doi:10.3969/j.issn.1000-436x.2013.z1.011

基于用户与节点规模的微博突发话题传播预测算法

王巍[1]，李锐光[2]，周渊[2]，杨武[1]
（1. 哈尔滨工程大学 信息安全研究中心，黑龙江 哈尔滨 150001; 2. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：突发话题传播建模与预测的主要目的是对网络中可能产生不良影响的、紧急性突发事件的后续传播进行控制。目前微博网络中的话题传播与预测研究尚处于起步阶段。通过对病毒传染模型、消息传播模型以及话题传播模型的深入研究，提出一种基于微博粉丝关系、用户活跃度和影响力的话题传播模型，将微博用户集合划分为感染用户、易染用户和免疫用户，分析感染用户和易染用户的粉丝关系，预测下个时间窗口内被感染的用户规模。沿用话题传播模型研究中的“内外场强”概念，通过研究发现“内场强”和“外场强”有特定的比例关系，基于用户群的规模大小，分别提出基于用户和节点规模的话题传播预测算法。相关实验表明，基于用户的算法预测更为准确但是时间复杂度较高，基于节点规模的算法则更适合大规模数据集的处理。

关键词：微博网络；话题传播；传播预测；节点规模
中图分类号：TP393.08 **文献标识码**：A **文章编号**：1000-436X(2013)Z1-0084-08

Microblog burst topic diffusion prediction algorithm based on the users and node scale

WANG Wei[1], LI Rui-guang[2], ZHOU Yuan[2], YANG Wu[1]
(1. Information Security Research Center, Harbin Engineering University, Harbin 150001, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, China)

Abstract: The main purpose of burst topic diffusion modeling and prediction is to control the subsequent large-scale dissemination of emergency incidents with adverse effect. Currently microblog topic diffusion and prediction is still in its infancy. The viral infection model, the message propagation model and topic propagation model were deeply studied and a topic diffusion model was proposed based on fans relationship, user activity and influence. By partitioning microblog users into infected users, tangible user and immune user, the relationship between infected and tangible user was analyzed to predict the scale of users which were infected in next time window. Following "internal and external field strength" concept in topic diffusion model, the proportional relationship between them was studied. Based on the scale of the user, topic diffusion prediction algorithms were proposed based on user and node scale respectively. Experiments show that the former can predict diffusion more accurately but with bad time complexity, and the latter node is more suitable for processing large data sets.

Key words: microblog network; topic diffusion; diffusion prediction; node scale

1 引言

随着社交网络越来越成熟，更多的人群开始习惯通过网络参与话题的讨论，并进行传播。而微博作为目前最为流行的社交媒体尤为受到关注。微博突发话题传播预测研究主要是为了遏制可能造成不良影响的突发事件的传播，在突发事件传播的早期对其进行控制。基于微博的突发事件传播预测主

收稿日期：2013-05-02
基金项目：国家自然科学基金资助项目（61170242, 61272536）；国家高技术研究发展计（“863”计划）基金资助项目（2012AA012802）；中央高校基本科研业务费专项基金资助项目（HEUCF100601）
Foundation Items: The National Natural Science Foundation of China (61170242, 61272536); The National High Technology Research and Development Program of China (863 Program)(2012AA012802); Fundamental Research Funds for the Central Universities(HEUCF100601)

要是通过分析微博中的用户关系对话题中消息的可能传播路径进行评估，从而给出某个突发话题在下个时间段的传播规模。

话题是消息的集合，所以所谓话题的传播其实就是消息集合的传播。关于社交网络中的话题传播相关研究已经成为热点[1,2]。本文提出的话题传播模型很大程度上借鉴了病毒传播的相关研究成果[3]，以形成微博网络的话题传播模型。现有突发话题传播预测研究尚处于初级阶段，2009 年清华大学的赵丽[4]等人提出一个基于博客网络的突发话题传播模型。2011 年重庆大学的孙留东[5]在赵丽提出的模型上进行了改进，使模型更符合实际的博客网络。2010 年 HE D[6]将突发的结果看成一个话题的动态过程，允许分层的突发模型。

在其他博客的突发话题传播及预测的相关研究中，大部分都是基于博客的拓扑结构来分析的[7]，和话题的具体内容以及博客用户之间的关系没有很大的关联，和本文的研究内容偏差较大。此外比较常见的分析话题传播规律的方法是根据话题包含的消息的传播给出的，即由改进的消息传播模型对话题的传播进行粗略的传播建模与预测。

