doi: 10.3969/j.issn.1005-3026.2015.03.011

基于观察点的信息源定位方法的准确率分析

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.东北大学 医学影像计算教育部重点实验室, 辽宁 沈阳 110819)

摘 要: 观察点部署位置与定位准确率之间关系的研究,对基于观察点的信息源定位方法具有重要的意义.从单信息源的信息定位过程入手,首先对理论传播延迟与实际传播延迟进行分析,可知:对于一个指定信息源,观察点间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就越高.进而分析网络中观察点部署位置与指定信息源定位准确率的关系,得出结论:随着信息源到观察点的距离差之和增加,对该信息源的定位准确率增大.在模型网络上进行仿真实验,实验结果验证了分析结论的准确性.

关 键 词: 社交网络;信息扩散;信息源定位;观察点部署;定位准确率
中图分类号: TP 399 **文献标志码**: A **文章编号**: 1005-3026(2015)03-0350-04

Analysis of Accuracy of the Locating Information Source Method Based on Observers

ZHANG Yu-bo[1], *ZHANG Xi-zhe*[1,2], *ZHANG Bin*[1,2]
(1. School of Information Science & 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: The research on the relationship between the position of the observers and the location accuracy has significant importance for the source location methods based on the observers. The information diffusion process for a single source was investigated as the starting point, and theoretical spread delay and real spread delay were then analyzed. It is found that for a fixed source, if the difference of the shortest path among observers is increasing, the similarity of both types of delay is increasing. Then the relationship between the position of the observers and the location accuracy of the fixed diffusion source was analyzed, and it was concluded that: when the sum of distance difference from the information source to the observers increases, the location accuracy for the source becomes larger. The conclusion is validated by simulation experiments on model network data.

Key words: social network; information diffusion; information source location; observers deployment; location accuracy

伴随着博客、微博等新型在线社交网络服务的出现,社交网络已经成为当前最重要的信息扩散途径之一[1].由于用户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容易形成网络级联效应[2].这种现象带来了巨大的商业和社会价值,已经成为当前的研究热点之一.

现有的工作从传播模型[3]、特征分析[4]、数据挖掘[5]等方面对社交网络中的信息传播过程进行了研究,目的是找到可以使信息传播影响力最大化[6]的方法.另一个重要问题是,如何根据现有的传播数据找到信息源,这对网络谣言和病毒控制等问题具有重要研究价值[7-8].

收稿日期: 2014-01-27
基金项目: 国家自然科学基金资助项目(61100090,61073062,71272216);中央高校基本科研业务费专项资金资助项目(N120804001,N120604003,N120404011,HEUCFT1208).
作者简介: 张聿博(1984-),男,辽宁沈阳人,东北大学博士研究生;张 斌(1964-),男,辽宁沈阳人,东北大学教授,博士生导师.

现有的定位方法[7-9]大多需要获取信息传播各个阶段的网络快照(该时刻网络中各节点所处的状态). 这样虽然可以定位信息源,但由于社会网络的规模过于庞大,在实际应用中很难实现.

一种可行做法是由 Pinto 等[10]提出的基于部署观察点的信息源定位方法. 通过在网络中部署少量观察点,记录其传播过程数据,并以此计算网络中节点为信息源的最大似然估计值,得到信息源. 这种方法的定位精度取决于观察点的部署位置,因此关于定位准确率与观察点部署位置关系的研究,对于如何确定观察点集的最优部署位置具有重要意义.

Pinto 等[10]所提出的定位方法,其核心是计算信息传播过程中各观察点间的理论传播延迟与实际传播延迟的相似程度. 因此,观察点间的理论传播延迟是否能够准确地反映出实际的传播情况,是定位准确与否的关键. 本文从实际信息源与观察点间的位置关系入手,分析了观察点部署位置与该源点的距离关系对于定位准确率的影响. 对于某一指定信息源,观察点到该信息源的距离差的和增大时,理论传播延迟可以更准确地反映信息传播过程中的真实情况,那么在信息源定位计算过程中,产生的误差就更小,对于该信息源的定位准确率也就更高.

