软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2837−2851 [doi: 10.13328/j.cnki.jos.004723]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向社交网络信息源定位的观察点部署方法*

张聿博[1], 张锡哲[1,2], 张 斌[1,2]

[1](东北大学 信息科学与工程学院,辽宁 沈阳 110819)

[2](医学影像计算教育部重点实验室(东北大学),辽宁 沈阳 110819)

通讯作者: 张锡哲, E-mail: zhangxizhe@ise.neu.edu.cn

摘 要: 准确地定位社交网络上的信息扩散源点,对于网络信息扩散控制具有重要的现实意义.现有的一种可行方法是通过在网络中观察点搜集的过程信息对扩散源进行定位,定位准确率与观察点的选择紧密相关.针对网络中的信息扩散源定位问题,提出了一种网络观察点优化部署方法.考虑单信息源的信息扩散过程,首先分析了特定信息源定位准确率与观察点部署位置之间的关系,以此为基础,发现了与任意信息源定位准确率相关的关键因素.提出基于 *r* 覆盖率的观察点部署策略,以观察点集合的 *r* 覆盖率作为目标函数,实现了 *r* 覆盖率优先观察点选取算法.在模型网络与实际网络上进行了实验,验证了该方法的有效性.提出的观察点部署策略对于网络谣言、计算机病毒的控制具有重要意义.

关键词: 社交网络;信息扩散;信息源定位;观察点部署;*r* 覆盖率

中图法分类号: TP311

中文引用格式: 张聿博,张锡哲,张斌.面向社交网络信息源定位的观察点部署方法.软件学报,2014,25(12):2837−2851. http://www.jos.org.cn/1000-9825/4723.htm

英文引用格式: Zhang YB, Zhang XZ, Zhang B. Observer deployment method for locating the information source in social network. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2837−2851 (in Chinese). http://www.jos.org.cn/1000-9825/4723.htm

Observer Deployment Method for Locating the Information Source in Social Network

ZHANG Yu-Bo[1], ZHANG Xi-Zhe[1,2], ZHANG Bin[1,2]

[1](College of Information Science and Engineering, Northeastern University, Shenyang 110819, China)

[2](Key Laboratory of Medical Image Computing, Ministry of Education (Northeastern University), Shenyang 110819, China)

Corresponding author: ZHANG Xi-Zhe, E-mail: zhangxizhe@ise.neu.edu.cn

Abstract: Locating information source accurately is important for controlling its diffusion on the social network. In previous studies, a feasible way is locating the source using process information collected by the observers. Thus, the accuracy rate is closely related to the observer positions. In this paper, an optimal deployment method for observer positions is proposed. Considering the information diffusion process for single source, it firstly analyzes the relationship between the accuracy rate for locating a specified source and the positions of observers. Based on the relationship, it finds a key factor which is related to the accuracy rate of locating any source. It then suggests a method to deploy the observer positions based on r-coverage rate. It chooses the r-coverage rate of the observers as the objective function to implement the r-coverage rate first observer selection algorithm. The proposed method is tested on model and real networks respectively. Results show that the proposed method is effective. The observer deployment method is significant in controlling internet rumors and computer virus.

Key words: social network; information diffusion; information source location; observer deployment; *r* coverage rate

* 基金项目: 国家自然科学基金(60903009, 71272216, 61073062, 61100090); 中央高校基本科研业务费(120404011, 120804001, 120604003); 黑龙江省普通高等学校青年学术骨干支持计划(1253G017)

收稿时间: 2014-04-17; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

伴随着 Internet 的普及,以及诸如博客(blog)、微博(micro-blog)等新型社交网络服务的大量出现,社交网络(social networks services,简称 SNS)已经成为社会大众获取新闻信息的重要渠道,进而成为当前最重要的信息扩散途径之一[1,2].社交网络的用户往往会将感兴趣的信息转发给其他用户,因此,社交网络上的信息传播很容易形成网络级联效应[3].这种现象在传染病的传播[4,5]、计算机病毒的蔓延[6]、病毒式营销[7−9]等环境中均有出现,已经成为当前的研究热点之一.

现有的工作从传播模型[10,11]、特征分析[12,13]、数据挖掘[14]等方面对社交网络中的信息传播过程进行研究,目的是找到可以使信息传播影响力最大化[15,16]的方法,并且给出了在网络中部署信息源的优化方案[17].相对于对传播范围以及传播路径的预测,另一个重要问题是如何找到网络中的信息传播源头,这对社交网络中的谣言控制和计算机病毒控制等问题具有重要研究价值[18].

对于源点定位问题,一类方法是获取网络中的传播子图,即,某一时刻全部或部分节点是否收到信息(或被感染)的状态子图,然后通过计算谣言中心度[18]、传播临界边缘概率[19]、模拟感染路径[20]等方法找到网络中是传播源的可能性最大的节点.此类方法虽然具有不错的定位准确率,但是需要获取网络中完整的传播子图,即,收到信息的节点及其之间的传播路径.这对于动态多变的大规模社交网络很难实现.

另一类定位方法依靠在网络中部署少量的观察点收集传播信息,然后进行统计计算,推断传播源点.近期工作[21]中提出的定位方法可用于大规模社交网络,在网络中部署少量的观察点,基于观察点收集的局部传播信息及网络中节点间的静态拓扑,估计各观察点信息到达时间的理论延迟,通过最大似然估计计算找到信息扩散源点.这种定位方法的定位精度和计算复杂度取决于观察点的部署策略.

对于观察点部署策略,目前还没有系统的研究工作.一种直观想法是,优先选择在传播过程中起重要作用的节点为观察点,例如高度、高介数、高紧密度的节点.Pinto 等人给出了关于度中心性优先和随机选取两种部署策略的定位实验结果[21].然而,目前还没有针对信息源定位问题的观察点部署策略的相关研究.因此,如何找出最有效的观察点部署方式,是一个亟待解决的问题.

本文首先从网络中单个候选源点入手,分析了观察点与特定源点的位置关系对于定位准确率的影响.进而推断出对于整个网络,观察点的部署位置与任意源点定位准确率之间的关系.然后,针对如何在网络中有效部署指定数量的观察点以达到定位精度最高的问题,给出了一种对观察点部署策略进行优化的方法;以一组观察点的 r 覆盖率的值做为目标函数,将观察点集选取问题转化为网络中节点集的 r 覆盖率最大化问题,从而得到一组定位准确率较高的观察点集合.最后,分别在模型网络和实际网络数据上对本文提出方法的有效性进行了验证,并对算法性能进行了分析.

本文的创新之处在于:得出了在网络中观察点位置与定位准确率之间的关系;并以此为基础,提出了一种基于 r 覆盖率优先的观察点部署策略优化方法;以观察点集合的 r 覆盖率作为优化算法的目标函数,得到观察点的优化部署方案.

本文第 1 节首先对网络信息传播源点定位的相关工作进行描述,然后介绍几种节点对信息传播影响力的度量指标.第 2 节给出本文所采用的传播模型和定位方法.第 3 节分析观察点部署位置与定位准确率之间的关系.第 4 节对基于 r 覆盖率优先的观察点集选取算法进行阐述.第 5 节在实际网络和模型网络数据上验证本文提出方法的有效性,并进行分析.第 6 节总结全文.

1 相关工作

基于在线社交网络的信息传播问题,是当前社会网络研究领域的一个热点问题.对于社会网络信息传播的研究主要集中于两个方面[22],分别是传播影响力的最大化[23]和错误信息的传播控制.其中,对于错误信息传播控制的研究主要集中于链接预测的研究[24],采用链接预测的方法进行错误信息传播控制的主要思想是,通过对传播趋势与影响力最大化两方面的研究,在信息传播过程进行中,将可能成为谣言传播的链路切断,从而达到控制传播范围、减小谣言影响的目的.

