影响路径中最后一个节点 u 与邻居节点间的最大影响概率 $P_{u\to v_{m-1}}>1.5\theta$, u 与其他邻居节点间影响传播概率小于 1.5θ。在影响路径 $ip(v)$ 中 v_2 是 v_1 的邻居, v_3 是 v_2 的邻居,依次类推。max()函数主要用于计算邻居节点中的最大影

万方数据

响概率；$nei(v)$ 表示节点 v 的邻居；$R(v)$ 表示节点 v 的影响传播范围，其值为影响路径中包含的节点数。

基于社区划分算法CBIP，CMA算法首先根据社区大小确定每个社区中需要挖掘的有影响力节点数 k'_i，其次对每个社区并行挖掘 k'_i 个有影响力节点，然后在每个社区内根据影响传播范围依次选择 k'_i 个节点。最后在全局社会网络中根据IPA[32]算法在边际节点集 B_C 中寻找 b 个最有影响力的节点集合 S'。

CMA算法具体步骤如下：

输入：网络 $G=(V,E,W)$，社区结构 $G=(C_1,C_2,\cdots,C_M)$。

输出：包含具有最大影响力的 k 初始传播节点集合 S。

1. 根据公式(7)计算 $k'_1,k'_2,\cdots,k'_M$
2. Parallel for $(i=1 \text{ to } M)$ do
3. for $(v \in C_i)$
4. if $(B(v)=0)$ then $v_0 \leftarrow \max(Degree(v)), S=\{v_0\}$
5. end for
6. for $(j=1 \text{ to } k'_i)$ do
7. 在 C_i 中选择最大影响传播范围的节点加入 S
8. 在未标记的节点中继续选择下一个节点以获得最大影响度
9. $j=j+1$
10. end for
11. end for
12. 根据IPA算法在 B_C 中寻找 b 个最有影响力的节点集 S'
13. 输出节点集 $S=S+S'$

4　实验

本文在Matlab7.1中编写程序仿真实验，主要采用基于MapReduce的分布式并行计算原理从影响传播范围、运行时间分析算法的性能，以验证CAM算法在当今大规模社会网络中的可行性和高效性。实验中的数据集主要通过微博API用网络爬虫程序获取，共获取32万个用户的信息，包含786万条边，平均聚类系数为1/16，节点的平均度为9，并建立网络拓扑结构图。通过社区划分算法将32万用户分成8 763个社区，对42万个社区采用分布式并行计算，同时计算每个社区中挖掘最有影响力 k 个节点所需的时间，缩短计算32万用户中最有影响力用户的时间。为了更加有力地说明本文提出的算法的性能，实验中将CMA算法与IPA、CGA进行比较，以更好地从不同角度对算法的性能作比较，同时也可以通过与这些算法的比较衡量提出算法的可行性和高效性。

如图2所示，描述不同阀值下社会网络中的社区数目。实验结果表明当阀值较小时，整个网络中的社区数目比较少，即每个社区中的节点数相对比较多，当阀值较大时，社区数目比较多，当阀值逼近1时，每个节点都是一个社区，而当阀值在[0.3,0.7]时，社区数目相对比较稳定，维持在6.0×10^2到1.0×10^3之间，波动平稳。社区数目之所以这样分布是因为本文提出的CBIP社区划分算法基于节点间影响概率，而网络中大部分节点间的影响概率介于[0.28,0.76]。社区数目在阀值为0.3时开始趋于稳定并且社区数目大小适中，有利于合理的挖掘有影响力的节点，所以本文中的阀值取0.3。

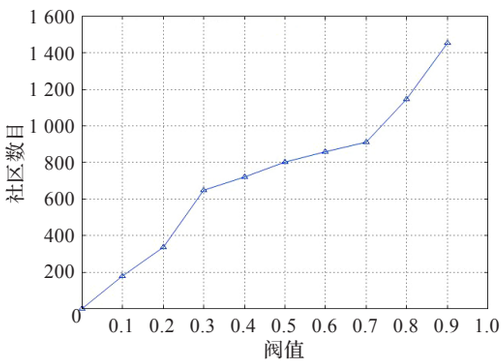

图2　不同阀值下社区数目

如图3所示，描述了三个算法在不同 k 值下影响范围的变化。通过对不同 k 值下影响范围值的分析和量化，得出影响传播范围 $\inf(f)=ak^{0.81}(a\in[1,1.5])$。相比CGA和IPA，在相同的 k 值下本文提出的CMA算法影响范围更广，而且随着 k 值不断变大CMA与其他两个算法的影响范围差距不断变大，这也验证了本文提出的算法能更好适应大规模社会网络，使得影响范围更广。CMA算法在影响范围中的优势主要在于本文将节点间的影响传播路径作为影响评估单元，同时通过阀值控制影响传播路径的长度，有效地增大了节点的影响范围，更充分体现了影响最大化，也说明了本算法在大规模社会网络中的可行性。

