

DOI: 10.7652/xjtuxb201702001

在线社会网络中多话题竞态传播分析与建模

周亚东¹, 刘丽丽¹, 张贝贝², 雷蕾³

(1. 西安交通大学智能网络与网络安全教育部重点实验室, 710049, 西安; 2. 西安理工大学计算机科学与工程学院, 710048, 西安; 3. 西安理工大学水利水电学院, 710048, 西安)

摘要: 针对在线社会网络中多个话题在传播过程中呈现出的竞争状态,进行了竞态传播过程的测量与分析,并建立了多话题竞态传播模型。基于多个话题数据进行了话题参与用户行为分析,发现较少用户会持续关注同一个热点话题,并且会有一定数量用户在多个同类话题间转移关注,从而使多个同类话题在并行传播时对吸引用户参与呈现出竞争态势。在分析结果的基础上,建立了考虑话题之间相互影响力以及话题吸引度的多话题竞态传播模型,该模型可有效描述多个同类热点话题在同时间段出现时各个话题之间的相互影响情况,以及各个话题在传播过程中人群参与规模的变化情况。在与实际数据的对比实验中,模型仿真结果的平均峰值出现时间的误差为 0.2 d,平均传播周期的误差为 2.4 d,话题间用户平均转出比例的误差为 1.2%,并且能复现参与人数的单峰性、长尾特性等话题传播的动态特性。上述实验结果表明,该模型可有效描述在线社会网络中的多话题竞态传播动态过程。

关键词: 在线社会网络;热点话题;传播模型;用户行为分析

中图分类号: TP393 **文献标志码:** A **文章编号:** 0253-987X(2017)02-0001-05

Analysis and Modeling for Competitive Diffusion of Multiple Topics in Online Social Networks

ZHOU Yadong¹, LIU Lili¹, ZHANG Beibei², LEI Lei³

(1. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; 3. Institute of Water Resources and Hydro-Electric Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Competitive diffusion processes are analyzed, and a model for the competitive diffusion of multiple topics is proposed to understand the competitive diffusion of multiple topics in online social networks. Analyzing real network data, it is found that just a few users continue to focus on the same topic, and many users transfer their attentions among multiple similar topics. In this case, similar topics attract users to participate in a competitive situation. Based on the analysis results, a diffusion model describing the dynamic diffusion process of multiple topics in the competitive situation is proposed. Experiment validates that the model reproduces the unimodality and long tail characteristics of dynamic changes of users who participate in multiple topics, and achieves a good performance. A comparison with actual data show that the error of average peak time of the proposed model is 0.2 day, the error of average diffusion period is 2.4 days, and the error of average transferring proportion among topics is 1.2%. These results show

收稿日期: 2016-10-07。 作者简介: 周亚东(1982—),男,讲师;张贝贝(通信作者),男,讲师。 基金项目: 国家自然科学基金资助项目(61572397,61502383,61571360);陕西省自然科学基金基础研究计划资助项目(2015JM6298)。

网络出版时间: 2016-12-30

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20161230.1629.002.html>

that the model can effectively describe the competitive diffusion of multiple topics in online social networks.

Keywords: online social network; hot topics; diffusion model; user behavior analysis

近年来,微博、社交网站、论坛等在线社会网络站点已成为网络用户公开传播和获取信息的重要平台,与新闻网站、电视、报纸等传统媒体相比,在传播公共热点信息方面更为高效,从而极大地改变了公众获取与传播信息的方式,同时也吸引国内外研究者对在线社会网络中的信息传播规律与应用展开了大量研究。对于在线社会网络中的信息传播过程,研究者通常以热点话题传播规律为主要研究问题,针对信息传播的测量分析^[1]、信息传播的影响因素^[2]、信息传播建模^[3]、信息传播最大化^[4]等问题取得了大量研究成果。在这些问题中,信息传播的分析与建模是揭示信息传播规律,并为各类基于信息传播的应用^[5]提供理论基础的重要研究问题。