相对于链接预测方法,如果能够沿着信息扩散的相反方向及时准确地定位信息扩散源头,那么对于网络中

的信息传播控制将具有重要意义.目前,对于源点定位方法的研究,主要分为两类:一类是基于传播过程中的传播子图的信息进行估计,另一类是基于网络中部分观察点观测的传播数据进行估计.

一种方法是 Shah 等人[18]提出的基于组合数最大似然估计的源点估计量,称为谣言中心度.该方法使用广度优先搜索树 BFS 得到以每个节点为源点时与其他全部节点首次感染相对应的广度优先搜索树,然后使用谣言中心度与 BFS 构建源点最大似然估计量.

另一种方法是基于网络快照数据进行扩散源定位,Zhu 等人[25]在 SIR 模型的基础上,应用样本路径方法对信息源探测问题进行了研究.对于给定的网络历史快照,不区分易感染节点和恢复节点,只是基于网络快照和网络拓扑寻找信息源.在样本路径方法中,预期信息源(即通过计算得到的是实际信息源可能性最大的节点)是最有可能形成网络快照中样本路径的根节点.对于一般网络,在计算过程中需要知道所有可能的样本路径,其数量级为 $O(t^N)$,N 是网络中节点的个数,t 是获得快照的次数.

同样是以 SIR 模型为研究基础,Andrey 等人[19]提出了一种基于动态信息传递方程式的信息源推理方法:首先,对于网络中每一个可能的扩散源点,应用 dynamic message-passing 算法(DMP 算法)计算网络中一个给定节点在快照中所处传播状态(S,I 或者 R)的概率;然后,应用 mean-field-like 近似法计算快照达到传播临界边缘的概率;最后,通过对概率排序得到预期源点.DMP 算法是一种基于动态方程的计算节点状态的方法,该方法具有较好的鲁棒性,在网络快照中节点信息有缺失的情况下依然可以进行定位.

此外,针对多感染源问题,文献[20]中提出了一种 NETSLEUTH 方法.该方法通过已知的网络快照来确定网络中的多个传播源点,采用最小描述长度原则识别传播源集合和病毒传播的路径,得到简化的网络感染子图.然后,识别给出快照中的可能扩散源集合.对于这些可能扩散源,基于最大似然估计的方法对病毒的扩散路径进行优化,并得到可能性最大的节点集,即为扩散源节点集.

虽然上述方法都可以对扩散源进行有效的定位,但都需要以获取扩散过程中的大量动态传播信息为前提(例如网络中全部节点的首次感染信息、多次网络快照信息、与每个网络节点的扩散交互概率等).对于社交网络这样庞大的网络规模来说,在实际应用中,获取网络中的这些动态扩散信息难度很大.

不同于上述方法,Pinto 等人[21]提出了一种在网络中部署少量观察点对信息源进行定位的方法,通过获取观察点记录的传播信息,计算网络中各节点是真实信息源的概率,达到对信息源进行定位的目的.该方法仅需要获取少量观察点的传播信息,占用网络资源非常少,在实际应用中具有较高的可行性,其定位准确率和消耗网络资源的程度,取决于在网络中所部署的观察点的位置和数量.

Brockmann 等人[26]在《Science》上也提出类似的定位方法,该方法创造性地将病毒传播过程中复杂的时空模式简化为均匀的波传播模式,将传统的地理距离用一种概率距离的形式取代,这种概率距离是通过计算节点间的扩散交互概率来实现的.同样是以 SIR 模式为传播模型,通过计算每个可能扩散源与其他全部节点的概率距离,得到这个可能信息源的概率距离图,最终找到实际扩散源.该方法的优势在于:流行病的相关参数与扩散的相关参数是相互分离的,即使流行病参数是未知的,依然可以使用.

当前,对于观察点部署策略的研究较少.一种直观的看法是:在网络中对信息传播的影响力越大的节点,其在信息传播过程中收到信息的可能性越大,记录的传播信息的有效性越高.当前研究中,主要是通过节点中心度对网络中节点的重要程度进行度量[27].其中,节点的度中心度[28](degree centrality)表示该节点的邻居节点个数,在一定程度上可以用于描述该节点的影响力大小.在社交网络研究中,度中心度最大的节点通常被认为是对于扩散过程影响最大的节点[29].

此外,学者们从最短路径的角度提出了紧密中心度[30](closeness centrality)和介数中心度[31](betweenness centrality).其中,紧密中心度表示该节点与网络中其他节点间的距离的反比,值越大,说明该节点与其他节点间的距离越近,可以用于描述该节点将信息发送给其他节点的速度;介数中心度表示网络中两两节点间的最短路径经过该节点的次数,值越大,说明信息传播过程中经过该节点的信息量越大,可以用于描述该节点在信息传播过程中的影响力.

还存在这样一类情况,即对于某一用户来说,他的朋友如果也是朋友,那么他们之间的联系会更为紧密,进

而趋向于形成一个社团.即:对于某一节点来说,其邻居节点间互相联系的可能性,描述了这些节点形成局部社团的趋势,这一趋势用聚集系数[32](clustering coefficient)表示,因此,聚集系数也可以用于描述一个节点影响力的大小.此外,文献[33]中提出了 *K*-核的概念.*K*-核的主要思路是:在网络中度为 1 的节点,其重要性较低;度值大的节点其重要性较高.其计算过程是将网络中度为 1 的节点及其临边全部去掉,去掉的节点其 *K*-核为 1,然后再去掉度为 2 的节点,以此类推.*K*-核描述了节点在网络中的核心程度,通过实验表明,*K*-核越高的节点,其对信息传播的贡献越大.

通过对上述节点中心度的描述可以得出:度中心度较大的节点,由于其邻居节点较多,因此其收到信息的可能性较大;紧密度中心度较大的节点,往往可以更早地收到信息;介数中心度较大的节点,由于经过该节点的信息量较大,因此其收到信息的可能性同样较高.文献[21]中分别采用随机选取策略和高度数优先选取策略进行了实验,结果表明:采用高度数优先选取策略选出的观察点,其定位准确率高于随机选取策略.然而,节点在信息传播过程中的影响力与定位准确性是否直接相关,暂时还没有明确的理论依据.因此,如何在社交网络中的选取一组优化的观察点部署以达到更高的定位准确率,仍然是一个尚待解决的问题.

2 传播模型与定位方法

2.1 传播模型

本文将一个社交网络记为有限无向网络 $G=(V,E)$,其中,$V=\{v_1,v_2,\ldots,v_N\}$是网络中 N 个节点的集合,$E=\{e_1,e_2,\ldots,e_L\}$是网络中 L 条边的集合.对于任意节点 $v\in G$,$N(v)$表示 v 的邻居节点集合,t_v 表示 v 首次收到某一指定信息的时间;对于每一条边 $e_i\in E$,都有对应的 θ_i 表示信息通过边 e_i 传播所需要的时间,即,信息从 e_i 的一端传送到另一端所需要的时间.在网络中选取 K 个节点作为观察点,用 $O=\{o_i \mid o_i\in G\}_{i=1}^{K}$ 表示观察点的集合.在任意时刻,节点 v 有两种可能状态:知情状态,即在当前时刻已经接收到信息;不知情状态,即未接收到信息.