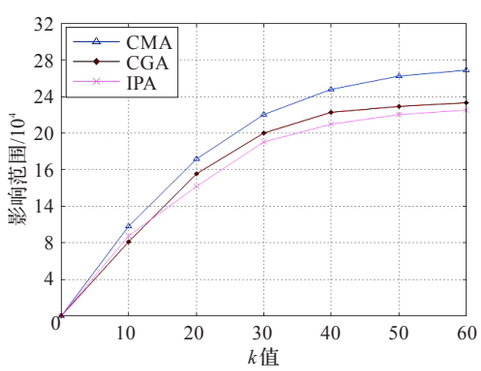

图3　影响范围

如图4描述了三个算法运行时间比较。在初始节点 k 比较小时，CMA算法运行时间比其他两个算法慢，这是因为CMA算法在初期对社区划分需要花费一定的时间。但是当初始节点越来越多时，CMA算法的优势很明显，运行时间比其他两个算法低很多，这是因为CMA根据影响路径计算影响传播范围，已经被作为节点影响路径的不会再被计算，随着 k 的不断增大，社区

万方数据

中需要计算的影响传播路径越来越少,所以运行效率比较高,此特性应用在大规模社会网络中可以减少运行时间,提高挖掘 k 个最有影响力节点的效率。这也说明了本文算法应用在大规模社会网络中的高效性。

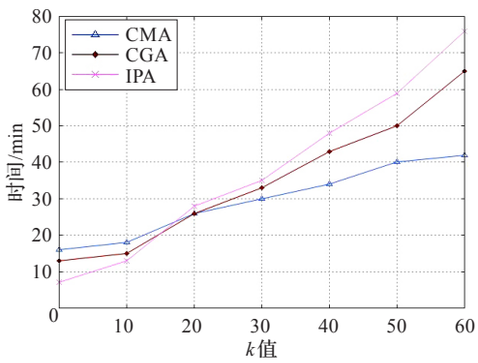

图4　运行时间对比

5　总结与展望

本文提出一种基于新型社区划分的大规模社会网络中影响力最大化算法CMA,与之前关于影响力最大化研究不同的是先基于影响传播概率和标签将社会网络划分成社区,达到化大为小的效果,然后基于CBI传播模型,综合考虑社区内节点和边界节点的影响力,减小因社区划分导致的整体与局部间的差距,以影响传播路径作为影响评估单元,通过并行处理同时挖掘每个社区中有影响力的节点。通过实验验证了提出算法的有效性和高效性,其提高了算法效率,可以很好地适应大规模的社会网络环境。本文接下来的工作会对算法作进一步的改进,在考虑算法效率的同时考虑算法的精度,并与实际的社交网络结合,挖掘出最有影响力的 k 个用户,为网络信息传播的有效控制提供理论依据和实践经验。

参考文献:

[1] Bakshy E,Eckles D,Yan R,et al.Social influence in social advertising:Evidence from field experiments[C]//Proceedings of the 13th ACM Conference on Electronic Commerce,Valencia,Spain,2012:146-161.

[2] Li G,Chen S,Feng J,et al.Efficient location-aware influence maximization[C]//Proceedings of the 2014 ACM Conference on Management of Data,Snowbird,Utah,2014.

[3] Luo Z L,Cai W D,Li Y J,et al.A pagerank-based heuristic algorithm for influence maximization in the social network[M]//Recent Progress in Data Engineering and Internet Technology.Berlin Heidelberg:Springer,2012:485-490.

[4] Chen W,Wang Y,Yang S.Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Paris,France,2009:199-208.

[5] Zhu Y,Wu W,Bi Y,et al.Better approximation algorithms for influence maximization in online social networks[J].Journal of Combinatorial Optimization,2013,43(10):1-12.

[6] Goyal A,Bonchi F,Lakshmanan L V S.A data-based approach to social influence maximization[J].Proceedings of the VLDB Endowment,2011,5(1):73-84.

[7] Chen Y,Chang S,Chou C,et al.Exploring community structures for influence maximization in social networks[C]//Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis held in conjunction with KDD,Beijing,China,2012:1-12.

[8] Yu H,Kim S K,Kim J.Scalable and parallelizable processing of influence maximization for large-scale social networks[C]//Proceedings of the 2013 IEEE International Conference on Data Engineering.[S.l.]:IEEE Computer Society,2013:266-277.

[9] 陈浩,王轶彤.基于阈值的社交网络影响力最大化算法[J].计算机研究与发展,2012,49(10):2181-2188.

[10] Galstyan A,Musoyan V,Cohen P,Maximizing influence propagation in network with community structure[J].Physical Review E,2009,79(5):711-715.

[11] Wang Y,Cong G,Song G,et al.Community-based greedy algorithm for mining top-*K* influential nodes in mobile social network[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.Washington,DC:ACM,2010:1039-1048.

[12] 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,Sierre,Switzerland,2010:1088-1094.

[13] 郭进时,汤红波,吴凯,等.基于社区结构的影响力最大化算法[J].计算机应用,2013,33(9):2436-2439.

[14] 冀进朝,黄岚,王喆,等.一种新的基于社区结构的影响最大化方法[J].吉林大学学报:理学版,2011,49(1):93-97.

万方数据