DOI:10.3969/j.issn.1673-4785.201305077
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130613.1325.001.html

基于内容的热点话题传播模型

韩忠明,张慧,张梦
(北京工商大学 计算机与信息工程学院,北京 100048)

摘　要:采用传染病模型对网络热点话题的传播进行建模具有重要的价值,但是现有的传染病模型并没有区分话题类型和不同用户传播话题的概率,为此提出一个基于内容的网络热点话题传播模型.模型中引入了用户对话题传播的敏感度,基于用户话题敏感度定义了单个用户传播话题的概率,融合话题的内容分类特性、用户传播概率、用户重入概率等因素,借鉴 SIRS 模型的基本思想,构建了话题传播模型(CSIRS).在无标度网络、小世界网络、随机网络和真实社会网络上作了不同实验,实验结果表明 CSIRS 模型不仅能够呈现一般传染病动力模型的传播模式,还能够呈现多个波动、小范围长时间传播、快速上升缓慢下降等社会网络热点话题的传播模式.该模型为融合网络结构和话题内容属性建模话题传播过程带来新的研究思路.

关键词:热点话题;传播模型;传染病模型;话题传播模型
中图分类号: TP391　　**文献标志码**:A　　**文章编号**:1673-4785(2013)03-0233-07

中文引用格式:韩忠明,张慧,张梦. 基于内容的热点话题传播模型[J].智能系统学报, 2013, 8(3): 233-239.
英文引用格式:HAN Zhongming, ZHANG Hui, ZHANG Meng. A hot topic propagation model based on topic contents[J]. CAAI Transactions on Intelligent Systems, 2013, 8(3): 233-239.

A hot topic propagation model based on topic contents

HAN Zhongming, ZHANG Hui, ZHANG Meng
(College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: Current researches suggest a new hot topic propagation modeling pattern, which uses the epidemic model, and adds great value to the industry. Unfortunately, the existing epidemic models make no distinction between topic type and topic propagation probability of different users. A new propagation model based on topic contents for network hot topics was proposed in this paper. User's sensitivity degree to topic propagation was introduced and based on this, the single user propagation probability was defined. By integrating factors such as topic contents classification characteristics, user propagation probability, user re-entry probability, and drawing on the basic idea of SIRS model, a topic propagation model (CSIRS for short) was built. Different experiments were conducted respectively in a scale-free network, small-world network, random network and real social network. The experimental results show CSIRS model can not only present the propagation pattern of a general dynamic model of infectious disease, but also present the propagation patterns of hot topics on social networks, such as multiple wave propagation patterns, small scale spreading with long tail propagation pattern and rapidly rising slowly falling propagation pattern. CSIRS model provides a new idea for modeling the topic propagation process by integrating network topology and topic attribute.

Keywords: hot topic; propagation model; epidemic model; topic propagation model

网络热点话题对虚拟社会和现实社会都具有极大的影响力,理解并建模热点话题在社会网络上的传播具有重要价值.网络热点话题传播与人类社会中的疾病传播具有高度的一致性,采用传染病传播模型来理解与分析网络中的消息传播具有广泛的基础.如采用传染病模型理解传感网络中的消息传播[1],在社会网络中结合复杂网络和传染病动力学

收稿日期:2013-05-28.　**网络出版日期**:2013-06-13.
基金项目:国家自然科学基金资助项目(61170112);教育部人文社会科学研究基金资助项目(13YJC860006);北京市属高等学校科学技术与研究生教育创新工程建设项目(PXM2012_014213_000037).
通信作者:韩忠明. E-mail: hanzm@th.btbu.edu.cn.

理论构建动力学演化方程组[2]等.

在传染病动力学中,仓室模型是最常用的模式,可根据不同的种群设置形成不同的模型,例如 SIR、SIS 和 SIRS 等,这些模型假设种群中个体被感染者感染的概率相等.随着小世界网络、无标度网络为代表的复杂网络研究的兴起,将传染病疾病传播与社会结构相结合,使传染病传播模型深入到个人层面[3-5].文献[6]用模拟方法证实了小世界网络能加快疾病传播进程.文献[5,7-9]是无标度网络上传染病模型的代表,用平均场方法研究了节点无限的无标度网上的 SIS 和 SIR 模型.文献[9]说明无标度网络抵抗传染病的能力很弱,可以在任意小的有效传播率下维持传播.文献[5]发现无标度网络条件下 SIR 模型的传播率阈值与节点度的指数截止值相关,随着指数截止值的增加而增大.SpikeM 模型[10]解决了 SI 模型的缺点,即模型服从幂律下降规律的同时保证待感染的节点数有限.

