博客信息传播的网络模型构建

张嘉龄, 李茂青

(厦门大学 信息科学与技术学院, 福建 厦门 361005)

摘　要: 介绍了博客网络中信息传播的博弈演化模型, 利用无标度网络的健壮性对模型进行了简化, 研究了在网络高效率的传播机制下信息的扩张或湮灭。将现实中博客网络的数据与仿真实验的数据进行对比, 发现实验结果和几类真实话题的传播过程基本吻合, 展望了这类信息传播模型的推广方向。

关键词: 复杂网络; 无标度; 小世界; 系统拓扑; 演化模型; 信息传播

中图分类号: TP393.02　　**文献标识码**: A　　**文章编号**: 1672-7800(2008)05-0067-03

1　博客网络拓扑结构

博客(Blog)是一种简易的个人信息发布方式, 任何人都可以注册, 进而完成个人网页的创建、发布和更新。博客充分利用了网络互动、更新即时的特点。发展至今, 博客既是表达个人思想的工具, 被称为“个人出版物”; 也是传播信息的工具, 被称为“新媒体”; 还是网络交流工具, 被称为“自组织网络生态”。现在, 博客正处于快速发展阶段, 是互联网上最受欢迎、发展最快的应用之一。

博客上总会挂上友好博客、知名博客、感兴趣博客的链接, 从而形成有向边, 而博客就是点, 合起来就是一张有向图。这张有向图在很大程度上是一个社会信息传递网络。笔者利用自动搜索程序获得了新浪博客网的一个子图, 该图是有向图, 有向边(u, v)表示博客u上有博客v的链接。这一子图的结点数N=8163个, 平均入度和平均出度为$\bar{k}^{in}=\bar{k}^{out}=\frac{1}{2}\bar{k}$=7.49, 累计入度出度的分布如图1所示。其中左图是累积出度分布, 横坐标是出度, 纵坐标是出度大于等于某一值的博客数的比例。曲线尾部(大约从k^{out}=11开始)符合幂律分布, 拟合的直线斜率为-1.28, 所以出度也将为幂律分布, γ_{out}=2.28。右图是累积入度分布, 其尾部(大约从k^{in}=8开始)符合幂律分布, 斜率为-1.13, 所以入度也将为幂律分布, γ_{in}=2.13。

提取的子图具有明显的无标度特征, 同时其聚类系数C=0.46(远大于具有同样结点总数和平均度的ER随机图的聚类系数$C=\bar{k}/N$=0.0018), 直径D=13.8, 平均路径长度L=4.73(约等于$\log_{\frac{(1-C)\bar{k}^{out}}{1.1}}\frac{(1-C)N}{1.1}$=6.37), 高聚类系数和小平均路径表明这是一个小世界网络。

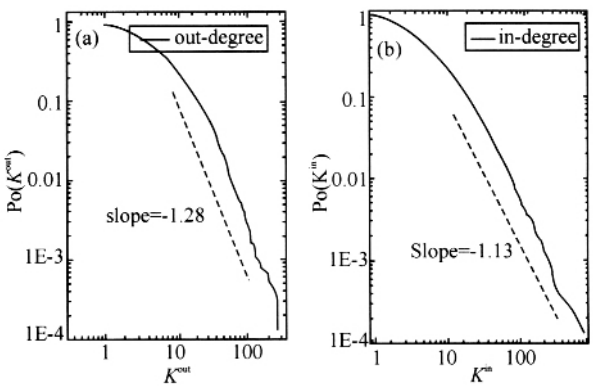


图1　累积出度分布和累积入度分布

2　信息传播的演化模型

考察某一特定话题在博客中传播时, 每一个博客都可能采取的策略有两种: 讨论(Discuss)和沉默(Silent, 不知道该话题或不想讨论)。每一博客都会观察它所链接的博客, 博客u和v之间, 如果同时讨论某一话题则享受度为E, 如果一方讨论一方沉默则享受度均为0, 如果两方都没讨论则享受度为1。这相当于一个协同博弈, E表示了这一话题是否让人满意(Enjoy), 是一个正值, 越大表示越让人感兴趣。对于结点u, 如果讨论则记其行动值为Au=1, 如果沉默则记其行动值为Au=-1。参与讨论的结点的比例$\rho=\frac{1}{N}\sum\frac{A_u+1}{2}$。