传播过程如图 1 所示,在某一未知时刻 t^*,选取 $s^*\in G$ 为源点,将消息 M 发送给其全部邻居节点 $N(s^*)$.在时刻 t_v,v 收到消息 M,若此时 v 为知情状态,则 v 不做任何操作;若 v 为不知情状态,则 v 变成知情状态,并将消息 M 发送给 $N(v)$.对于节点 $u\in N(v)$,若 e_j 为 u 与 v 之间的边,则 u 在时刻 $t_v+\theta_j$ 接收到由节点 v 发送的消息,此时若 u 是不知情状态,则 u 变为知情状态,并将消息发送给 $N(u)$;否则不做任何操作.依此类推,直到网络中的节点均为知情状态为止.在传播过程中,观察点需要记录信息传播的过程,用 $\varphi=\{(o_i,v,t_{v,o_i})\}$ 表示观察点 o_i 记录的传播信息,其中,$v\in N(o_i)$表示将信息发送给 o_i 的节点,t_{v,o_i} 表示 v 将信息发送给 o_i 的时间.当有多个节点将信息发送给观察点时,只记录首次收到信息的节点和时间.

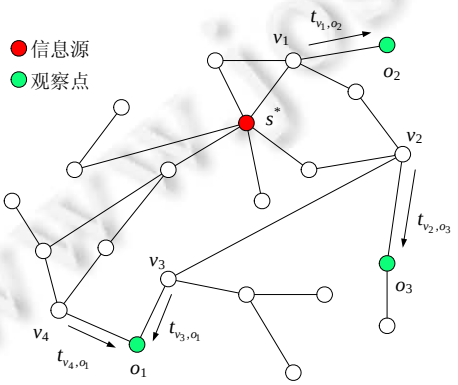


Fig.1 Diagram for information diffusion process

图 1 信息传播过程示意图

2.2 定位方法

本文考虑 Pinto 等人[21]提出的定位方法,称网络中可能的信息源点为候选源点.该方法假设从各候选源点发布信息,依据传播模型计算各观察点收到信息的理论时间,然后与各观察点收到信息的实际时间做对比,找到最可能的候选源点即为实际信息源.然而由于无法获得信息扩散的初始时间,所以无法直接计算各观察点收到信息的理论时间.因此,需要通过对比各观察点收到信息的实际时间延迟与理论时间延迟来找到最符合的候选源点.

用 S 表示候选源点集合,假设观察点不会是实际信息源,那么其余节点均有可能是实际信息扩散源点,则有 $S=\{s_i \mid s_i \in G, s_i \notin O\}_{i=1}^{N-K}$.当信息传播到某一时刻 t,设当前有 K_a 个观察点处于知情状态,用 $\{t_k\}_{k=1}^{K_a}$ 表示知情观察点首次收到信息的实际时间集合,用 $d=\{d_1,d_2,...d_{K_a-1}\}^T$ 表示知情观察点的实际传播延迟向量,其中,d_i 表示观测点 o_{i+1} 与观察点 o_1 首次收到信息的实际时间差,o_1 为第 1 个收到信息的观察点,则有:

$$[d]_k=t_{k+1}-t_1 \tag{1}$$

由中心极限定理可得,网络中边的传播延迟 θ_i 满足 $\theta \sim N(\mu,\sigma^2)$.用 $p(u,v)$ 表示 u 到 v 之间的最短路径,$|p(u,v)|$ 表示这条最短路径的长度,假设某一候选源点 $s_i \in S$ 为实际信息源,则各知情观察点首次收到消息的理论时间 $\{\tilde{t}_k\}_{k=1}^{K_a}$ 为

$$\tilde{t}_k = t^* + \sum_{e_i \in p(s,o_k)} \theta_i = t^* + \mu \cdot |\, p(s_i,o_k)\,| \tag{2}$$

知情观察点间的理论传播延迟向量 $\mu_s=\{\mu_{s_1},\mu_{s_2},...,\mu_{s_{K_a-1}}\}^T$ 为

$$[\mu_s]_k = \tilde{t}_{k+1}-\tilde{t}_1 = \mu \cdot (|\, p(s_i,o_{k+1})\,| - |\, p(s_i,o_1)\,|) \tag{3}$$

应用多元正态分布概率密度计算 d 与 μ_s 的相似度 $\hat{s}$,公式如下:

$$\hat{s} = \frac{\exp\left(-\frac{1}{2}(d-\mu_s)^T \Lambda_s^{-1}(d-\mu_s)\right)}{\sqrt{|\,\Lambda_s\,|}} \tag{4}$$

$$[\Lambda_s]_{k,i} = \sigma^2 \cdot \begin{cases} |\, p(o_1,o_{k+1})\,|, & k=i \\ |\, p(o_1,o_{k+1}) \cap p(o_1,o_{i+1})\,|, & k \neq i \end{cases} \tag{5}$$

对 S 中的节点逐个计算 $\hat{s}$,得到 $\max \hat{s}$ 的候选源点,即为预期源点.

3 观察点部署位置与定位准确率的关系

对于网络 G 和观察点集合 $O=\{o_i\}_{i=1}^K$,定义信息源点定位的准确率为:

定义 1(特定源点的定位准确率). 令信息扩散源点为 s_i,独立进行 n 次信息传播,若基于定位算法得到的预期源点 $\hat{s}=s_i$,则认为定位命中.记 n 次实验中定位命中的次数为 m,则称基于观察点集合 O,s_i 的定位准确率为

$$P_{O_{si}} = m/n.$$

定义 2(任意源点的定位准确率). 随机选取网络中 x 个候选源点 s_i,独立进行 x 次信息传播,记命中次数为 y,则称网络 G 基于观察点集合 O 的定位准确率为 $P_O=y/x$.

由于实际网络中无法预知传播源点,因此,本文主要考虑针对任意源点的定位准确率,且假设网络 G 不随时间变化.因为定位准确率在很大程度上受观察点的个数及部署策略影响,所以本文从研究网络中特定源点的定位准确率与观察点部署位置之间的关系入手,分析观察点部署与定位准确率的关系.

3.1 观察点部署位置对特定源点定位准确率的影响

本文所采用的定位方法[21]是建立在信息在节点间按照最短路径进行传播的假设基础上的,通过计算候选源点到观察点间最短路径长度的差值 $|p(s_i,o_{k+1})|-|p(s_i,o_1)|$,估计信息到达时间的理论值;并以此为参考,与信息到达时间的实际观测值进行对比,通过计算相似度,找到实际信息源.候选源点与观察点之间的最短路径长度之差,是估算信息到达理论时间的基础.

如图 2 所示,信息源 s 到观察点 o_1,o_2,o_3 的最短路径为$|p(s,o_1)|=1,|p(s,o_2)|=4,|p(s,o_3)|=2$,则 o_2 和 o_3 与 o_1 的理论传播延迟为

$$\mu_1=\mu\cdot(4-1)=3\mu \tag{6}$$

$$\mu_2=\mu\cdot(2-1)=\mu \tag{7}$$

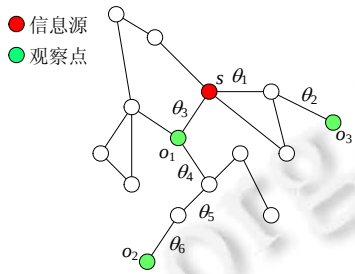


Fig.2 Diagram for the propagation delay

图 2 传播延迟示意图

由图 2 可以得出,o_1,o_2,o_3 的实际收到消息的时间为 $t_{o_1}=t^*+\theta_3,t_{o_2}=t^*+\theta_3+\theta_4+\theta_5+\theta_6,t_{o_3}=t^*+\theta_1+\theta_2$,则 o_2 和 o_3 与 o_1 的实际传播延迟为

$$d_1=\theta_4+\theta_5+\theta_6 \tag{8}$$

$$d_2=\theta_1+\theta_2-\theta_3 \tag{9}$$

其中,μ为网络中边的传播延迟θ_i的均值.以μ_2和d_1为例,$\theta_4,\theta_5,\theta_6$的均值越接近$\mu$,$\mu_1$和$d_1$就越接近,那么理论传播延迟与实际传播延迟的相似度就越高.即,$\theta_4,\theta_5,\theta_6$可以被认为是网络中信息传播延迟的一组抽样.