除采用传染病传播动力模型外,自激霍克斯过程(self-excited Hawkes process)[11]等随机过程理论建模话题也是一个研究思路.Crane 用带参数的自激霍克斯泊松过程对 YouTube 视频评论进行建模[12].文献[13]假设一个暴力事件服从一个自激霍克斯过程,并实现了复杂事件的自激霍克斯过程参数估计以及事件预测.信息级联模型(information cascading model)[14-18]采用了网络上的级联事件机制建模复杂网络上的消息传播,它将网络中的事件看成由一系列的级联活动组成,一个参与者独立观察他人的行为并做出自己的决策,这些决策依赖于用户的偏好.

无论是传统的仓室模式,还是基于复杂网络的传染病模型,都假设个体感染不同疾病的感染概率相同,不同类型的疾病对易感人群的效应也是相同的.然而,这 2 个假设在社会网络中都不太合理,网络用户对不同类型的热点话题的敏感程度不同,如对于一些负面的、批判性的热点话题,用户的参与性较强;对于明星类热点话题参与性较低.同理,具有不同特征的用户对话题的敏感程度也不同,即个体感染疾病的概率不同.

针对上述不足,本文提出基于内容的网络热点话题传播模型,该模型改进了传统传染病动力模型的缺点,考虑了不同节点对不同话题的敏感度.在小世界、无标度、随机网络与真实社会网络上进行了不同的试验,实验结果表明了考虑用户和话题之间的关系能够呈现出多个波动、小范围长时间传播、快速上升缓慢下降等传播模式,验证了模型的有效性.

1 基于话题内容的传播模型

为了克服传染病感染模型的缺点,假设影响话题在社会网络上传播的因素有:

1)话题吸引力指数:用来表达话题的内容特性,用 θ 表示.不同类型话题吸引用户参与的强度不同,则 θ 值不同.

2)网络结构:由用户构成的社会网络.

3)话题保持时间长度:一个节点能感知到话题的时间长度,如同疾病的携带期,用 π 表示.感染用户在话题保持时间长度内,其状态一直为感染状态.

4)节点对话题的敏感度:每个用户对话题的敏感程度都不同,设用户 i 对话题 T 的敏感度为 β_i.

5)用户重入概率,用 δ 表示,用户可以多次参与话题,也就是可以多次感染.为了简单起见,设用户重新进入易感群的概率为固定值.

采用 SIRS 模型作为基础模型,提出一个基于内容的话题传播模型(content based SIRS model, CSIRS),其基本模型中的种群迁移如图 1 所示.

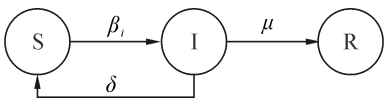


图 1 CSIRS 模型种群迁移示意

Fig.1 Population migration in CSIRS

从图 1 可以看出,热点话题中涉及到的用户分为 3 类种群:潜在参与用户(易感者 S)、参与用户(已感染者 I)和免疫用户(恢复者 R).一个用户在易感状态(S)下,以不同的概率进行传播(感染).一个用户在感染状态(I)下,以特定的概率转化为免疫用户(R)或者重新进入易感状态(S).

在 CSIRS 模型中,假设每个用户传播话题的概率不同,下面给出 CSIRS 的传播概率.设在一个社会网络中的用户总数为 N,一个用户 i 的相邻用户个数为 N_i,该用户对话题的敏感度为 β_i,则该用户在传染时刻成为感染者的概率,也就是用户传播话题的概率为

$$p_i = 1 - (1 - \beta_i)^{\sum_{j=1}^{N_i}\delta_j}.$$

式中:δ_i 为用户 i 的相邻用户中为感染者的符号函数,也就是

$$\delta_i = \begin{cases} 0, & \text{用户 } j \text{ 不是感染者;} \\ 1, & \text{用户 } j \text{ 为感染者.} \end{cases}$$

根据 CSIRS 模型的假设条件和传播概率,设计传播过程算法,步骤如表 1 所示.传播过程算法分为 2 个部分:1)初始化参数,随机产生初始话题的发起者;2)进入迭代循环,达到迭代次数上限或者所有

节点状态为 R 时算法结束,在循环体中,对网络中的每个节点进行状态判断,并进行相应的操作.