结点u的总收益如式(1):

$$P_u=(1-\omega)\alpha L_u+\frac{\omega}{k_u^{out}}\sum_{v\in I_u^{out}}\frac{(1+A_u)(1+A_v)E+(1+A_u)(1+A_v)}{4}\quad(1)$$

其中L是博客点击率, 统计结果表明正比于博客的入度, 在这里设为$L_u=1000k_u^{in}$, α取0.00001, 入度越高也即越知名的博

作者简介: 张嘉龄(1981~), 男, 福建厦门人, 厦门大学信息科学与技术学院硕士研究生, 研究方向为复杂系统、复杂网络; 李茂青(1953~), 男, 福建福安人, 厦门大学信息科学与技术学院教授、博士生导师, 研究方向为管理系统工程、人工智能、决策支持系统。

客,一般其策略被学习的可能性越大,这点在后面的策略转移公式里即可发现。Γ^{out}_u是结点u发出有向边指向的所有结点的集合。ω越大表示个人讨论时越关注别人做法,ω为0时则表示别人策略与自己无关。

改变行动的概率(Probability of Changing Action)如式(2):

$$pca(u)=\frac{\sqrt{\sum_{v\in\Gamma^{out}_u}(A_u-A_v)^2P_v}}{\sqrt{\sum_{v\in\Gamma^{out}_u}(A_u-A_v)^2P_v}+\sqrt{4P_u\sum_{v\in\Gamma^{out}_u}(A_u-A_v)^2P_v}} \quad (2)$$

也就是讨论和沉默哪一方的总收益高,则观察结点就采用大概率跟随那一方的策略。如果用平均收益,则没有意识到人的从众心理,故而采用两方总收益的平方根来计算策略转移概率。

点击量越大,P越高,参考它的结点如果与之同策略,则不易改变策略(分母大了);如果与之异策略,则容易改变策略与之趋同。

并不是所有博客都参与每轮决策,很多博客的主人可能不在线,或者在线但不关注这一话题,也就不会改变对话题的策略。所以每轮只有b(0<b<1,下文计算时取10%)的博客参与行动更新。

在演化过程中,新建联结的概率取决于某人对这一话题的兴趣以及要指向的博客的知名度(与入度点击量强相关),断开联结的概率很小,因为保持联结的代价不高。新添和消亡结点数量相对于原网络是小数,并且网络演化时间其实取得较短,如每个小时决策一次。10%的结点参与状态更新,更新24*90=2160次即三个月,博客总数应基本无变化。无标度网络具有较好的健壮性,即使网络部分结点失效,也不影响信息传播。因为在等概率失效时,被破坏的只有少数连接的非集散结点。进一步讨论,还可以利用无标度网络的脆弱性来遏制信息传播,即去除少数大入度的博客和少数信息源头。在本文算例中,暂不考虑点和边的加减,这一演化模型可以整理成算法1。

表1 算法1

算法 1	博客网络的消息传播 Diffuse(有向图,各结点状态,E,ω,b)
1	for u = 1 to N do
2	根据式(1)计算结点 u 的收益 Pu
3	end for
4	for u = 1 to N do
5	生成随机数 0≤p1<1
6	if p1< b //判断结点是否参与更新
7	生成随机数 0≤p2<1
8	if p2<Pca(u) //判断结点是否改变行动
9	Au= - Au
10	end if
11	end if
12	end for

对于每一组参数E,ω,ρ_0,b,先初始化结点状态,随机选择ρ_0个结点,令其行动值为1(Discuss),其余结点为-1(Silent);然后模拟消息传播,先演化2160次,网络具有小世界特性,使得系统可以较快达到平衡态;ρ稳定,再演化100次,并统计参与讨论结点的比例。为了消除初始状态为1的结点具体分布对结果的影响,重复这一过程10次,最后得到该组参数E,ω,ρ_0,b对应的ρ。演化均衡算法如算法2。