1　传播模型与定位方法

1.1　传播模型

将一个社交网络记为一个有限无向图 $G=(V,E)$,其中 V 是节点集合,E 是边的集合. 对于任意节点 $v\in V$,$N(v)$ 表示 v 的邻居节点集合,t_v 表示 v 首次收到信息的时间;对于边 $e_i\in E$,有 θ_i 表示信息通过边 e_i 传播所需要的时间. 在 G 中选取 K 个节点作为观察点,用 O 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,在当前时刻已收到信息;不知情状态,尚未收到信息.

传播过程如图 1 所示,在某一未知时刻 t^*,选取 s^* 为源点,将消息发送给 $N(s^*)$ 中全部节点. 在时刻 t_v,v 收到消息,若此时 v 为知情状态,则不做任何操作,否则 v 变成知情状态,并将消息发送给 $N(v)$. 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,然后执行类似操作,依次类推,直到网络中的节点均为知情状态为止. 在传播过程中,观察点还需要记录信息传播的过程,记为 $\varphi=\{(o_i,v,t_{v,o_i})\}$,其中 v 表示首次将信息发送给 o_i 的邻居节点,t_{v,o_i} 表示 v 将信息发送给 o_i 的时间.

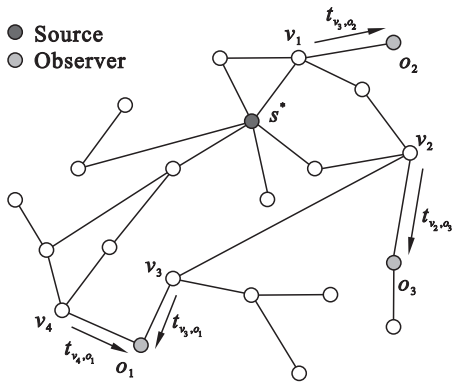

图 1　传播过程

Fig. 1　The propagation process

1.2　定位方法

本文采用 Pinto 等[10]提出的方法定位信息源. 假设各候选源点为实际信息源,计算这种情况下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作比较,最符合的候选源点即为实际信息源. 本文假设观察点不会是实际信息源,那么网络中的全部非观察点均为候选源点. 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 $\boldsymbol{d}=[d_1,d_2,\cdots,d_{K_{(a-1)}}]^{\mathrm{T}}$ 表示知情观察点的实际传播延迟向量,其中 d_i 表示观察点 o_{i+1} 与观察点 o_1 首次收到信息的实际时间差,o_1 为第一个收到信息的观察点,则有 $[\boldsymbol{d}]_k=t_{k+1}-t_1$.

由中心极限定理可得[10],θ_i 满足 $\theta\sim N(\mu,\sigma^2)$. 用 $p(u,v)$ 表示 u 与 v 之间的最短路径,$|p(u,v)|$ 表示其长度,假设某一候选源点 s_i 为实际信息源,则各知情观察点首次收到消息的理论时间为

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

理论传播延迟向量 $\boldsymbol{\mu}_s=\{\mu_{s_1},\mu_{s_2},\cdots,\mu_{s_{K_{(a-1)}}}\}^{\mathrm{T}}$ 为

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

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

$$\hat{s}=\frac{\exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu}_s)^{\mathrm{T}}\Lambda_s^{-1}(\boldsymbol{d}-\boldsymbol{\mu}_s)\right)}{\sqrt{|\Lambda_s|}};$$

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

对候选源点逐个计算 $\hat{s}$,得到 $\max\hat{s}$ 的候选源点,即为预期源点.

2 观察点位置与定位准确率的关系

本文对指定信息源定位准确率做如下定义:

定义 在图 G 中部署一组观察点 O,指定某一候选源点 s_i 固定为信息源,独立进行 n 次信息传播,若通过定位计算得到的预期源点为 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 $P_{O_{s_i}}=m/n$.

本文所采用的定位方法,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.