在上述的定位方法中,我们通过对观察点理论延迟的分析认为:虽然θ_i是随机分布的,但由大数定律可得:当抽样样本较大时,抽样值会趋于接近其算数平均值.因此,对于网络中的一个指定信息源来说,其到观察点间的最短路径的差值越大,该点的理论传播延迟与实际传播延迟的相似度越高,那么这个点在定位过程中被选为实际信息源的概率越高,即,定位准确率越高.

对于一组观察点$O=\{o_i\}_{i=1}^K$,取s为某一指定候选源点,$p(m,n)$表示节点m与n之间的最短路径,假设o_1为距离s最近的观察点,有如下定理:

定理 1. 设$l(s,O)=\sum_{i=2}^{K}(|p(s,o_i)|-|p(s,o_1)|)$,不同的两个观察点集合$O_1$和$O_2$,其相对于$s$的定位准确率分别为$P_{O_1s}$和$P_{O_2s}$,那么当$l(s,O_1)>l(s,O_2)$时,有$P_{O_1s}>P_{O_2s}$.

证明:以网络G中某一$s\in G$为候选源点,消息在未知时刻t^*开始传播,o_1和o_i分别在时刻t_1和t_i收到消息.因为网络中各边传播延迟θ_i满足$\theta\sim N(\mu,\sigma^2)$,则有:

$$t_1=t^*+\sum_{\theta_i\in p(s,o_1)}\theta_i \tag{10}$$

$$t_i=t^*+\sum_{\theta_i\in p(s,o_i)}\theta_i \tag{11}$$

$$t_k-t_1=\sum_{\theta_i\in p(s,o_i)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i \tag{12}$$

设$\bar{\theta}_O$为基于O的$p(s,o_i)$和$p(s,o_1)$上边的传播延迟θ_i的算术均值,则有:

$$\bar{\theta}_O=\left(\sum_{\theta_i\in p(s,o_i)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i\right)\Big/(|p(s,o_i)|-|p(s,o_1)|) \tag{13}$$

由期望与方差的性质可知:

$$\begin{aligned}E(\bar{\theta}_O)&=E\left[\left(\frac{1}{|p(s,o_i)|-|p(s,o_1)|}\left(\sum_{\theta_i\in p(s,o_i)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i\right)\right)\right]\\&=\frac{1}{|p(s,o_i)|-|p(s,o_1)|}\left[\sum_{\theta_i\in p(s,o_i)}E(\theta_i)-\sum_{\theta_i\in p(s,o_1)}E(\theta_i)\right]\\&=\frac{1}{|p(s,o_i)|-|p(s,o_1)|}(|p(s,o_i)|\cdot\mu-|p(s,o_1)|\cdot\mu)\\&=\mu\end{aligned}\tag{14}$$

$$\begin{aligned}D(\bar{\theta}_O)&=D\left[\left(\frac{1}{|p(s,o_i)|-|p(s,o_1)|}\left(\sum_{\theta_i\in p(s,o_i)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i\right)\right)\right]\\&=\frac{1}{(|p(s,o_i)|-|p(s,o_1)|)^2}\left[\sum_{\theta_i\in p(s,o_i)}D(\theta_i)+\sum_{\theta_i\in p(s,o_1)}D(\theta_i)\right]\\&=\frac{1}{(|p(s,o_i)|-|p(s,o_1)|)^2}(|p(s,o_i)|\cdot\sigma^2+|p(s,o_1)|\cdot\sigma^2)\\&=\frac{|p(s,o_i)|+|p(s,o_1)|}{(|p(s,o_i)|-|p(s,o_1)|)^2}\sigma^2\end{aligned}\tag{15}$$

利用切比雪夫不等式可得:

$$P\{|\bar{\theta}_O-\mu|<\varepsilon\}\geqslant 1-\frac{(|p(s,o_i)|+|p(s,o_1)|)\sigma^2}{(|p(s,o_i)|-|p(s,o_1)|)^2\varepsilon^2}\tag{16}$$

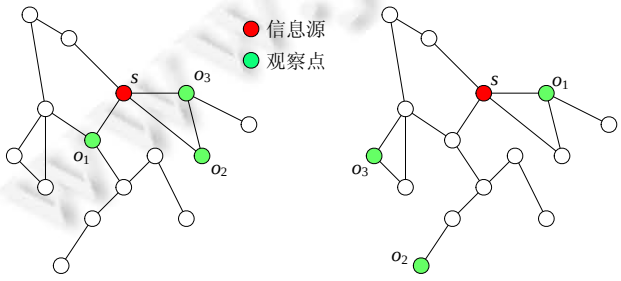
其中,ε为任意正数,当$|p(s,o_i)|-|p(s,o_1)|\to\infty$时,有$\frac{|p(s,o_i)|+|p(s,o_1)|}{(|p(s,o_i)|-|p(s,o_1)|)^2}\to 0$,因此有:

$$\lim P\{|\bar{\theta}_O-\mu|<\varepsilon\}=1\tag{17}$$

说明当$|p(s,o_i)|-|p(s,o_1)|\to\infty$时,算术均值$\bar{\theta}_O$无限接近数学期望$\mu$,有$[d]_k\approx[\mu_s]_k$.

因此,当 $l(s,O_1)>l(s,O_2)$时,有$|\bar{\theta}_{O_1}-\mu|<|\bar{\theta}_{O_2}-\mu|$,即,$\bar{\theta}_{O_1}$比$\bar{\theta}_{O_2}$更接近于$\mu$,因此,基于 O_1 的实际信息传播延迟与理论信息传播延迟间的误差更小.因为本文采用的信息定位方法是通过计算理论信息传播延迟相对于实际信息传播延迟的概率密度分布来实现的,因此,实际信息传播延迟与理论信息传播延迟间的误差越小,定位准确率越高.所以,对于 O_1 和 O_2,有 $P_{O_1s}>P_{O_2s}$.

定理 1 表明:对于某一特定信息源,观察点到该信息源的距离差的和较大时,理论传播延迟可以更准确地反映出信息传播过程中的真实情况,指定信息源在计算过程中的相似度也更高,被选为实际信息源的概率也就更大.也就是说,对于该源点的定位准确率也就更高.如图 3 所示,图 3(a)中信息源与观察点的距离差之和为 0,图 3(b)中信息源与观察点的距离之差为 5,对于 s 来说,图 3(b)中的观察点部署具有更高的定位准确率.



(a) 低定位准确率 (b) 高定位准确率

Fig.3 Diagram for the accuracy rate of locating specified source

图 3 指定信息源定位准确率示意图

3.2 观察点部署位置对任意源点定位准确率的影响

由定理 1 可以得出:指定信息源与观察点之间的距离差与定位准确率有关,如果有一组观察点集合可以满足,对于每一个指定候选源点来说,均具有较高的定位准确率,那么这组观察点部署对于任意信息源的定位准确率较高.以定理 1 中的结论为基础,得到定理 2.