表2 算法2

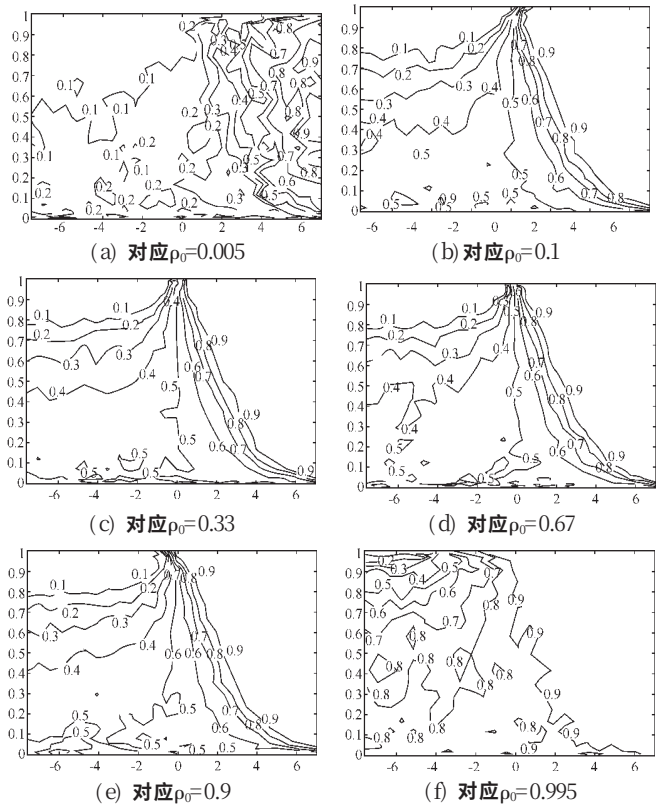
算法 2	消息传播的均衡 Equilibrium()
1	给 E,ω,ρ_0,b 赋值,ρ←0
2	for tinitial = 1 to 10 do
3	随机选择 ρ_0 个结点,令其行动值为 1,其余结点为-1
4	for $t_{diffuse}$ = 1 to 2160 do
5	调用算法 1 更新结点行动值
6	end for
7	for $t_{diffuse}$= 1 to 100 do
8	调用算法 1 更新结点行动值,并计算当前 ρ
9	ρ_t= ρ_t+ ρ
10	end for
11	end for
12	ρ= ρ_t/(100*10)
13	输出均衡时的 ρ

3 主要结果及讨论

图2显示了不同ρ_0下的(E,ω)空间中的ρ。可以看出,稳态时某一话题的讨论比例受到话题兴趣度E、个人对别人策略关注度ω及初始讨论结点比例ρ_0的影响。仿真结果与直觉经验相一致:当其它因素不变时,话题越让人感兴趣,最终讨论的人越多;个人对别人策略如果全然不关注,那么话题也不容易散播开来,但ω高则更凸显话题兴趣度E的重要性,如果话题兴趣度趋向于零,那么高ω只会让大家一块无视这一话题,个别讨论的人会渐渐自觉无趣而中止讨论;最初讨论的人越多,意味着话题跟人们生活越贴近,话题最终也会被相对更多的人讨论。对于话题兴趣、话题贴近生活与否的判断以及在意别人看法的程度,均取决于人的心智模式,实际情况应该是每个人都与众不同,但这里为了简化模型把参数都先简单化成相同值。这种平均意义上的讨论,依然能带来有意义的定性结论。

图2(a)~(f)都是多次初始化网络结点策略所得的平均值。由图2容易知道,话题E值较高,即便初始讨论者极少,也可以引起许多博客的谈论。实验数据表明:纵然E值不高(譬如说E=1),只要初始讨论者足够有声望(有高的点击率L,或者说高的入度kin),就有一定的概率让更多的人乃至全网络的人参与讨论该话题;反之,很热门的话题也有一定可能沉寂下来。