本文通过对病毒传染模型、消息传播模型以及话题传播模型的深入研究，提出一种基于微博粉丝关系、用户活跃度和影响力的话题传播模型，通过将微博用户集合划分为感染用户、易染用户和免疫用户，分析感染用户和易染用户的粉丝关系，预测下个时间窗口内被感染的用户规模。

此外，本文沿用话题传播模型研究中的“内外场强”概念，通过研究发现“内场强”和“外场强”有特定的比例关系，基于用户群的规模大小，提出基于用户的话题传播预测和基于规模的话题传播预测。

2　微博突发话题传播

本文通过分析微博突发话题传播规律，总结出影响话题传播的相关参数，具体如下。

话题感染力：表示当话题推送给用户时，话题导致该用户转发或者发布相关微博的力度。

话题知名度：也称作话题的影响力，表示在本时间段内能够看到该话题相关微博的用户数。

话题热度：话题在当前窗口的突发权值。

用户知名度：表示在突发话题传播中用户的影响力。

用户活跃度：表示用户在本时间段内能够看到话题的概率，或者说是用户看到话题的次数。

用户感染度：表示用户在被感染力度大于其接收力时才会转发或发布相关微博。

通过总结微博本身的特性，结合话题传播的相关规律，对微博中突发话题的传播规模进行预测。

下面首先通过具体的实例来分析微博中突发话题传播的一些特性。

突发话题的传播大致符合如图 1 所示的话题传播曲线。在话题突发后很短的时间内，话题的传播规模达到顶峰，随后传播的规模随着时间推移逐渐减小，直到话题不再有影响力为止。

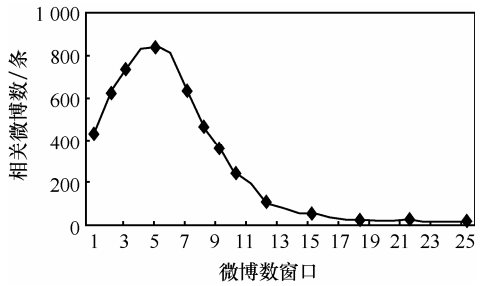

图 1　突发规模小的话题传播曲线

对于突发规模大的突发事件，话题传播的时间跨度会相应地增加，主要表现为如图 2 所示的传播曲线，话题的规模在达到巅峰后很长的一段时间内影响力不减。

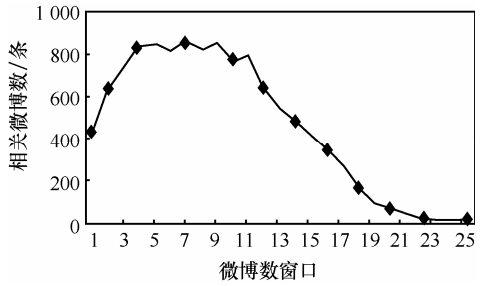

图 2　突发规模大的话题传播曲线

此外，还存在一些突发性强的话题，其影响规模会出现 2 个或更多个峰值的情况，原因是这些话题在传播过程中发生漂移，加入了一些新的突发特征词，这些话题传播可能会呈现如图 3 所示的传播曲线。从表现来看，这种情况不符合传统的话题传播规律，因为它和认知中突发话题的传播曲线形式相差很大，用常规的模型无法模拟

出这种话题的演变情况，这种情况可以看做是 2 个突发话题在某个阶段的重叠，本文采用突发特征词表述话题传播模型中的突发话题，如果在话题传播过程中话题加入了新的突发词，就代表有一个新的突发话题和原有话题共同作用于话题传播模型。