在信息传播的分析与建模方面,已有研究成果主要基于话题信息内容、用户行为和社会网络拓扑结构等,分析话题相关信息传播过程,建立描述参与人群、传播路径等方面变化的演变模型。例如,Jure等对实际在线社会网络中的话题信息传播进行了测量分析,发现话题信息在传播过程中具有单峰性、快速传播性及长尾效应等特点^[6]。Xiong等对用户的浏览行为和转发行为做了深入研究,提出了考虑用户在浏览信息前后状态变化机制的信息传播模型^[7]。Yang等在SIS模型^[8]的基础上,提出了可根据当前微博信息的流行度预测未来某一时间话题信息传播规模的线性影响模型^[9]。Farajtabar等基于用户创建新链接的行为和此行为引起的网络拓扑改变,提出了一种概率生成模型,可模拟话题信息传播的路径演变过程^[10]。Bourigault等将用户嵌入潜在空间并计算用户传播信息的概率,从而提出了一种可描述信息传播基本过程的嵌入式级联模型^[11]。

然而,包括上述研究成果在内,目前已有的信息传播的分析与建模研究是基于话题信息在传播过程中彼此独立的假设,即只考虑单个话题的传播过程,而不考虑多个话题在同一时间、同一网络中传播时的相互影响情况。由于网络信息量的爆炸式增长,严重的信息过载会导致话题之间对用户关注度存在激烈竞争,因此多个话题在竞争状态下的传播情况应是非常值得研究的问题。

基于新浪微博API,本文通过随机选取粉丝数较大的若干节点,采用“滚雪球”的方式采集了大量

数据。数据集中包括1 797 941位用户6个月的微博数据,共有69 450 325条微博文本。在此数据集中,选取36项参与人数最多的话题作为研究数据。

基于该数据集分析发现,同类型话题之间在吸引用户参与讨论方面产生了比较显著的相互影响。通过对用户参与多个热点话题信息讨论和传播的过程进行分析,发现在一个时间段内用户在单个话题上参与的次数和持续参与时间较少,并且还发现用户在同类话题间以比较大的比例转移参与,而这一现象在不同类话题间不显著。基于以上分析,我们提出了一种考虑话题之间相互影响力以及话题吸引力的多话题竞态传播模型,可用于有效描述多个同类热点话题在同时间段出现时,各个话题之间的相互影响情况以及各个话题在传播过程中人群参与规模的变化情况。

1 多热点话题传播中用户行为分析

1.1 参与热点话题用户的行为特征分析

从用户行为角度分析用户在参与热点话题时的发文数量、间隔时间和持续时间等重要行为特征,结果如图1~图3所示,图中均使用对数坐标系。图1是用户在单个热点话题的发文数的互补累计分布函数图,分析发现90%以上的参与热点话题用户在单个话题的发文数量仅为1条,这表明用户参与单个热点话题的发文数量十分有限,即用户在对某热点话题发文之后,对其兴趣会降低,因此大部分用户不会在同一热点话题上多次发文。图2给出了用户在参与的单个热点话题上的发文间隔时间的分布,分析发现用户再次参与同一个话题的可能性会随时间骤减,即用户参与某个话题后,间隔时间越长,越不可能再就该话题发文。图3给出了用户在同一话题上连续发文持续时间的分布,分析发现90%的热点话题用户在单个话题上的发文持续时间为1 d,即大部分用户在参与一个热点话题之后兴趣发生转移,进而参与其他话题或不再发文。

基于图1~图3所呈现的分析结果可知,用户在参与热点话题发文时,对同一话题的参与次数、间隔时间和持续时间均表明大部分用户的兴趣点不会持续长时间停留在某个固定话题上,而是会在较短的时间内发生转移。