定理 2. 设网络 G 中任意候选源点 s_i 到距离其最近的观察点的距离为 $p_{\min s_i}$,对于一组观察点 O,候选源点集合 S 中 $p_{\min s_i}$ 的最大值 $\max\{p_{\min s_i}\}_{i=1}^{N-K}=r_O$,那么对于两个观察点集合 O_1 和 O_2,其对应的定位准确率为 P_{O_1} 和 P_{O_2},那么当 $r_{O_1}>r_{O_2}$ 时,有 $P_{O_1}<P_{O_2}$.

证明:对于一组观察点 O,取 G 中任意两个候选源点 s_i 和 s_j,o_i 和 o_j 分别表示距离 s_i 和 s_j 最近的观察点,则有 $|p(s_i,o_i)|=p_{\min s_i}\leqslant r_O$,$|p(s_j,o_j)|=p_{\min s_j}\leqslant r_O$.那么 s_i,s_j 和 o_j 构成了一个三角形,根据三角形边的性质,有:

$$|p(s_i,o_j)|\geqslant|p(s_i,s_j)|-p_{\min s_j} \tag{18}$$

其中,当 o_j 在 $p(s_i,s_j)$ 上时,$|p(s_i,o_j)|=|p(s_i,s_j)|-p_{\min s_j}$.因此,当 s_i 为候选源点时,s_i 到 o_i 与 s_i 到 o_j 之间的路径差满足:

$$|p(s_i,o_j)|-|p(s_i,o_i)|\geqslant|p(s_i,s_j)|-p_{\min s_j}-p_{\min s_i} \tag{19}$$

$$l(s_i,O)=\sum_{j=1,j\neq i}^{K}(|p(s_i,o_j)|-|p(s_i,o_1)|)\geqslant\left[\sum_{j=1,j\neq i}^{K}|p(s_i,s_j)|-\sum_{j=1,j\neq i}^{K}p_{\min s_j}-(K-1)p_{\min s_i}\right] \tag{20}$$

取网络中节点的平均路径长度为 R,因为 $p_{\min s_i}\leqslant r_O$,$p_{\min s_j}\leqslant r_O$,所以:

$$l(s_i,O)\geqslant(K-1)(R-2r) \tag{21}$$

那么,对于两个观察点集合 O_1 和 O_2,当 $r_{O_1}>r_{O_2}$ 时,有 $l(s_i,O_1)<l(s_i,O_2)$.由定理 1 可以得出:当 $l(s_i,O_1)<l(s_i,O_2)$ 时,有 $P_{O_1s_i}<P_{O_2s_i}$.可以得出:对于某一指定信息源 s_i,当 $r_{O_1}>r_{O_2}$ 时,对于 O_1 和 O_2,定位准确率 $P_{O_1s_i}<P_{O_2s_i}$;并且对于每一个候选源点 s_i,均有 $P_{O_1s_i}<P_{O_2s_i}$,那么 $P_{O_1}<P_{O_2}$. □

定理 2 表明:对于一组观察点集合来说,若对于每一个候选源点,距离其最近的观察点与该节点之间的距离较小,那么这组观察点的定位准确率较高.如果一组观察点能够满足在任意候选源点的一个较小范围内,均存在至少一个观察点,那么这组观察点部署即为一组优化部署.

3.3 基于*r*覆盖率的观察点优化部署策略

由定理 2 可以得出:对于一组观察点来说,距离观察点距离较小的候选源点越多,则这组观察点的定位准确率越高.对于指定数量的观察点,若以一个指定距离为半径(这个距离要尽可能的小),以观察点集合中的点为圆心做若干个圆去覆盖图中的候选源点,那么能够覆盖候选源点最多的一组观察点为定位准确率最高的一组观察点,即为一组最优的观察点部署.为了得到最优的观察点部署,本文提出通过计算一组观察点集合的 r 覆盖率来衡量这组观察点的定位准确率.观察点集合的 r 覆盖率定义如下:

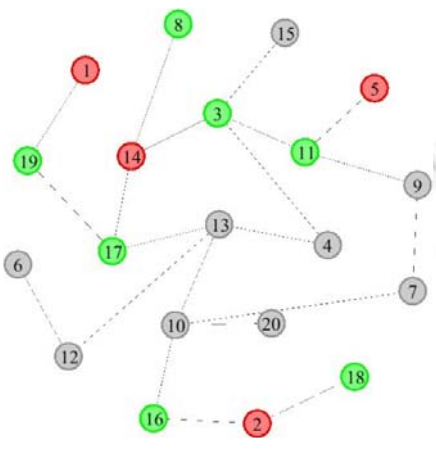


Fig.4 Diagram for 1 cover set

图 4 1 覆盖集合示意图

定义 3(*r* 覆盖率). 在网络 G 中,对于某一观察点 o_i,所有满足 $|p(s,o_i)|\leqslant r$ 的节点的集合 T_{o_i} 称为观察点 o_i 的 r 覆盖集合.集合 $\bigcup_{i=1}^{K}T_{o_i}$ 称为观察点集合 O 的覆盖集合,称 $C_O=\left|\bigcup_{i=1}^{K}T_{o_i}\right|\Big/N$ 为观察点集合 O 的 r 覆盖率.

如图 4 所示,以一组观察点的 1 覆盖率为例,在网络中选取观察点集合为{1,2,5,14},则满足 $|p(s,o_i)|\leqslant 1$ 的候选源点集合{1,2,3,5,8,11,14,16,17,18,19}为该观察点集合的一个 1 覆盖集合,其 1 覆盖率为 0.55.

随着 C_O 的增大,可以有更多的候选源点满足在距离其 r 范围内存在至少一个观察点,那么对于一个观察点集合 O 来说,随

着 C_O 的增大,其定位准确率 P_O 提高.

因此,可以选择 r 覆盖率作为观察点集合的评价标准.对于相同数量的观察点集合,高 r 覆盖率的集合具有更高的定位准确率.那么,观察点的优化部署问题就可以转化为 r 覆盖率的优化问题.

4 r 覆盖率优先的观察点集选取算法

考虑观察点优化部署问题,在网络中寻找 k 个观察点,使得利用这 k 个观察点进行源点定位的准确率最高.根据前述分析,可以将观察点集优化选取问题建模如下:

用 n 维 0-1 向量$\{x_1,x_2,\ldots,x_N\}$表示图 G 中节点是否被选为观察点的状态,其中,x_i=0 表示节点 i 未被选为观察点,x_i=1 表示节点 i 被选为观察点,k 表示在 G 中可以部署的观察点的个数,C_O 表示被选中的观察点集合 O 的覆盖率.那么,达到 r 覆盖率最大的 k 个节点的集合即为一组优化部署,可以有约束条件和目标函数为

$$\begin{cases}\sum_{i=1}^{N} x_i \leqslant k \\ x_i \in \{0,1\},(i=1,2,...,N)\end{cases} \tag{22}$$

$$\max f(x_1,x_2,\ldots,x_N)=\max C_O \tag{23}$$

其中,公式(22)是约束条件,公式(23)是目标函数.显然,上述问题是一个集合覆盖问题,该问题已经被证明是一个 NP 完全问题[34].对于集合覆盖问题,一种有效的解决方法是应用 Greedy 算法[35],其思想是每一步操作都达到最优,即在选取观察点过程中,每增加一个观察点,都使观察点集的覆盖率达到最大,直至得到最终的一组优化观察点集.本文采用 Greedy 算法的思想解决观察点集选取问题,用 k 表示观察点集规模,V 表示网络中节点集合,O 表示观察点集合,$C(\cdot)$表示覆盖集,H 表示候选源点集合,n 表示网络中任意节点.具体算法如下:

算法 1. r 覆盖率优先的观察点集选取算法.