表 1　传播过程算法

Table 1　Propagation process algorithm

序　号	伪代码
1)	初始化
2)	初始化参数
3)	随机产生初始的话题发起者
4)	While: 循环
5)	For i in Networks: 对网络的每个节点
6)	If i = = "I": 如果节点状态为 I
7)	If i 在保持时间范围内:
8)	保持
9)	Else: 以概率 δ 进行重入
10)	Elif i = = "S": 如果节点状态为 S
11)	获取相邻节点
12)	计算参与概率 p_i
13)	以概率 p_i 感染

2　实验结果与分析

为了分析 CSIRS 模型的传播模式,采用 Python 语言实现了一个仿真系统,Python 中包含一个 SimPy(simulation in Python)包,基于 SimPy 验证 CSIRS 模型的效果.实验分为 2 个部分:1)仿真实验,用来分析和考察 CSIRS 模型在不同类型复杂网络下的传播模式;2)真实社会网络实验,用来分析和考察 CSIRS 模型在一个真实社会网络下的传播模式.

所有的实验都运行在一个平台下,平台实验环境为 Intel I5 M540 2.53 GHz CPU、4GB 内存、300GB 硬盘,操作系统为 Windows7,Python 版本为 2.7.

2.1　实验设置

在仿真环境中,用户对不同话题的敏感程度难以界定,为了真实地仿真用户受话题吸引的程度,采用模拟标签的方式模拟用户对不同话题的敏感度.

对话题随机设定 3~8 个标签,类似地,对用户也随机设定 3~8 个标签,然后计算用户标签和话题标签的相似度作为用户对话题的敏感度.设话题标签为 $T_{\text{tags}}=\{\text{tag}(1),\text{tag}(2),\cdots,\text{tag}(n)\}$,用户标签为 $U_{\text{tags}}=\{\text{tag}(1),\text{tag}(2),\cdots,\text{tag}(n)\}$,则用户 i 对话题的敏感度定义为

$$\beta_i=\frac{2\times|T_{\text{tags}}\cap U_{\text{tags}}|}{|T_{\text{tags}}|+|U_{\text{tags}}|}. \tag{1}$$

直观分析式(1),用户标签和话题标签的相似度越高,用户对话题的敏感度越高,也就越容易传播话题,这与现实一致.

2.2　仿真实验结果与分析

实验采用 3 种网络形态对 CSIRS 模型的传播效果进行仿真分析,3 种网络形态分别是:

1)无标度网络(scale-free network, FS).无标度网络具有很强的异质性,少量的节点占据了大量的边,因此这些少数节点对于网络的性质会有很大的影响.

2)小世界网络(small-world network, WS).小世界网络具有小世界特性(较小的平均最短路径)及聚类特性(较大的聚类系数).采用 WS 模型生成小世界网络,其中重连概率 p 设为 0.2.

3)随机网络(Erdos-Renyi network, ER).随机网络的结构具有很大的变化,通过 Erdos-Renyi 模型生成随机网络,其中连接概率设为 0.02.

为了比较 CSIRS 模型在 3 种网络上的不同传播效应,对 3 种网络都采用相同的参数设置,如表 2 所示.其中,话题保持时间长度 π 是一个随机变量,假设每个用户对话题的保持时间长度是一个正态分布,为了简单起见,设分布的均值为 7.