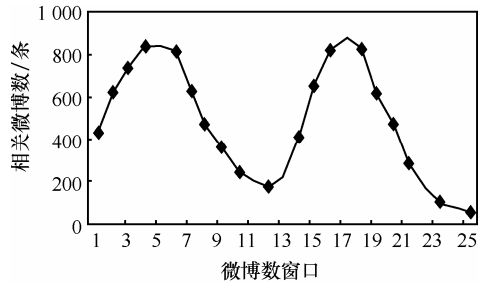

图 3 具有演变特性的话题传播曲线

再来分析话题感染力。通过图 1 和图 2 所示的话题传播曲线，很容易误认为话题的感染力是一条先急剧上升再缓慢下降的类抛物线。其实不然，通过对实际情况进行分析，话题的感染力应该是随时间推移一直减小的，表示为如图 4 所示的话题感染力衰减曲线。如此可能会有一个疑问，假如在突发话题传播的开始阶段话题感染力是随时间推移逐渐下降的，那么突发话题的传播规模为何会先增加再减少，甚至在传播开始阶段规模呈现几何级数。出现这种情况的原因是传播是累加的过程，虽然话题感染力在话题传播的起始阶段减小了，但是突发话题的转发人数却在急速地增加，当未感染节点在短时间内频繁地接收到其他用户推送给他的相同话题时，即使此话题的感染力在减小，但是多个感染力的累加导致微博用户大量地转发相关微博。

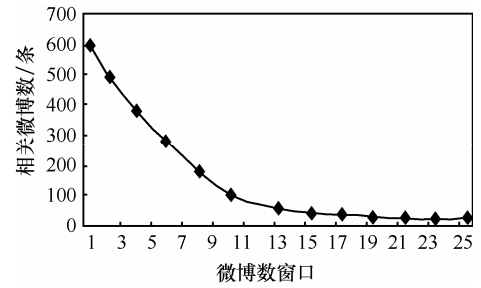

图 4 突发话题感染力衰减曲线

3 基于用户的微博突发话题传播预测

3.1 突发话题传播模型建立

面向微博的消息传播模型可以用一颗树表示，根节点为原始消息，通过微博中的粉丝关系、转发关系和评论关系生成一颗消息传播树，话题的传播可以看做是消息集合的传播，描述如图 5 所示。

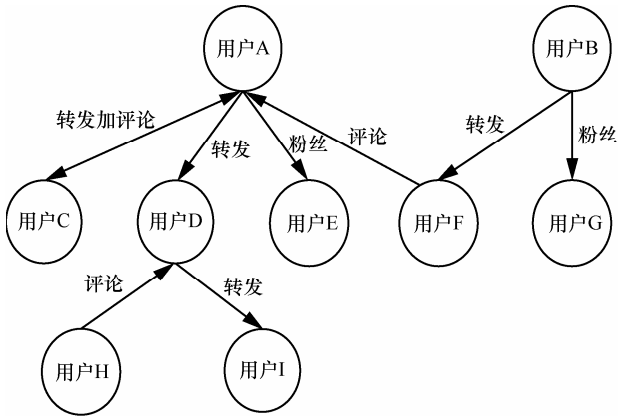

图 5 微博用户关系

其中用户 A 和 B 为初始感染节点，C、D、E、F 均为 A 的粉丝，其中 C 和 D 转发了 A 发布的话题相关微博，C 在转发的同时还对微博进行了评论，F 评论了此微博，而且 F 同时也是用户 B 的粉丝，他还转发了 B 发布的话题相关微博，E 代表既未转发也未评论该条微博的用户 A 的粉丝，用户 G 是 B 的粉丝，H 和 I 是 D 的粉丝。

本文结合病毒传染模型、线性阈值模型以及赵丽提出的话题传播模型，提出一个更适合描述微博突发话题传播规律的话题传播模型。首先把微博网络中的所有用户节点分为 3 个部分：感染用户、易染用户和免疫用户。

其中感染用户和免疫用户是很容易界定的。感染用户是在本窗口前已经发送话题相关微博的用户，而免疫用户是通过本文传播模型预测不会发送相关微博的用户。

易染用户的阶段较为复杂，具体描述如下。

$$Y_i(window_n)=\begin{cases}0,\text{在第}n\text{个窗口节点}i\text{是易染节点}\\1,\text{在第}n\text{个窗口节点}i\text{是感染节点}\\2,\text{在第}n\text{个窗口节点}i\text{是免疫节点}\end{cases} \quad (1)$$

突发话题跟踪的初始阶段，微博网络中只出现感染用户和易染用户，其中感染用户集包括突发话题检测中发布相关微博的用户，剩余所有用户均为易染用户。跟踪窗口内每个易染用户都更新 2 个属性值——感染度值和活跃度值。活跃度值是由该用户平均发布微博的时间间隔决定，对于特定窗口中