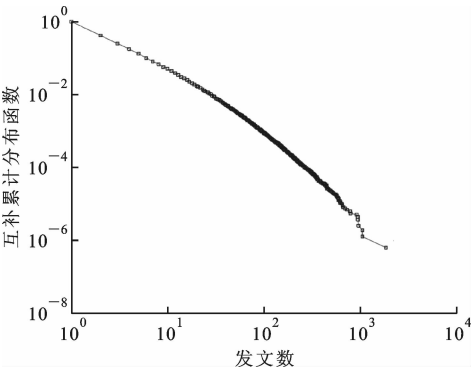


图 1 用户在单个热点话题上的发文数分布

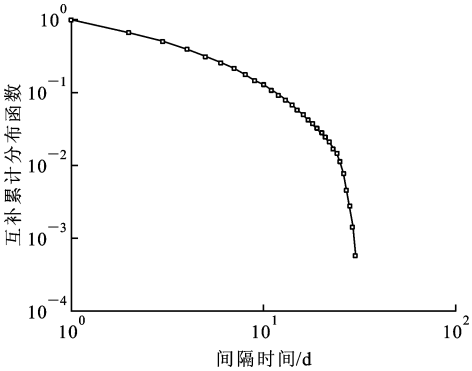


图 2 用户在同一个热点话题上的发文间隔时间分布

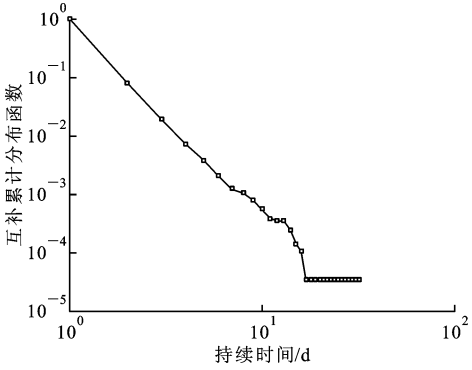


图 3 用户在同一个热点话题上的连续发文持续时间分布

1.2 用户参与多个热点话题的行为特征测量

对于用户兴趣点的转移情况,分析了用户在退出热点话题后参与其他同类和不同类话题的转入转出比例分布,发现用户在同类热点话题间的转移比例远大于在不同类热点话题之间的转移比例。图 4 给出了退出话题用户转向其他同类和不同类话题的比例分布,分析发现 50% 以上的热点话题中有 10% 以上的用户会转移到其他同类话题,而仅有约 1% 的热点话题中有 10% 以上的用户会转移到其他非同类热点话题。

上述分析表明,在同类话题间存在较显著的竞态传播现象,即用户参与话题时会被同类话题吸引,

从而使得多个同类话题在并行传播时对用户参与的资源呈现出竞争态势。同时,此现象在多个不同类话题间并不显著存在。

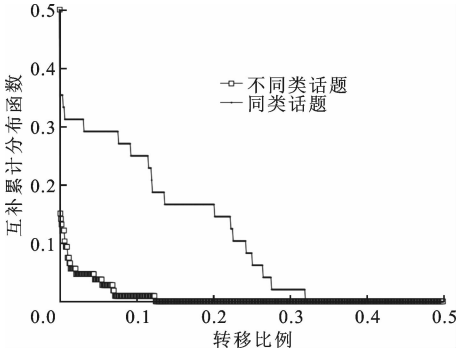


图 4 退出话题用户的转移比例的分布

2 社会网络中多话题竞态传播模型

为刻画多个同类热点话题竞态传播时各热点话题传播规模的动态演变过程,基于传染病动力学模型,本文提出一种考虑话题之间相互影响力以及话题吸引度的多话题竞态传播模型。基于上述实际数据的测量分析结果,该模型充分考虑了用户参与话题的发文数与话题出现时长呈现幂率分布、用户参与同一个话题的发文数有限、存在部分用户会多次参与同一个话题讨论等用户行为特点,多个同类话题同时传播时的竞争态势,以及不同类话题相互基本不存在竞争的现象,比现有的单话题模型更能准确描述话题的实际传播情况。