表 2　参数设置

Table 2　Parameter settings

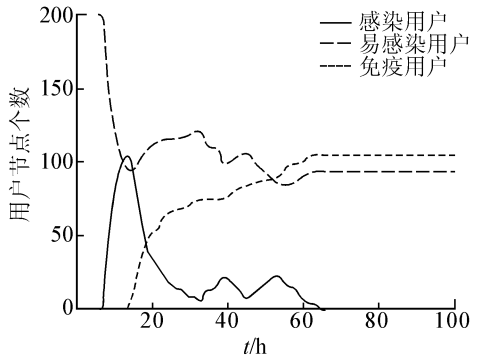
参　数	值
节点个数 N	200
话题吸引力指数 θ	0.5
话题保持时间长度 π	$N(7,1)$
用户重入概率 δ	0.5
用户对话题的敏感度 β_i	$\beta_i=\frac{2\times\lvert T_{\text{tags}}\cap U_{\text{tags}}\rvert}{\lvert T_{\text{tags}}\rvert+\lvert U_{\text{tags}}\rvert}$

3 种不同的网络在相同的参数配置下各执行 3 次仿真实验,并将 3 次结果进行综合比较.图 2~4 分别是无标度网络、小世界网络和随机网络上的实验结果.综合分析图 2~4 的传播模式,可以得出如下结论.

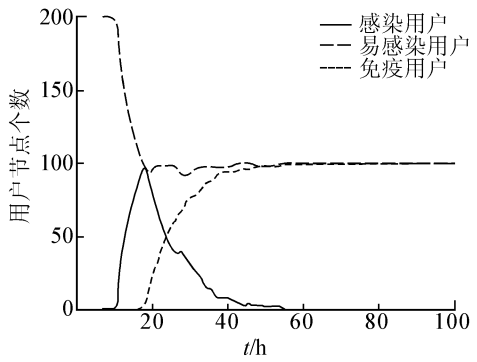
1)CSIRS 模型在不同的网络上,话题传播呈现不同的模式.CSIRS 模型在小世界网络上感染的用户最多,速度也快,但消息的消退速度也最快;而在无标度网络上传播的速度最慢,消退的速度也最慢.

2)从图 2 可以看出,CSIRS 模型在无标度网络上的传播能够呈现出多个波动.由于无标度网络中度很高的节点很少,多数节点的度都很小,因此如果消息无法传播到度高的节点上,那么话题的传播范围会很小,如图 2(c)所示.如果话题传播过程中传

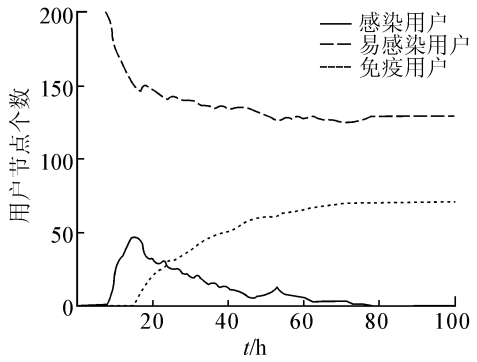
播到一些度高的节点,那么能引起多个波动,如图 2(a)所示.另外由于无标度网络的结构复杂性,因此话题在无标度网络上的存活期长,而且可能在较长时间的静默期后重新传播,如图 2(c)所示.图 2(b)则呈现出一般的话题传播模式,但也有一些小的波动.另一方面,从图 2 的 3 个子图可以看出,这 3 个传播模式具有较大的差异,其原因在于节点对话题的敏感度和节点的度没有必然的相关性,因此话题传播范围和速度的决定因素不仅是网络的物理性质,还有节点对话题的敏感程度等,这个现象与文献[19]中的分析一致.