的特定用户，其活跃度值是一个定值，感染度值是指特定话题对用户的感染度，包括微博中其他用户对其影响力和微博外现实世界对其影响力 2 部分，即话题感染的“内场强”和“外场强”。“内场强”主要是由用户关注的感染用户产生，由感染用户的影响力、突发话题的突发权值以及一个衰减系数 3 部分决定；“外场强”是指来自现实世界的影响，是由该话题本身的特性和一个衰减系数来决定。这里有 2 个衰减系数：话题感染力衰减系数用 p_{topic} 表示，用于描述突发话题感染力随时间逐渐减小的特性；用户感染度值衰减系数用 p_{user} 表示，用于描述突发话题对用户在相邻2个窗口的感染度值的衰减情况。

假设用户 i 在第 n−1 个窗口发布了指定话题相关微博，在第 n 个窗口该用户就会对其所有粉丝产生一个感染度值，即该用户的所有粉丝都会有一定的概率来转发或者评论此微博。首先，忽略突发话题给微博消息的传播带来的影响，则该用户的粉丝是否转发此消息就完全取决于用户之间的关系，即需要知道该用户的每个粉丝转发和评论该用户消息的概率。此概率计算需要较高的时间复杂度，在实验用户数量很少的情况下，可以通过详细计算用户之间的关系提高预测的准确度，但是对于微博突发话题传播，计算每 2 个用户之间转发和评论的概率显然不现实。基于此本文把用户 i 和其所有粉丝之间的关系度看作是相同的，则用户 i 的粉丝 j 在第 n 个窗口内被用户 i 相关微博消息影响的概率为

$$fc_{i,j}^{n}=\frac{for\_num_i+\sqrt{com\_num_i}}{follow\_num_i} \tag{2}$$

其中，for_num_i 是用户 i 的平均转发数，com_num_i 是用户 i 的平均评论数，$follow\_num_i$ 为用户 i 的粉丝数。

微博网络中很多用户往往在转发的同时对消息进行评论，或者在评论的同时转发，这样在计算 i 对粉丝的影响力时就不能简单地把转发数和评论数相加。用户转发话题相关微博后，肯定会对该话题的传播起到积极的作用，而用户评论也可能会导致用户发送相关微博，所以采用开根号来表示评论的影响。

式(2)是在不考虑突发话题的突发权值时用户对粉丝的影响力，对于突发话题的传播来说，这么计算显然是不准确的。微博中和突发话题相关消息的转发数和评论数往往会比一般微博消息的转发和评论多，用户发布的微博对粉丝的感染度值和相关话题的突发权值有很大的关系，同时还应该考虑突发话题的衰减等因素。经综合分析本文提出用户 i 对任意粉丝 j 的感染度值为

$$P_{i,j}^{n}=fc_{i,j}^{n}\times \text{lb}(w_{\text{topic}}^{n-1})\times(p_{\text{topic}})^{n-1} \tag{3}$$

其中，w_{topic}^{n-1} 为话题在第 n−1 个窗口内的突发权重。

3.2　基于用户的传播预测

用户 j 关注的所有感染用户都会给 j 一个感染度，这样假设用户 i 是用户 j 关注的用户，并且用户 i 在第 n−1 个窗口内变为感染用户，即 $Y_i(\text{window}_{n-1})=1\,\&\&\,Y_i(\text{window}_{n-2})=0$，则用户 j 在第 n 个跟踪窗口的感染度值为：

$$P_j^n=\sum_{i\in \text{following}(j)\&\&\,Y_i(\text{window}_{n-1})=1\&\&\,Y_i(\text{window}_{n-2})=0} P_{i,j}^{n}+P_j^{n-1}\times p_{\text{user}} \tag{4}$$

其中，P_j^{n-1} 是用户 j 在第 n−1 个跟踪窗口内的感染度值，p_{user} 表示用户感染度值衰减系数。用户 j 在第 2 个跟踪窗口的感染度值为

$$P_j^2=\sum_{i\in \text{following}(j)\&\&\,Y_i(\text{window}_1=1)} P_{i,j}^{2} \tag{5}$$