令 $S(t)$ 和 $S(t-1)$ 分别表示在 t 时刻和 $t-1$ 时刻所有对话题有兴趣但还未参与话题讨论的用户数(初始状态集), $I_1(t), I_2(t), \dots, I_n(t)$ 表示在 t 时刻参与话题 1, 2, $\dots, n$ 讨论的用户数(参与状态集), $R_1(t), R_2(t), \dots, R_n(t)$ 表示 t 时刻退出话题 1, 2, $\dots, n$ 讨论的用户数(退出状态集)。考虑到话题传播过程可认为是离散时间动态过程,模型的数学形式以差分方程的形式表达,如式(1)~式(3)所示。

$$S(t) - S(t-1) = - \sum_{i=1}^n \varphi(S) I_i(t-1),$$
$$\varphi(S) = k_i S \tag{1}$$

式中: $\varphi(S)$ 为话题接触感染率,即认为对话题 i 感兴趣的用户群体 S 中每个用户在接触到话题 i 时,平均会有 $\varphi(S)$ 的概率接触到话题 i 并进行传播。显然在 $t-1$ 时刻, $S(t-1)$ 中用户能接触到话题的概率既取决于 $S(t-1)$ 的大小,又取决于话题 i 的参与状态集 $I_i(t-1)$ 的大小,还取决于话题的吸引度 k_i , 因此易感人群中在 t 时刻的用户流失数即为被各话

题吸引参与的用户总数

$$I_i(t) - I_i(t-1) = \varphi(S)I_i(t-1) - \gamma_i I_i(t-1) + \alpha_i R_i(t-1) + \alpha'_i (R(t-1) - R_i(t-1)) + \sum_{j \neq i} \beta_{ji} I_j(t-1) - \sum_{i \neq j} \beta_{ij} I_i(t-1) \quad (2)$$

$$R_i(t) - R_i(t-1) = \gamma_i I_i(t-1) - \alpha_i R_i(t-1) - \alpha'_i (R(t-1) - R_i(t-1)) \quad (3)$$

式中: α_i 表示 t 时刻话题 i 退出状态集 $R_i(t-1)$ 中的用户回到话题参与状态集 $I_i(t)$ 中的概率, 即话题 i 退出用户的激活率; α'_i 表示 t 时刻其他话题退出状态集 $R(t-1) - R_i(t-1)$ 中的用户参与热点话题 i 的概率, 即话题 i 对其他话题退出用户的吸引力; $R(t-1)$ 表示 $t-1$ 时刻全部话题的退出状态集, $R_i(t-1)$ 表示 $t-1$ 时刻话题 i 的退出状态集; β_{ij} 表示 $t-1$ 时刻话题 i 中参与用户在 t 时刻参与话题 j 讨论的话题转移率, 即话题 j 对话题 i 参与用户的吸引力; γ_i 表示 t 时刻话题 i 参与用户的退出率, 即 t 时刻不再参与话题 i 讨论的概率。

式(3)表示的是话题 i 的退出状态集的变化情况。该式描述了以下情况: 当用户在参与话题 i 讨论后, 会以 γ_i 的概率在下一时刻不再参与话题 i 讨论, 进入话题退出状态, 同时位于话题退出状态的用户会以 α_i 的概率再次参与话题 i 讨论, 也会以概率 α'_i 参与其他话题的讨论, 因此话题 i 的退出状态集的变化取决于这 3 方面因素。

此外, 位于各个状态的用户集不能小于 0, 因此各状态还具有约束条件

$$S(t) \geq 0; I_i(t) \geq 0; R_i(t) \geq 0 \quad (4)$$

对于式(1)~式(3)中的参数 k_i 、 α_i 、 α'_i 、 β_{ij} , 根据测量分析显示其应为变量, 并且是与话题传播时间长短存在近似服从指数衰减的非线性关系, 因此对这些参数做出假设

$$\left. \begin{aligned} k_i(t) &= k_{i0} q_1^{t-t_0} \\ \alpha_i(t) &= \alpha_{i0} q_2^{t-t_0} \\ \alpha'_i(t) &= \alpha'_{i0} q_3^{t-t_0} \\ \beta_{ij}(t) &= \beta_{ij0} q_4^{t-t_0} \\ \gamma(t) &= 1 - q_5^{t-t_0} \end{aligned} \right\} \quad (5)$$