(a)第 1 次实验



(b)第 2 次实验

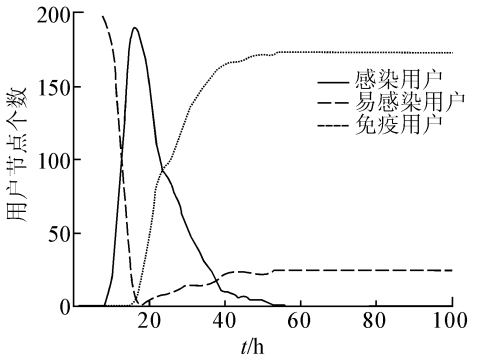


(c)第 3 次实验

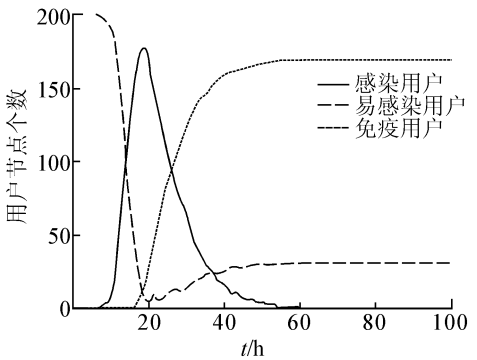
图 2 无标度网络上的仿真结果

Fig.2 Simulation results under scale-free network

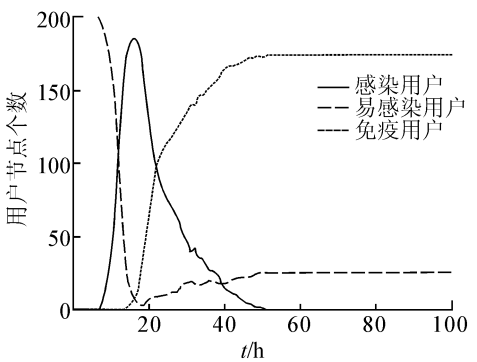
3)图 3 中 3 个子图呈现出基本一致的传播模式,话题快速上升后以近似幂律的形式下降.话题的存活期较短,3 次传播的存活期都小于60 h,而在无标度网络和随机网络中存活期都超过了60 h.这个结果与 Watts 等发现疾病在小世界网络中传播速度比规则网络更快、传播范围更广[20]的结果一致.



(a)第 1 次实验



(b)第 2 次实验



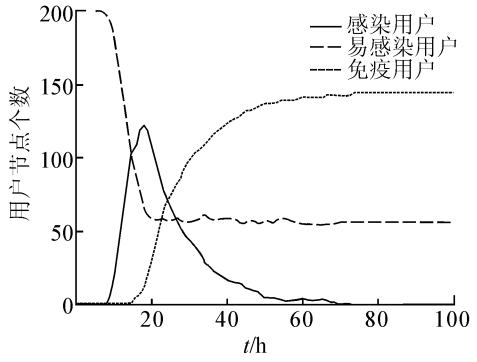
(c) 第 3 次实验

图 3 小世界网络上的仿真结果

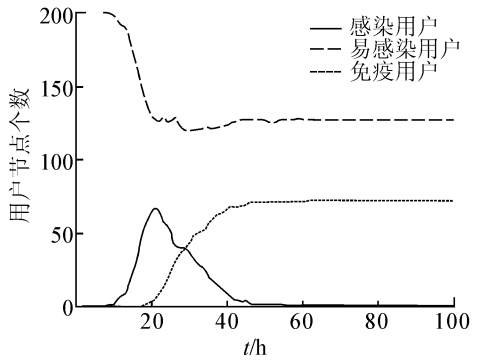
Fig.3 Simulation results under small-world network

4)从图 4 可以看到,CSIRS 模型在随机网络上传播的范围和传播速度都介于无标度网络和小世界网络之间.随机网络上的话题传播长尾特性比在小世界网络上的传播明显,这说明由于内容和用户之间存在不同的敏感程度,话题在传播高峰期后得以在较长的时间内继续小范围的传播.此外,随着随机网络的连接概率增大,随机网络的边数将不断增加,节点间的

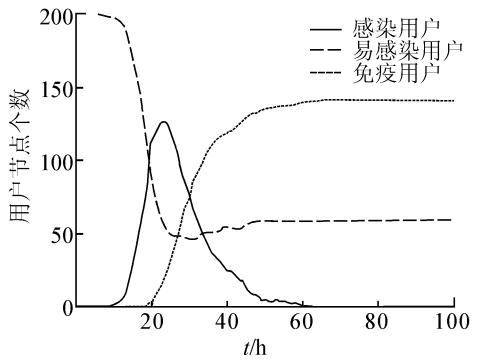
平均路径不断减小;因此,模型传播的范围将不断扩大,速度也会明显加快.