由式(4)和式(5)可以计算出下个跟踪窗口内每个易染用户的感染度值，用户 j 的感染度值可以简单地理解为用户 j 转发或发送话题相关微博的概率。需要注意的是，这里的感染度值只是微博网络内部用户对 j 的作用，即“内场强”。现实生活中，用户不仅受到微博网络的影响，微博网络外的影响作用同样巨大，甚至要高于微博对用户的影响。在微博突发话题传播预测的研究中无法准确的计算“外场强”的值，本文假设“内外场强”在每个窗口内都存在一个确定的比例。突发话题对易染用户的外场强 E 是由突发话题的属性和衰减系数 2 个因素共同决定的。本文定义在第 n 个窗口内跟踪话题的外场强 E_{topic}^{n} 为

$$E_{\text{topic}}^{n}=\frac{E_{\text{topic}}^{n-1}}{P_j^{n-1}+E_{\text{topic}}^{n-1}}\times\frac{num_{\text{topic}}^{n-1}}{window\_num^{n-1}}\times\frac{P_j^n}{P_j^{n-1}} \tag{6}$$

其中，num_{topic}^{n-1} 表示第 n−1 个窗口内讨论话题 topic

的微博数，$window\_num^{n-1}$ 是第 $n-1$ 个窗口的总微博数。

由式(4)和式(6)可知，用户 j 在第 n 个窗口内的总感染度值为

$$PE_j^n = P_j^n + E_{\text{topic}}^n \tag{7}$$

微博中易染用户有感染度值和活跃度值 2 个属性，现在已经得出易染用户的感染度值，下面给出用户活跃度值的计算公式为

$$RES_j^n = \frac{wb\_num_j}{wb\_time_j} \times window\_time^n \tag{8}$$

其中，wb_num_j 是用户 j 的微博总数，wb_time_j 表示用户从注册微博到当前时刻的时间，$window\_time^n$ 是第 n 个窗口的时间间隔。

式(8)含义是该用户发布微博的平均时间间隔。然后给出用户 j 会转发或发送突发话题相关微博的条件为，用户的感染度值和活跃度值乘积大于 1。

在某个窗口内当用户感染度值和其活跃度值之积大于 1 时，就认为该用户在本窗口内会被感染。然后和实际的实验结果进行对比，如果实验中该用户未被感染，而按本文的预测用户应该被感染，则把该用户归入免疫用户集，表示该用户不会再关注该话题。免疫用户的准确划分也是论文验证准确性的一个验证标准。基于用户的微博突发话题传播预测描述如算法 1。

算法 1 基于用户的话题传播预测

输入：网络用户图

输出：预测第 n 个窗口的感染用户集合

for each user $j \in$ 易染用户集

for each user $i \in$ 上一跟踪窗口的感染用户集

compute $P_{i,j}^n$

$P_j^n += P_{i,j}^n$

end for

$P_j^n += P_j^{n-1} \times p_{\text{user}}$

compute E_{topic}^n，PE_{j}^n，RES_j^n

if($PE_j^n + RES_j^n > 1$)

预测 j 在第 n 个窗口被感染

end for

4 基于节点规模的微博突发话题传播预测

微博中的用户数量庞大，其中大部分的用户均为易染用户，尤其在跟踪的前几个窗口，易染用户数可能数以百万计。对于算法 1 来说，遍历所有的易染用户显然会消耗太多的时间。上述所提基于用户的话题传播预测不适用于用户数量庞大的情况，在小规模数据中话题传播预测的准确度高。

基于用户的微博突发话题传播预测模型不适合于大规模微博网络，本节提出一种针对大规模用户群体的突发话题传播与预测模型——基于规模的传播预测，该模型不再去遍历数量庞大的易染用户集，而是通过计算感染用户对易染用户总的感染度值，估算下个窗口话题传播的整体规模。对于在第 $n-1$ 个窗口内的感染用户 i，由式(4)和式(5)可知，第 n 个窗口内用户 i 给其所有粉丝的总感染度值为

$$Pl_i^n = \sum_{j \in \text{follower}(j) \,\&\&\, Y_j(\text{window}_{n-1})=0} P_{i,j}^n \tag{9}$$