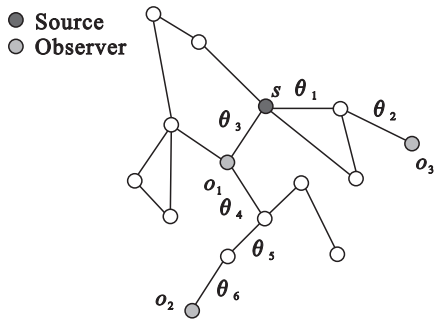

图 2 传播延迟示意图

Fig. 2 Schematic diagram for propagation delay

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 $|p(s,o_1)|=1$,$|p(s,o_2)|=4$,$|p(s,o_3)|=2$,则理论传播延迟为 $\mu_2=\mu(4-1)=3\mu$,$\mu_3=\mu(2-1)=\mu$. 而 o_1,o_2,o_3 的实际收到消息的时间为 $t_{o_1}=t^*+\theta_3$,$t_{o_2}=t^*+\theta_3+\theta_4+\theta_5+\theta_6$,$t_{o_3}=t^*+\theta_1+\theta_2$,则实际传播延迟为 $d_2=\theta_4+\theta_5+\theta_6$,$d_3=\theta_1+\theta_2-\theta_3$. 其中,$\mu$ 为网络中边的传播延迟 θ_i 的均值. 以 μ_2 和 d_2 为例,$\theta_4,\theta_5,\theta_6$ 的均值越接近 μ,μ_2 和 d_2 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.

本文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,但由大数定律可知,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就越高,那么这个节点在定位过程中被选为实际信息源的概率(即定位准确率)也就越高. 由此可得定理 1.

定理 1 设各观察点到信息源 s 的距离差的和为 $l(s,O)=\sum_{i=2}^{K}(|p(s,o_i)|-|p(s,o_1)|)$,对于不同的两个观察点集合 O_1 和 O_2,那么当 $l(s,O_1)>l(s,O_2)$ 时,O_1 的理论传播延迟较小.

证明 选图 G 中某一候选源点 s,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满足 $\theta\sim N(\mu,\sigma^2)$,则有

$$t_k-t_1=\sum_{\theta_i\in p(s,o_k)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i.$$

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

$$\bar{\theta}_O=\Big(\sum_{\theta_i\in p(s,o_k)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i\Big)/(|p(s,o_i)|-|p(s,o_1)|).$$

由期望与方差的性质可知:

$$E(\bar{\theta}_O)=E\left[\frac{\sum_{\theta_i\in p(s,o_k)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i}{|p(s,o_i)|-|p(s,o_1)|}\right]=\mu,$$

$$D(\bar{\theta}_O)=D\left[\frac{\sum_{\theta_i\in p(s,o_k)}\theta_i-\sum_{\theta_i\in p(s,o_1)}\theta_i}{|p(s,o_i)|-|p(s,o_1)|}\right]=\frac{|p(s,o_i)|+|p(s,o_1)|}{(|p(s,o_i)|-|p(s,o_1)|)^2}\sigma^2.$$

利用切比雪夫不等式可得

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

其中,ε 为任意正数,因此有

$$\lim_{g\to\infty}P\{|\bar{\theta}_O-\mu|<\varepsilon\}=1.$$

说明当 $|p(s,o_{k+1})|-|p(s,o_1)|\to\infty$ 时,算术均值 θ_O 无限接近数学期望 μ,有 $[\boldsymbol{d}]_k\approx[\boldsymbol{\mu}_s]_k$.

因此,当 $l(s,O_1)>l(s,O_2)$ 时,有 $|\theta_{O_1}-\mu|<|\theta_{O_2}-\mu|$,即 θ_{O_1} 比 θ_{O_2} 更接近于 μ,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.

根据定理 1,可以得到如下推论.

推论 1 对于指定信息源 s,当两组观察点集满足 $l(s,O_1)>l(s,O_2)$ 时,那么对于 s 的定位准确率满足 $P_{O_{1s}}>P_{O_{2s}}$.

本文采用的信息定位方法,是通过计算理论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,所以对于某一指定信息源,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.

实际传播延迟与理论传播延迟间的误差越小，该源点在计算过程中的相似度越高，被选为实际源点的概率越大. 那么，对于信息源的定位准确率也就更高. 因此对于 O_1 和 O_2，有 $P_{O_{1s}} > P_{O_{2s}}$.

3　实验与分析

本文选取 ER 模型和 BA 模型生成模型网络进行实验，其中 ER 网络的节点度符合正态分布[11]，BA 网络符合幂律分布[12]. 实验数据如表 1 所示. 其中，N 表示网络中节点的个数；L 表示网络中边的条数；AD 表示网络中节点的平均度；ND 表示网络的直径.