输入:观察点规模 k,网络 G;

输出:一组覆盖率优先的观察点集.

```
BEGIN
1.  O=ϕ, H=V, size(O)=0, size(C(O))=0;        //初始化
2.  if size(O)<k
3.     O′=O;
4.     for (n∈H)                               //得到可以满足 maxC(O+{n})的节点 n
5.        if size(C(O+{n}))>size(C(O′))
6.           O′=O+{n};
7.        end if
8.     end for
9.     O=O′, H=H−{n};
10. end if
11. return O                                   //得到一组优化观察点集
12. END
```

5 仿真实验及数据结果分析

5.1 实验数据

为了验证基于 r 覆盖率优先的观察点部署策略优化方法的有效性,本文在模型网络和实际网络上进行实验,对本文提出的方法进行验证.

其中,ERNetwork1-ERNetwork7 网络是通过随机模型(ER 模型)生成的模型网络;SFNetwork1-SFNetwork4 网络是通过无标度模型(Scale-free 模型)生成的模型网络;Political-blogs[36],UCIonline[37]是实际网络.N 表示网络

中节点的个数;*L* 表示网络中边的条数;AD(average degree)表示网络中的平均度信息,基于此信息,可以看出网络中节点之间边的稠密程度;ND(network diameter)表示网络的直径,代表了网络中所有可以到达的节点对之间路径最长的那条路径长度.具体参数见表 1.

Table 1 Experimental data

表 1 实验数据

Network name	*N*	*L*	*AD*	*ND*
ERNetwork1	1 000	4 924	9.848	5
ERNetwork2	2 994	8 863	5.921	9
ERNetwork3	5 293	7 076	2.674	22
ERNetwork4	994	2 510	5.05	9
ERNetwork5	1 000	3 846	7.692	7
ERNetwork6	1 000	6 289	12.578	5
ERNetwork7	1 000	7 513	15.026	4
SFNetwork1	1 000	5 000	10	5
SFNetwork2	2 946	8 999	6.109	9
SFNetwork3	5 282	7 411	2.806	19
SFNetwork4	966	2 550	5.28	9
UCIonline	1 893	13 835	14.617	8
Political-Blogs	1 222	16 714	27.355	8

5.2 *r*覆盖率与定位准确率的关系验证实验

以模型网络数据为基础进行实验,对 *r* 覆盖率与定位准确率之间的关系进行验证.分别随机选取了 80 组不同 1 覆盖率的观察点集合进行实验.从每组观察点对应的候选源点集合中独立随机选取 1 000 次信息源点进行信息传播,然后对信息源进行定位,得到该组观察点在对应图上的定位准确率.实验结果如图 5 所示,可以看出:随着观察点集合的覆盖率升高,其定位准确率也随之提高.

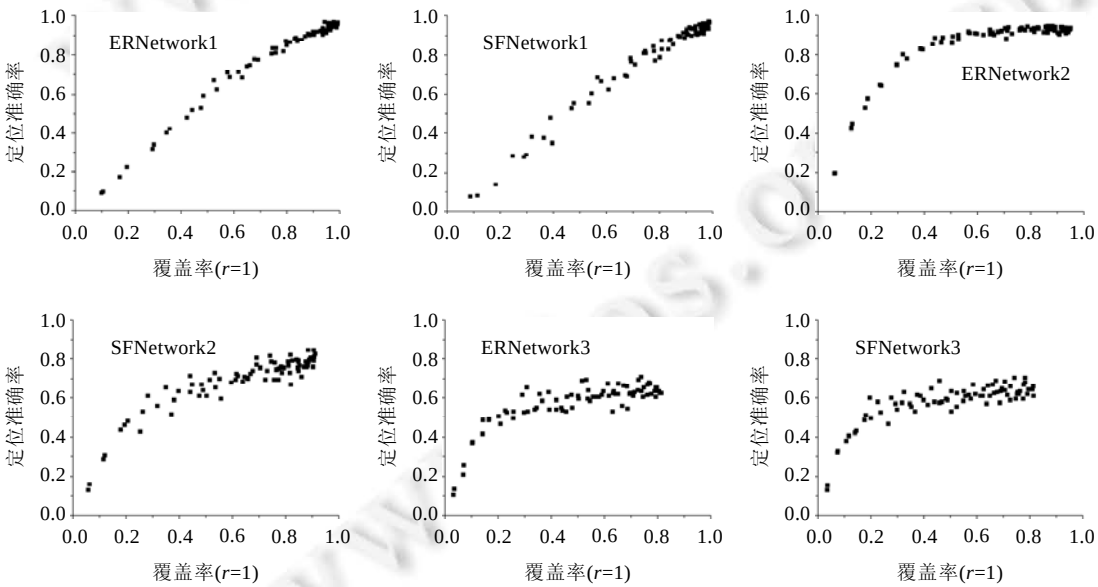


Fig.5 Verification experiment result of the relationship between *r* coverage and location accuracy

图 5 *r* 覆盖率与定位准确率关系验证实验结果

5.3 算法有效性验证实验

为了验证本文提出方法的有效性,分别以高介数优先(high-betweenness)、高度数优先(high-degree)、随机选取(random)以及本文提出的 *r* 覆盖率优先(high-C_O)这 4 种不同观察点部署策略进行信息源定位对比实验.其

中,随机选取和高度数优先是在文献[21]中提出的.节点的度属性表示该节点邻居节点数量,可以描述该节点在静态网络中产生的直接影响力[29].类似地,介数也是一种重要的网络节点特征向量属性,节点的介数属性表示网络中节点对之间的最短路径经过某一指定节点的条数,可以描述节点对网络信息传播的控制能力[31].本文选取这 3 种部署策略,是为了从不同角度与本文所提出的观察点部署策略进行对比.

为了模拟实际应用过程中的信息传播情况,本文在实验过程中每次定位实验均在网络中的非观察点集合内随机抽取一个节点作为信息源,即网络中除观察点外任意节点皆有可能是信息源.具体实验过程是:在网络中以 4 种不同策略分别选取观察点,比例为 5%~10%(每次增加 0.5%),然后从非观察点集合随机选取信息源进行信息传播,然后对信息源进行定位计算.每种部署策略的在每个比例下进行 3 000 次独立实验,得到对应的定位准确率.其中,当 r=1 时,在模型网络上的对比实验结果如图 6 所示.可以看出:在选定的 4 个模型网络上,在同一观察点数量的情况下,通过 1 覆盖率优先策略选取的观察点集合的定位准确率和覆盖率总体上高于现有的基于网络中节点中心性(度、介数)优先的观察点部署策略以及随机选取策略得到的观察点集合.