图2(b)~(e)的极端相似,结合不同的初始传播状态,可以看出,初始讨论人群(或者看成是截取的一个中间状态)比例对结果影响并不是很大,除非这一信息在没有被传播开来前就湮灭或者被压制了。图2(f)则可以用来描述一类有人为推动和人为散播的高ρ_0值的信息传播。



(a) 对应ρ_0=0.005　(b) 对应ρ_0=0.1

(c) 对应ρ_0=0.33　(d) 对应ρ_0=0.67

(e) 对应ρ_0=0.9　(f) 对应ρ_0=0.995

图2　(E, ω) 空间中的ρ_e等高线

图2 (a) ~(f) 分别对应ρ_0等于0.005, 0.1、0.33、0.67、0.9和0.995; 对于(E, ω) 空间中的ρ_e, 横坐标是$\log_2$E, 纵坐标是ω 等高线图中右边表示ρ_e很大, 左上方的ρ_e很小。

图3显示了固定ω和ρ_0时, ρ_e和E的关系。可以看出, 该曲线在E=1附近有一个阶跃, 也就是E大于2时该话题才能够引起大多数人共同讨论, 小于0.25时则很容易沉寂。

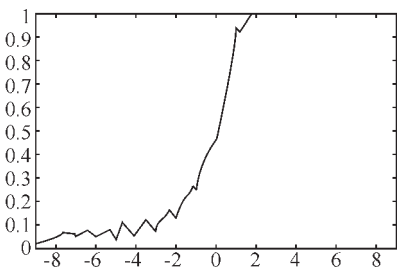


图3　(ω, ρ_0)=(0.8, 0.33) 时的E-ρ_e关系(横坐标是$\log_2$E, 纵坐标是ρ_e)

4　结束语

初始参与讨论的结点比例ρ_0在一定程度上显示了这一话题是否贴近生活, 越贴近生活的话题ρ_0越大。但对于具体初始哪些结点参与, 现阶段并没有很好的设定方法, 这一模型的源头还有待进一步探讨。

话题的粒度需要很好的把握, 这需要更多的实证研究来支持。如以 '娱乐新闻 '为标题来划分一类话题, 则粒度太大; 如果以某电视剧中某一集中的事件来划分话题, 一般来说粒度又太小。根据不同粒度的话题, 甚至可以提出完全不同的模型。

话题还有一个生命周期, 或长或短, 如饮食话题就有长周期, 某项科学技术周期长度次之, 某个明星的话题一般来说周期短。处于不同生命周期的话题, 应该有不同的初始网络状态和发展趋势。

本文是复杂网络应用于现实生活中的一个实例, 通过数值仿真得到了信息传播的一个大致特点。相信随着这方面的研究越来越深入, 复杂网络中的消息传播机制会越来越得到重视。

参考文献:

[1] Erd?s P, Rényi A. On random graphs[J]. Publications Mathematical Debrecen, 1959, 6: 290~297.

[2] Watts D J, Strogatz S H. Collective dynamics of "small-world" networks[J]. Nature, 1998, 393: 440~442.

[3] Barabási A-L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286: 509~512.

[4] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45: 167~256.

[5] Boccaletti S, Latora V, et al. Complex networks: Structure and dynamics[J]. Physics Reports, 2006, 424: 175~308.

[6] Nowak M A, May R M. Evolutionary games and spatial chaos[J]. Nature, 1992, 359: 826~829.

[7] Abramson G, Kuperman M. Social games in a social network[J]. Physical Review E, 2001, 63: 030901.

[8] Flake G W, Lawrence S, et al. Self-Organization and Identification of Web Communities[J]. IEEE Computer, 2002, 35(3): 66~71.

[9] Gruhl G, Guha R, et al. Information diffusion through blogspace [C]. Proceedings of the 13th International Conference on World Wide Web (New York), 2004, 13: 491~501.

[10] 吴金闪, 狄增如.从统计物理学看复杂网络研究[J].物理学进展, 2004(1).

[11] 周涛, 傅忠谦.复杂网络上传播动力学研究综述[J].自然科学进展, 2005(5).

(责任编辑: 赵　峰)