为便于理解, 以话题 I_1 的 t 时刻传播过程为例, 以上数学过程可以表示为如图 5 所示的传播模型。

3 实验分析

为了验证多个同类热点话题信息的传播过程是否与实际话题传播过程相符, 对上述推导出的模型

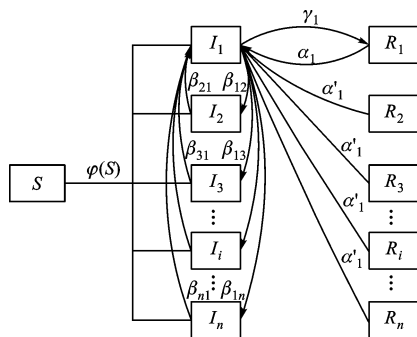


图5 多话题竞态传播模型

进行实验分析。在实验中, 考虑到实际网络中用户参与热点话题的情况, 对各参数和话题规模做出相关假设。

对于模型的各项相关参数, t 表示热点话题 i 出现后的时间, 并且在进行仿真实验时需给定参数 k_i 、 α_i 、 α'_i 、 β_{ij} 的初始值 k_{i0} 、 α_{i0} 、 α'_{i0} 、 β_{ij0} 。在实验中, 依据各项参数的物理意义, 分别基于实际数据进行估计。模型参数中, k_{i0} 、 α_{i0} 、 α'_{i0} 、 β_{ij0} 的估计值分别为 0.000 4、0.14、0.28、0.06, q_1 、 q_2 、 q_3 、 q_4 、 q_5 的估计值分别为 0.3、0.1、0.73、0.98、0.1。例如, 对于话题转移率 β_{ij0} 及其对应的参数 q_4 , 通过测量实际数据中各个话题参与用户在传播过程中的转移率, 并采用最小二乘法拟合得到 β_{ij0} 与 q_4 的估计值为 0.06 和 0.98。

实际网络中, 每天都有许多热点话题出现, 这些话题的初始出现时间和当天参与人数具有随机性。关于热点话题出现时间及初始参与人数的分布形式, 已有研究中并未有定论, 常假设为均匀分布和多项分布等分布形式。在本文实验中, 主要为验证所提模型能否有效描述同时传播的多个热点话题之间的相互影响, 而话题出现时间及初始参与人数的具体分布形式对此影响不大, 因此为简化实验过程, 假设二者均服从最简单的均匀分布, 并假设热点话题的出现时间在 $[0, 30]$ 区间内均匀分布, 话题初始参与人数在 $[0, 100]$ 区间内均匀分布。

基于以上参数设定结果, 对本文所提出的多话题竞态传播模型进行实验分析, 得到各话题在设定时间内参与话题的人数的实时变化情况, 模型实验结果如图 6 所示, 实际多热点话题传播结果如图 7 所示。对比图 6 和图 7 分析表明, 各话题的传播过程中参与该话题讨论的用户数先迅速增加后迅速减小, 且部分话题后期衰减较慢, 具有长尾效应; 而在实际微博网络中, 热点话题也满足单峰条件, 且由于部分用户长时间参与某个话题的讨论, 因此实际微博网络中话题传播也会表现出长尾分布情况。经过

对比可知,仿真结果可较好地拟合实际网络中话题的传播过程。为分析模型是否可描述多个同类型话题之间在吸引用户参与讨论方面的相互影响,将模型中的用户转出到其他话题中的比例与实际话题中用户转出比例的累积概率分布进行对比,如图 8 所示。从对比结果可以看出,模型实验中的转出比例对实际话题的转出比例分布拟合较好。