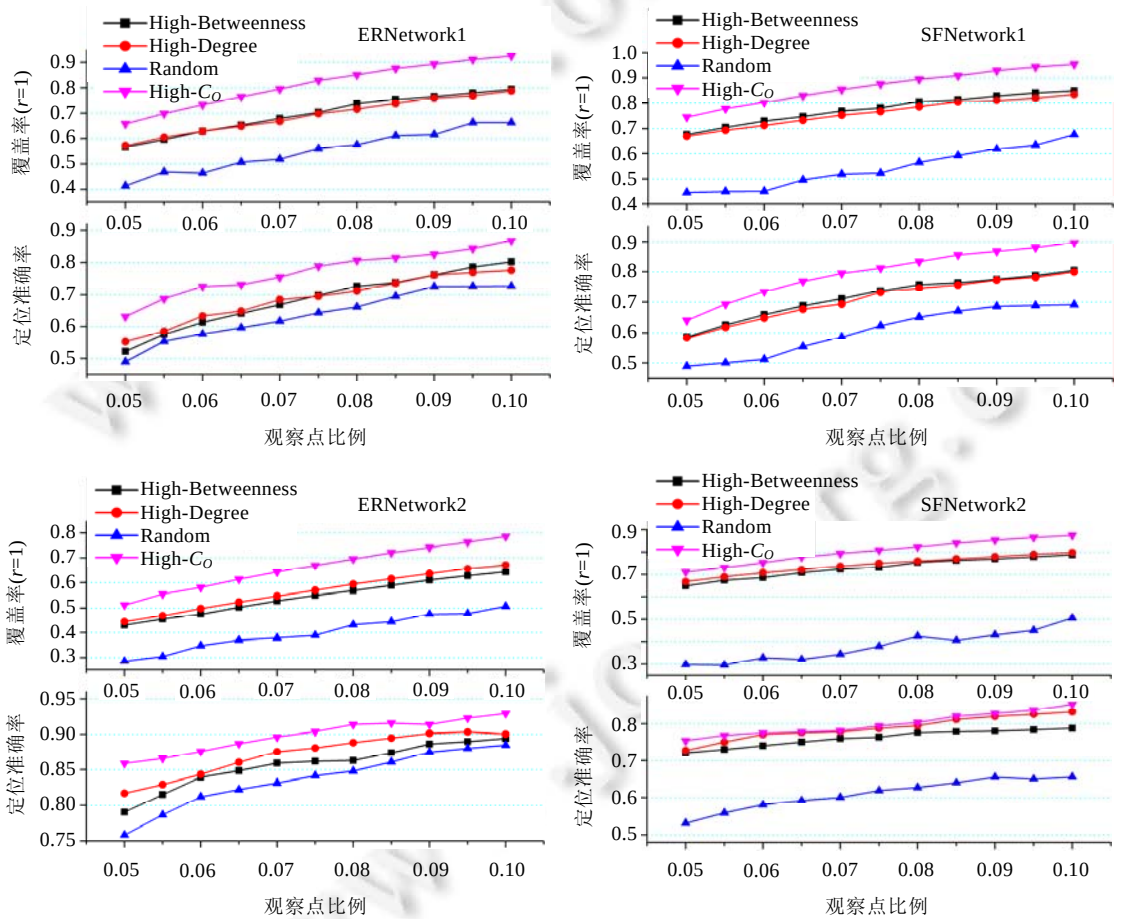


Fig.6 Compared experiment result of the model network location accuracy (r=1)

图 6 模型网络定位准确率对比实验结果(1 覆盖率)

进一步地,为了全面验证本文所提出方法的有效性,在模型网络 ERNetwork4 与 SFNetwork4 上进行 2 覆盖率对比实验,观察点比例选取 1%~5%(每次增加 0.5%),结果如图 7 所示.可以看出:当 r=2 时,因为覆盖半径增大,可以用更少的观察点达到全覆盖效果.基于 2 覆盖率优先选取出的观察点集,其 2 覆盖率与定位准确率全都高于其他 3 种观察点部署策略.

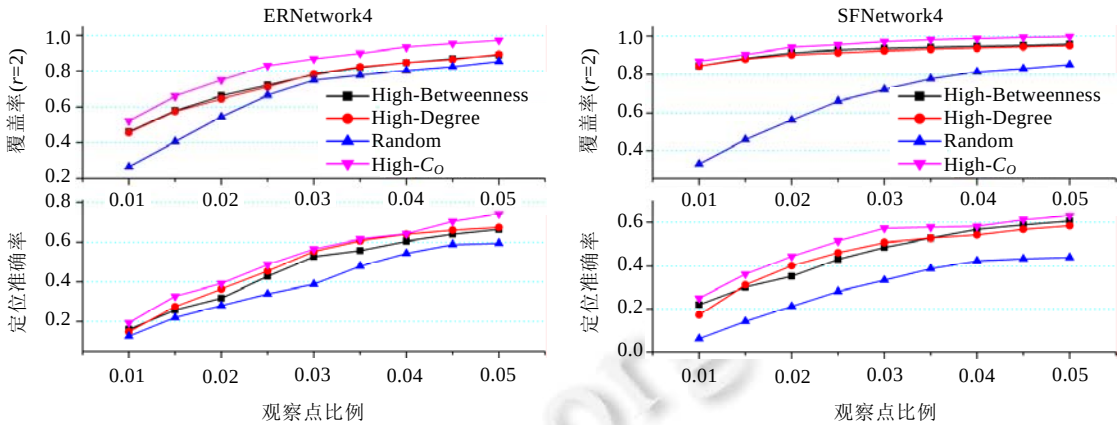


Fig.7 Compared experiment result of the model network location accuracy (r=2)

图 7 模型网络定位准确率对比实验结果(2 覆盖率)

同样地,按照在模型网络上应用的 4 种观察点部署策略,在实际网络上选取观察点集合进行实验.由于在实际网络上观察点比例超过 25%以后,上述 4 种策略所选取的观察点集合的 1 覆盖集合均可达到覆盖网络中的所有节点,因此在实际网络的验证实验中,观察点比例选取 5%~25%(每次增加 5%).实际网络实验结果如图 8 所示,在选定的两个实际网络上,通过 1 覆盖率优先策略选取的观察点集合的定位准确率高于其他 3 种策略.

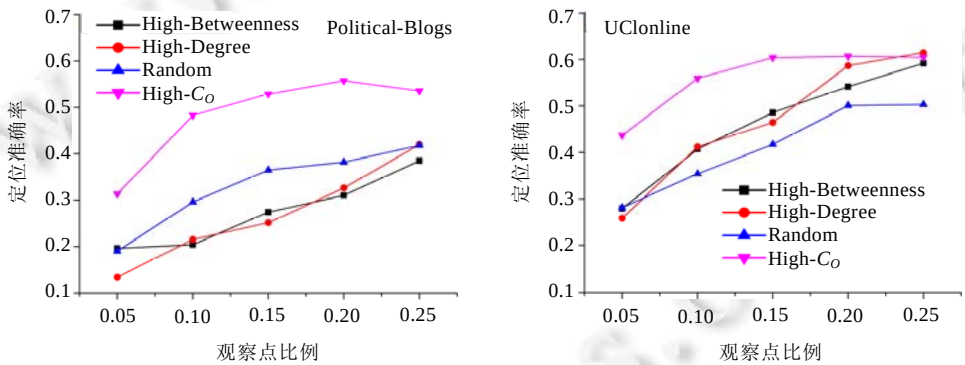


Fig.8 Compared experiment result of the real network location accuracy

图 8 实际网络定位准确率对比实验结果

可以得出结论:在观察点数量相同的情况下,本文提出的优化部署策略具有更高的定位准确率.即对于一个指定的定位准确率(例如,要求定位准确率达到 90%的情况下),本文提出的优化部署策略需要的观察点数量更少,那么在计算过程中的时间消耗更小,充分验证了基于 r 覆盖率优先的观察点部署策略的有效性与可行性.

5.4 算法性能验证实验

为了验证对算法执行效率产生影响的因素,实验数据选取 4 个节点规模相同、网络平均度不同的模型网络数据 ERNetwork1,ERNetwork5,ERNetwork6 和 ERNetwork7.应用 r 覆盖率优先的观察点集选取算法进行实验,取 r=1,观察点比例取 10%,实验结果如图 9 所示.从实验结果可以看出:选取相同数量观察点的情况下,随着网络平均度的增大,计算所需要的时间也增大;并且从曲线斜率的变化可以看出:观察点集中已被选中的观察点越多,再次增加一个观察点所需要的时间就越长.