(a)第 1 次实验



(b)第 2 次实验



(c)第 3 次实验

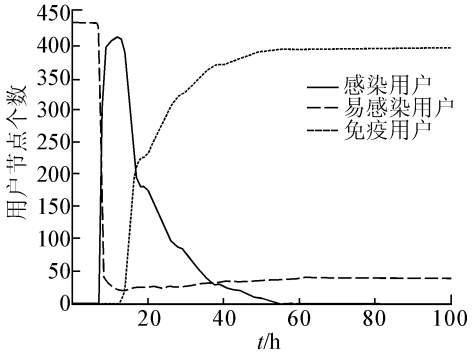
图 4　随机网络上的仿真结果

Fig.4　Simulation results under random network

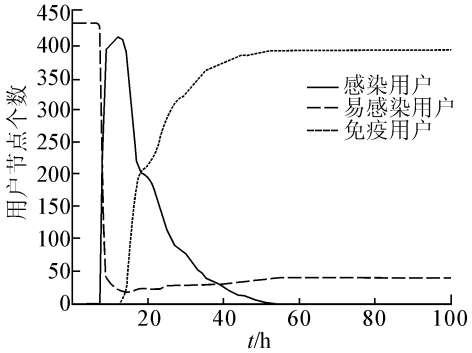
通过不同网络在相同的参数配置下得到的不同仿真实验结果说明,CSIRS 模型不仅能呈现出一般传染病动力模型所呈现的传播模式,还能够呈现出多个波动、小范围长时间传播、快速上升缓慢下降等社会网络话题传播模式.这充分说明了考虑话题特性、用户对话题的敏感程度等因素能够较好地刻画社会网络上的热点话题传播模式.

2.3　真实网络上的传播实验

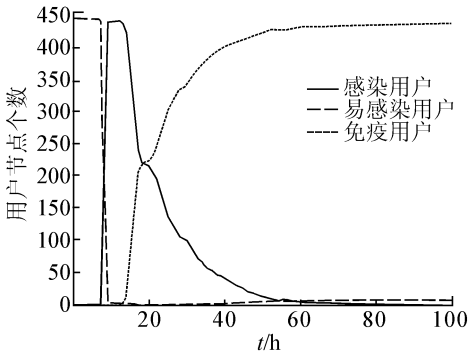
为了评估 CSIRS 模型在真实网络上的效应,从人人网上采集了一个用户的真实社会网络,网络中的节点个数为 445 个.真实社会网络具有无标度和小世界的混合特性,采用与仿真实验相同的参数设置.实验结果如图 5 所示.



(a)第 1 次实验



(b)第 2 次实验



(c) 第 3 次实验

图 5　真实社会网络上的实验结果

Fig.5　Experimental results under real social network

从图 5 可以看出,真实社会网络上 CSIRS 模型呈现出基本一致的传播模式,其差异在于传播过程的上升速度更快,以及传播下降过程中具有局部波动和更明显的长尾效应.CSIRS 模型在真实社会网络上的传播速度非常快,超过了无标度网络和小世界网络下的传播速度,其原因在于实验采集的是一个用户真实的社会网络,平均路径较短,因此很容易形成话题的快速漫延.CSIRS 模型下降方式呈现先快后慢的特征,这与现实话题传播一致.另外如图 5(c)所示,CSIRS 模型在真实社会网络上的传播长尾效应要大于小世界网络,这也符合话题在现实网络上可能被少数用户在较长时间内局部传播的情况.

3 结 论

本文基于内容构建了一个话题传播模型(CSIRS),在无标度网络、小世界网络、随机网络以及真实社会网络上做了不同的实验,研究结果表明:

1)CSIRS模型在不同结构的复杂网络上呈现丰富的传播特征,尤其是在无标度和真实社会网络上的传播能体现出真实话题传播的特征;

2)CSIRS模型在小世界网络和真实社会网络上的传播模式基本一致,这说明在具有小世界特征的复杂网络上的话题传播具有相似的模式;但在实际社会网络中话题传播的形态还有很多,还需要进一步在模型中加入不同的因素.

现实的社会网络中同时具有无标度、小世界等特性,其热点话题的传播还受到外部因素、内部相互激励、噪声等多种因素的影响,因此如何融合更多的因素,构建更加精确的模型,在更多实际的网络上进行实验都是值得研究的问题.

参考文献:

[1]刘晓凤,黄刘生,吴俊敏,等.传感网络定位中基于传染病模型的信息传播研究[J].小型微型计算机系统, 2009, 30(4): 647-651.
LIU Xiaofeng, HUANG Liusheng, WU Junmin, et al. Study of the false information spread in localization algorithms for wireless sensor networks using epidemic model[J]. Journal of Chinese Computer Systems, 2009, 30(4): 647-651.