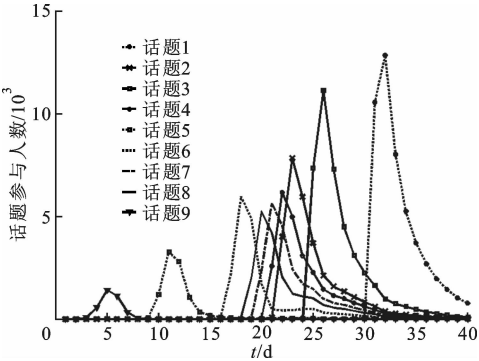


图 6 多话题竞态传播模型的实验结果

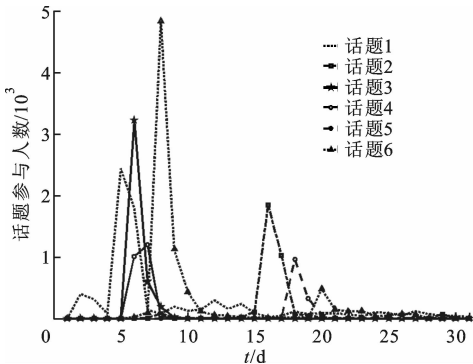


图 7 实际网络中的热点话题传播结果

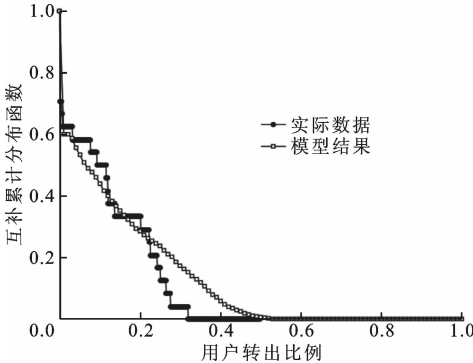


图 8 实际数据与模型结果的用户转出比例对比

为进一步分析模型对话题实际传播过程的描述能力,采用多个话题的平均峰值出现时间、平均传播周期和话题间的用户平均转出比例 3 项指标,对实际数据与模型结果进行了量化对比分析,结果见表 1。通过分析对比结果可知,所提模型能够有效描述实际话题传播过程的动态特性,并且描述话题间用

户转移的误差也较小。

表 1 模型结果与实际数据的对比

对比指标	平均峰值出	平均传播	话题间的用户平
	现时间/d	周期/d	均转出比例/%
实际数据	1.7	20.7	15.9
模型结果	1.9	23.1	14.7
绝对误差	0.2	2.4	1.2

4 结 论

本文通过对多个热点话题参与用户行为进行分析,发现参与热点话题讨论的用户具有单个话题参与次数少、易转移、参与话题类型集中的特点,进一步分析发现话题具有同类话题转移比例大,不同类话题转移比例小的特点,由此发现了在线社会网络中多个话题共同传播时存在竞态传播现象。在此基础上,提出了一种考虑话题之间相互影响力以及话题吸引力的多话题竞态传播模型。通过将实验仿真分析结果与实际数据分析结果相对比,验证了该模型可较好地描述实际话题传播过程中的单峰性、长尾特性,以及同类话题间相互影响的特性。与已有的单话题传播模型相比,该模型更贴合实际网络中的话题传播状况,可为网络舆情分析、在线网络营销等领域的应用提供更好的理论支撑。

参考文献:

[1] CHA M, MISLOVE A, GUMMADI K P. A measurement-driven analysis of information propagation in the Flickr social network [C]//Proceedings of the 18th International Conference on World Wide Web. New York, USA: ACM, 2009: 215-221.

[2] 王晨旭,管晓宏,秦涛,等. 微博消息传播中意见领袖影响力建模与应用研究 [J]. 软件学报, 2015, 26 (6): 1473-1485.

WANG Chenxu, GUAN Xiaohong, QIN Tao, et al. Modeling on opinion leader's influence in microblog message propagation and its application [J]. Journal of Software, 2015, 26(6): 1473-1485.

[3] GUILLE A, HACID H, FAVRE C. Information diffusion in online social networks: a survey [J]. ACM Sigmod Record, 2013, 42(2): 17-28.

[4] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network [J]. Theory of Computing, 2015, 11(4): 105-147.

(下转第 39 页)