第 $n-1$ 个窗口内发布话题相关微博的所有用户都会产生一个感染度值，所有符合条件的用户总感染度值和前 $n-1$ 个窗口的感染度值之和，表示所有易染用户在第 n 个窗口内被感染的百分比，预测窗口内微博网络总的感染百分比计算公式为

$$P^n = \sum_{Y_i(\text{window}_{n-1})=1 \,\&\&\, Y_i(\text{window}_{n-2})=1} Pl_i^n + Pl^{n-1} \times p_{\text{user}} \tag{10}$$

外部场强还是由式(6)给出，假设现在网络中易染用户总数为 N，预测在第 n 个窗口内的感染用户数为：

$$aff\_num^n = Pl^n + N \times E_{\text{topic}}^n \tag{11}$$

基于规模的微博突发话题传播预测只需遍历感染用户集，在针对庞大数据集的话题传播预测中时间复杂度大大地降低。算法 1 可以相应地改为算法 2。

算法 2 基于规模的话题传播预测

输入：微博用户图

输出：预测第 n 个窗口的感染用户规模

for each user

$i \in Y_i(\text{window}_{n-2}) = 0 \;\&\&\; Y_i(\text{window}_{n-1}) = 1$

compute Pl_i^n

$P^n += Pl_i^n$

end for

set $N =$ 易感染用户数

compute aff_num^n

由算法 2 给出基于规模的突发话题传播预测第 n 个窗口内的传播用户人数为 aff_num^n。基于规模的预测模型很好地解决了大数据网络中话题的传

播预测问题，缺点是只能给出预测窗口中的预测规模，不能对具体的用户是否被感染进行分析。2 种方法各有优缺点，第一种方法是针对小数据网络详细的突发话题传播预测，可以分析每个易染用户的感染情况；第二种方法时间复杂度低，适用于实际的大规模网络，预测结果为下个窗口的传播规模。

5　实验过程与结果分析

5.1　实验目的

本实验的目的是验证基于微博的突发话题跟踪与突发话题传播预测的准确性。重点验证以下几点：本文所提基于特征字的突发话题跟踪技术是否能有效而准确地跟踪突发话题；基于用户的微博突发话题传播预测与基于规模的微博突发话题传播预测的准确性和时间复杂度。

5.2　实验环境

linux 32 位操作系统，1 G 内存，mysql 数据库系统。微博突发话题跟踪算法和微博突发话题传播预测均采用 *c* 语言编写。

5.3　实验数据

修改基于 Linux 的爬虫 Larbin 爬取的大量微博数据（包括用户的属性、微博消息的属性和用户之间的关系）。主要围绕以下 5 个突发事件进行抓取。

事件 1：北京时间 2012 年 5 月 28 日 10 点 22 分，河北省唐山市市辖区、滦县交界(北纬 39.7，东经 118.5)发生 4.8 级地震，震源深度 8 km。

事件 2：2012 年 06 月 30 日 5 点 07 分新疆维吾尔自治区伊犁哈萨克自治州新源县、巴音郭楞蒙古自治州和静县交界(北纬 43.4 度，东经 84.8 度)发生 6.6 级地震，震源深度 7 km。

事件 3：北京时间 2012 年 7 月 28 日 3 点 12 分举行伦敦奥运开幕式。

事件 4：北京时间 2012 年 8 月 13 日 4 点 12 分举行伦敦奥运闭幕式。

事件 5：2012 年 8 月 6 日 21 时 24 分，中国选手陈一冰以 15.80 的成绩获得亚军。

5.4　实验过程与结果分析

本实验主要对比清华大学的赵丽等人所提话题传播模型和本文话题传播模型在预测微博中话题传播时的准确度。通过预测结果和真实跟踪结果的对比，验证本文所提传播模型更符合微博话题的传播规律。实验的参数设置为：窗口大小为 2 000 条微博，话题感染度的衰减系数 p_{topic}=0.98，用户的感染度衰减系数 p_{user}=0.95，其他变量均可通过实验统计得到。其中预测误差是指预测结果和真实结果的差值，2 种模型预测结果如下。

事件 1：唐山地震

如图 6，图 7 所示，原有预测模型指赵丽等人所提的话题传播模型，其各个窗口的预测规模始终小于实际传播人数，原因是原有模型主要是针对于博客网络，博客网络中没有所谓的粉丝关系，假设用户 A 在第 n−1 个窗口内发布相关博文，那么在第 n 个窗口每个用户都可能通过 A 获知并发布类似博文。而在实际微博网络中，关注人数多的用户更容易看到相关的突发事件，故原有模型会降低那些关注数高的用户的被感染度，从而使预测结果偏小，遗漏本来可能的感染。