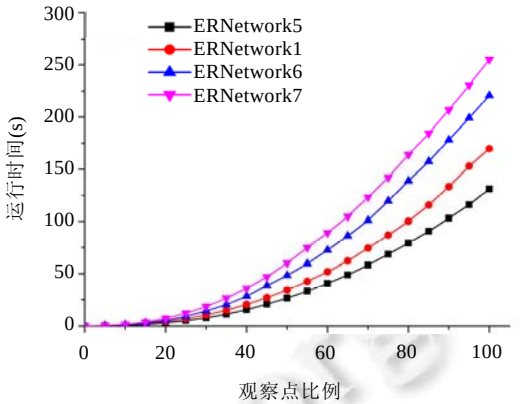


Fig.9 Result of the running time of the algorithm

图 9 算法运行时间实验结果

6 结束语

针对信息源定位领域中的观察点部署策略优化问题,本文首先分析了观察点集合与特定源点的相对位置关系对定位准确率产生的影响,并以此为基础,对网络中的观察点部署位置与任意信息源的定位准确率的关系进行研究.提出了基于 r 覆盖率优先的观察点部署策略,并设计了一种基于 r 覆盖率优先的观察点集选取算法.通过与现有观察点部署策略在模型网络和实际网络中进行对比实验,充分验证了本文所提出方法的有效性,并对算法性能进行了分析.此外,基于 r 覆盖率优先的观察点部署策略主要是针对如何提高定位准确率而提出的,并未考虑缩短定位时间的问题.因此,如何提高定位速度,是下一阶段我们需要研究的内容.

References:

[1] Zinoviev D, Duong V, Zhang HG. A game theoretical approach to modeling information dissemination in social networks. In: Proc. of the IMCCIC. 2010. 407−412.

[2] Zinoviev D, Duong V. A game theoretical approach to broadcast information diffusion in social networks. In: Proc. of the 44th Annual Simulation Symp. Society for Computer Simulation Int'l, 2011. 47−52.

[3] Easley D, Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010,6(1):6.1. [doi: 10.1017/S0266466609990685]

[4] Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks. Physical Review E, 2001,63(6):19. [doi: 10.1103/PhysRevE.63.066117]

[5] Pastor-Satorras R, Vespignani A. Epidemic dynamics in finite size scale-free networks. Physical Review E, 2002,65(3):14. [doi: 10.1103/PhysRevE.65.035108]

[6] Jagatic T, Johnson NA, Jakobsson M, Menczer F. Social phishing. Communications of the ACM, 2007,50(10):94−100. [doi: 10.1145/1290958.1290968]

[7] Chen W, Wang C, Wang YJ. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 1029−1038. [doi: 10.1145/1835804.1835934]

[8] Chen W, Wang YJ, Yang SY. Efficient influence maximization in social networks. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2009. 199−208. [doi: 10.1145/1557019.1557047]

[9] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media. In: Proc. of the 19th Int'l Conf. on World Wide Web. ACM Press, 2010. 591−600. [doi: 10.1145/1772690.1772751]

[10] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978,83(6):1420−1443.

[11] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001,12(3):211−223.

[12] Cha M, Mislove A, Adams B, Gummadi KP. Characterizing social cascades in flickr. In: Proc. of the 1st Workshop on Online Social Networks. ACM Press, 2008. 13−18. [doi: 10.1145/1397735.1397739]

[13] Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B. Measurement and analysis of online social networks. In: Proc. of the 7th ACM SIGCOMM Conf. on Internet Measurement. ACM Press, 2007. 29−42. [doi: 10.1145/1298306.1298311]

[14] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2002. 61−70. [doi: 10.1145/775047.775057]

[15] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2003. 137−146. [doi: 10.1145/956750.956769]

[16] Budak C, Agrawal D, El Abbadi AE. Limiting the spread of misinformation in social networks. In: Proc. of the 20th Int'l Conf. on World Wide Web. ACM Press, 2011. 665−674. [doi: 10.1145/1963405.1963499]

[17] Wang Y, Cong G, Song GJ, Xie KQ. Community-Based greedy algorithm for mining top-*k* influential nodes in mobile social networks. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 1039−1048. [doi: 10.1145/1835804.1835935]

[18] Shah D, Zaman T. Detecting sources of computer viruses in networks: Theory and experiment. ACM SIGMETRICS Performance Evaluation Review, 2010,38(1):203−214. [doi: 10.1145/1811099.1811063]

[19] Lokhov AY, Mézard M, Ohta H, Zdeborova L. Inferring the origin of an epidemy with dynamic message-passing algorithm. Eprint arXiv: 1303.5315. 2013.

[20] Prakash BA, Vreeken J, Faloutsos C. Spotting culprits in epidemics: How many and which ones. ICDM, 2012,12:11−20.

[21] Pinto PC, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. Physical Review Letters, 2012,109(6): 068702. [doi: 10.1103/PhysRevLett.109.068702]

[22] Agrawal D, Budak C, Abbadi AE. Information diffusion in social networks: Observing and influencing societal interests. Proc. of the VLDB Endowment, 2011,4(12):1−5.

[23] Weng JS, Lim EP, Jiang J, He Q. Twitterrank: Finding topic-sensitive influential Twitterers. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2010. 261−270. [doi: 10.1145/1718487.1718520]

[24] Budak C, Agrawal D, El Abbadi AE. Structural trend analysis for online social networks. Proc. of the VLDB Endowment, 2011, 4(10):646−656.

[25] Zhu K, Ying L. Information source detection in the SIR model: A sample path based approach. In: Proc. of the Information Theory and Applications Workshop (ITA). IEEE, 2013. 1−9. [doi: 10.1109/ITA.2013.6502991]

[26] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. Science, 2013,342(6164): 1337−1342. [doi: 10.1126/science.1245200]

[27] Ghosh R, Lerman K. Predicting influential users in online social networksp. In: Proc. of the 4th KDD Workshop on Social Network Analysis. Washington, 2010

[28] Freeman LC. Centrality in social networks conceptual clarification. Social Networks, 1979,1(3):215−239.

[29] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Physical Review Letters, 2001,86:3200−3203.

[30] Newman MEJ. A measure of betweenness centrality based on random walks. Social Networks, 2005,27(1):39−54.

[31] Freeman LC. A Set of Measures of Centrality based on Betweenness. Sociometry, 1977. 35−41.

[32] Holland PW, Leinhardt S. Transitivity in structural models of small groups. Comparative Group Studies, 1971,2(2):107−124.

[33] Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA. Identification of influential spreaders in complex networks. Nature Physics, 2010,6(11):888−893.

[34] Lemke CE, Salkin HM, Spielberg K. Set covering by single-branch enumeration with linear-programming subproblems. Operations Research, 1971,19(4):998−1022.

[35] Kearns MJ. The Computational Complexity of Machine Learning. Cambridge: MIT Press, 1990. 68−73.

[36] Adamic L, Glance N. The political blogosphere and the 2004 US election: divided they blog. In: Proc. of the 3rd Int'l Workshop on Link Discovery. ACM Press, 2005. 36−43. [doi: 10.1145/1134271.1134277]

[37] Opsahl T, Panzarasa P. Clustering in weighted networks. Social Networks, 2009,31(2):155−163. [doi: 10.1016/j.socnet.2009.02.002]

张聿博(1984－),男,辽宁沈阳人,博士生,主要研究领域为社交网络,信息源定位.
E-mail: 273274535@qq.com

张锡哲(1978－),男,博士,副教授,CCF 会员,主要研究领域为复杂网络分析.
E-mail: zhangxizhe@ise.neu.edu.cn

张斌(1964－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为Web信息处理,服务计算,数据挖掘.
E-mail: zhangbin@ise.neu.edu.cn