[2]张彦超,刘云,张海峰,等.基于在线社交网络的信息传播模型[J].物理学报, 2011, 60(5): 050501.
ZHANG Yanchao, LIU Yun, ZHANG Haifeng, et al. The research of information dissemination model on online social network[J]. Acta Physica Sinica, 2011, 60(5): 050501.

[3] KEELING M J, EAMES K T D. Networks and epidemic models[J]. Journal of the Royal Society Interface, 2005, 51(2): 295-307.

[4] RILEY S. Large scale spatial transmission models of infectious disease[J]. Science, 2007, 316(5829): 1298-1301.

[5] NEWMAN M E J. Spread of epidemic disease on networks [J]. Physical Review E, 2002, 66(1): 1103-1115.

[6] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks [J]. Nature, 1998, 393 (6684): 440-442.

[7] MAY R M, LLOYD A L. Infection dynamics on scale-free networks[J]. Physical Review E, 2001, 64(6): 066112.

[8] LLOYD A L, MAY R M. How viruses spread among computers and people[J]. Science, 2001, 292(5520): 1316-1317.

[9] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic spreading in scale-free networks[J]. Physical Review Letters, 2001, 86(14): 3200-3203.

[10] MATSUBARA Y, SAKURAI Y, PRAKASH B A, et al. Rise and fall patterns of information diffusion: model and implications[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 6-14.

[11] HAWKES A G, OAKES D. A cluster representation of a self-exciting process[J]. Journal of Applied Probability, 1974, 2(11): 493-503.

[12] CRANE R, SORNETTE D. Robust dynamic classes revealed by measuring the response function of a social system[J]. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653.

[13] STOMAKHIN A, SHORT M B, BERTOZZI A L. Reconstruction of missing data in social networks based on temporal patterns of interactions[J]. Inverse Problems, 2011, 27(11): 115013.

[14]赵丽,袁睿翕,管晓宏,等.博客网络中具有突发性的话题传播模型[J].软件学报, 2009, 20(5): 1384-1392.
ZHAO Li, YUAN Ruixi, GUAN Xiaohong, et al. Bursty propagation model for incidental events in blog networks [J]. Journal of Software, 2009, 20(5): 1384-1392.

[15] LESKOVEC J, MCGLOHON M, FALOUTSOS C, et al. Cascading behavior in large blog graphs[J]. 2007 SIAM International Conference on Data Mining. Minneapolis, USA, 2007, 551-556.

[16] KUMAR R, MAHDIAN M, MCGLOHON M. Dynamics of conversations[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA: ACM, 2010: 553-562.

[17] PRAKASH B A, CHAKRABARTI D, FALOUTSOS M, et al. Threshold conditions for arbitrary cascade models on arbitrary networks[C]//2011 IEEE 11th International Conference on Data Mining. Vancouver, Canada, 2011: 537-546.

[18] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]//2010 IEEE 10th International Conference on Data Mining. Sydney, Australia, 2010: 599-608.

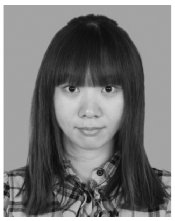
[19]王延,郑志刚.无标度网络上的传播动力学[J].物理学报, 2009, 58(7): 4421-4425.
WANG Yan, ZHENG Zhigang. Spreading dynamics on scale-free networks[J]. Acta Physica Sinica, 2009, 58 (7): 4421-4425.

[20] NEWMAN M E J, WATTS D J. Scaling and percolation in the small world network model[J]. Physical Review E, 1999, 60(6): 7332-7342.

作者简介:

韩忠明,男,1972年生,副教授,博士后,主要研究方向为数据仓库与数据挖掘、生物信息学.主持和参与国家自然科学基金、教育部人文社科基金等项目多项,发表学术论文50余篇.

张慧,女,1989 年生,硕士研究生,主要研究方向为互联网检索与数据挖掘、社会网络.



张梦,女,1991 年生,硕士研究生,主要研究方向为互联网检索与数据挖掘、社会网络.