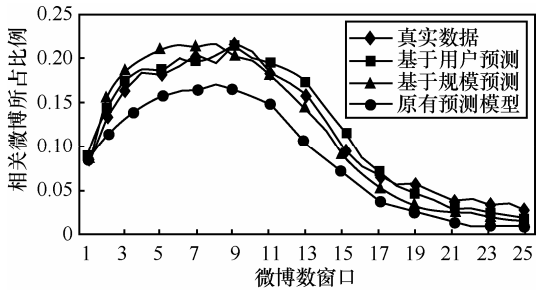

图 6　“唐山地震”预测结果对比

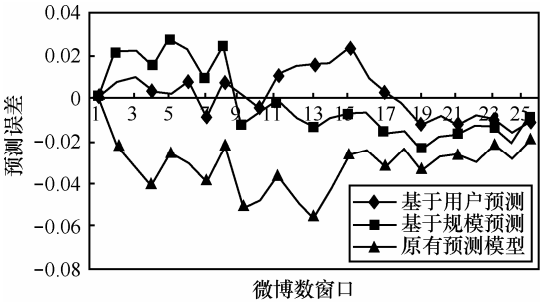

图 7　“唐山地震”预测误差对比

本文所提基于规模的预测模型，实验结果在第 8 个跟踪窗口前高于实际传播人数，第 11 个窗口后又低于实际传播人数，但是误差始终小于原有模型的误差。原因是基于规模的话题模型同样也将已染用户的感染力平均地分配给每个易染用户，和原有模型不同的是它考虑了微博中用户的影响力及用户的活跃度，使预测误差减小。

相比较而言，基于用户的预测模型，实验结果和实际传播情况非常接近，没有明显的过大或过小的预测，证明在充分考虑微博中的粉丝关系和用户

活跃度、影响力等因素后，预测的结果可以无限地趋向于实际情况。

事件 2：奥运开幕式

如图 8，图 9 所示，对于突发事件“奥运开幕式”的预测结果，和“唐山地震”的预测结果相似，验证了本文所提模型在准确度上比原有模型有所提升。但是本文所提模型和原有模型均存在一个严重的问题，如以下实验所示。

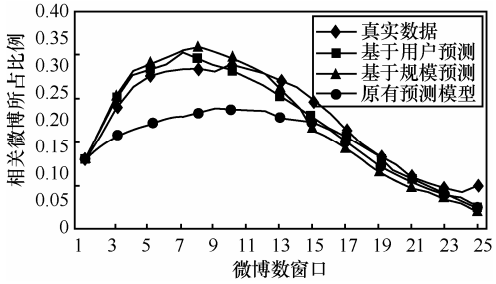

图 8 “奥运开幕式”预测结果对比

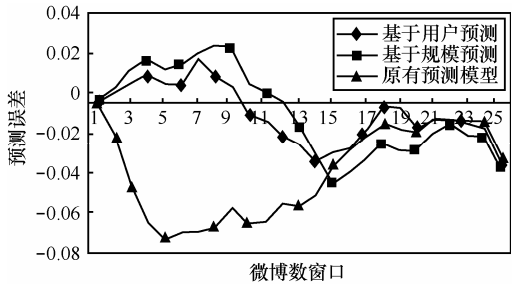

图 9 “奥运开幕式”预测误差对比

事件 3：陈一冰获亚军

如图 10，图 11 所示实验结果，在预测实验的前期，结果和上 2 次实验类似。然而，在第 13 个窗口以后，本文预测模型和原有模型均出现很大的误差，预测规模远远低于实际传播人数。可能的原因是现有的基于衰减系数的衰减模型只适合于一般的突发事件，对于某些持续时间长的突发事件会有较大的预测误差。但是本文所提预测模型的误差还是小于原有模型，原因是本文考虑了突发事件的规模，一般情况下，突发的规模越大，持续的时间越长。