表 1　实验数据
Table 1　Experimental data

Network	N	L	AD	ND
ERNetwork1	1 000	3 983	7. 966	6
ERNetwork2	1 000	6 144	12. 288	5
BANetwork1	1 000	3 990	7. 980	6
BANetwork2	1 000	5 979	11. 958	5

具体实验过程是：在网络中选定一个指定节点为信息源，然后选取 100 组不同的观察点集（观察点占节点总数 5% ）. 从指定信息源进行信息扩散，然后进行定位计算作为一次独立实验，每组观察点集进行 2 000 次独立的定位实验.

实验结果如图 3 所示. 其中，横轴表示观察点集的 $l(s,O)$ 值，纵轴表示该组观察点集对于候选源点 s 的定位准确率，曲线部分表示线性回归方程曲线. 可以看出，随着 $l(s,O)$ 值增加，定位准确率也随之提高. 当 $l(s,O)$ 增大时，该组观察点的理论信息传播延迟可以更好地反映实际信息传播延迟，这样在信息源定位过程中产生的误差更小，因此其定位准确率也就更高，验证了上节结论的准确性.

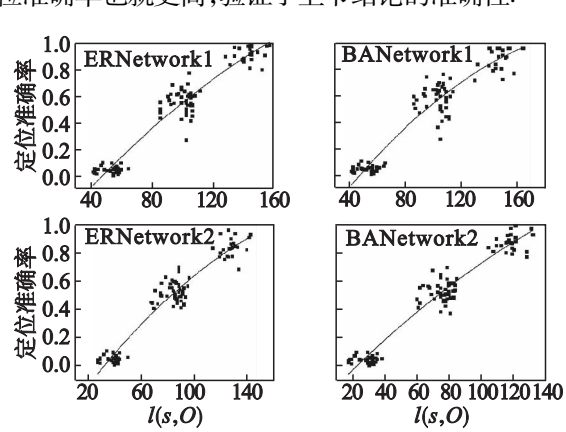

图 3　模型网络实验结果
Fig. 3　Experimental result of the model network

4　结　　语

本文分析了观察点间的理论传播延迟与实际传播延迟之间相似度的计算方法，并以此为基础，得到了观察点与信息源的相对位置对定位准确率产生的影响. 对于一个指定信息源，其到各观察点的距离差之和较大时，往往具有较高的定位准确率. 如果一组观察点对任意信息源均具有较高的定位准确率，那么这组观察点就是一组优化观察点集合. 显然，观察点部署与定位准确率之间关系的研究，对于网络中观察点部署的优化问题，具有重要意义.

参考文献：

[1] Zinoviev D, Duong V. A game theoretical approach to broadcast information diffusion in social networks [C]// Proceedings of the 44th Annual Simulation Symposium. San Diego: Society for Computer Simulation International, 2011: 47 – 52.

[2] Easley D, Kleinberg J. Networks, crowds, and markets. [M]. Cambridge: Cambridge University Press, 2010: 483 – 505.

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

[4] Cha M, Mislove A, Adams B, et al. Characterizing social cascades in flickr[C]// Proceedings of the First Workshop on Online Social Networks. New York: ACM, 2008: 13 – 18.

[5] Wang Y, Cong G, Song G. Community-based greedy algorithm for mining top-*k* influential nodes in mobile social networks[C]// Proceedings of the 16th ACM SIGKDD. Washington D C: ACM, 2010: 1039 – 1048.

[6] Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks[C]// Proceedings of the 20th International Conference on World Wide Web. Hyderabad: ACM, 2011: 665 – 674.

[7] Shah D, Zaman T. Detecting sources of computer viruses in networks: theory and experiment[C]// ACM SIGMETRICS Performance Evaluation Review. New York: ACM, 2010: 203 – 214.

[8] Budak D, El Abbadi A. Information diffusion in social networks: observing and influencing societal interests[J]. *Proceedings of the VLDB Endowment*, 2011, 4(12): 1 – 5.

[9] Prakash B A, Vreeken J, Faloutsos C. Spotting culprits in epidemics: how many and which ones? [C]// The 12th International Conference on Data Mining (ICDM). Brussels: IEEE, 2012: 11 – 20.

[10] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks [J]. *Physical Review Letters*, 2012, 109(6): 068702.

[11] Erd6s P, Rényi A. On random graphs I[J]. *Publicationes Mathematicae*, 1959, 6: 290 – 297.

[12] Barabási A L, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509 – 512.