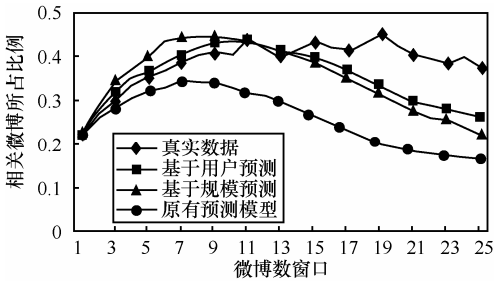

图 10 “陈一冰亚军”预测结果对比

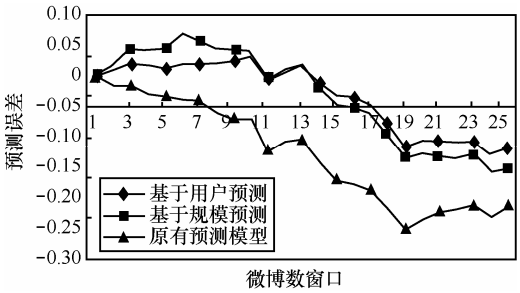

图 11 “陈一冰亚军”预测误差对比

6 结束语

通过研究病毒传染模型、微博中的粉丝关系和用户活跃度、影响力属性，本章提出了 2 种基于微博的突发话题预测技术——基于用户的突发话题传播预测和基于规模的突发话题传播预测。将微博用户划分为 3 个集合——感染用户、易染用户和免疫用户。基于用户的话题传播预测需要遍历所有的易染用户，统计每 2 个用户之间的关注关系，可以预测具体用户在下个窗口的感染概率，预测结果准确；基于规模的话题传播预测只需要遍历感染用户集，适合大规模网络突发话题的传播预测，预测结果有一定偏差。

通过实验验证了本文所提突发话题传播预测模型的准确性。基于用户的话题传播预测准确度高，基于规模的话题传播预测时间复杂度低，和原有模型相比准确度有所提升。通过实验发现现有预测模型中的衰减模型存在一定的问题，预测大规模突发时有一定的预测误差，有待后续改进。

参考文献：

[1] KUO T T , HUNG S C , LIN W S. Exploiting latent information to predict diffusions of novel topics on social networks[A]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics[C]. Stroudsburg, PA, USA, 2012. 344-348.

[2] ROMERO M.D, MEEDER B, KLEINBERG J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter[A]. Proceedings of the 20th International Conference on World Wide Web[C]. New York, NY, USA, 2011.695-704.

[3] 韩兰胜. 计算机病毒的传播模型及其求源问题研究[D]. 武汉：华中科技大学. 2006.

HAN L S. Research on Propagation Model and Source Seeking of Computer Virus[D]. Wuhan:Huazhong University of Science and Technology, 2006.

[4] 赵丽, 袁睿翕, 管晓宏. 博客网络中具有突发性的话题传播模型[J]. Journal of Software, 2009, 20(5):1384-1392.
ZHAO L, YUAN R X, GUAN X H. Bursty propagation model for incidental events in blog networks[J]. Journal of Software, 2009, 20(5):1384-1392.

[5] 孙留东. 博客网络中突发话题传播模型及网络特性研究[D]. 重庆: 重庆邮电大学, 2011.
SUN L D. Research on Evolution Emergment Events and Network Property in Blog[D]. Chongqing: Chongqing University of Posts and Telecommunications, 2011.

[6] HE, D, PARKER D S. Topic dynamics: an alternative model of bursts in streams of topics[A]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York, NY, USA, 2010.443-452.

[7] 孙胜平: 中文微博客热点话题检测与跟踪技术研究[D]. 北京: 北京交通大学, 2011.
SUN S P. Research on Chinese Micro-Blog Hot Topic Detection and Tracking[D]. Beijing: Beijing Jiaotong University, 2011.

作者简介：

王巍（1974-），男，山东夏津人，博士，哈尔滨工程大学副教授，主要研究方向为网络与信息安全、数据挖掘。

李锐光（1979-），男，山西阳泉人，硕士，国家计算机应急技术处理协调中心工程师，主要研究方向为舆情分析、移动互联网安全。

周渊（1972-），男，江苏无锡人，博士，国家计算机网络与信息安全管理中心教授级高级工程师，主要研究方向为信息安全。

杨武[通信作者]（1974-），男，辽宁宽甸人，博士，哈尔滨工程大学教授，主要研究方向为网络与信息安全、数据挖掘。E-mail: yangwu@hrbeu.edu.cn。