计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190584
57(6): 1312-1322, 2020

# NT-EP:一种无拓扑结构的社交消息传播范围预测方法

刘子图 全紫薇 毛如柏 刘 勇 朱敬华
(黑龙江大学计算机科学技术学院 哈尔滨 150080)
(Vimotus_liu@163.com)

# NT-EP: A Non-Topology Method for Predicting the Scope of Social Message Propogation

Liu Zitu, Quan Ziwei, Mao Rubai, Liu Yong, and Zhu Jinghua
(*College of Computer Science and Technology, Heilongjiang University, Harbin* 150080)

**Abstract** Predicting the scope of a message accurately in social networks is an important part of public opinion analysis, which has received extensive attention in the field of data mining. Most of the current research mainly uses social network topology and user action logs to predict the spread of social messages. It is usually easy to obtain action log about users in real applications, but the topology of the social network (for example, the friend relationship between users) is not easy to obtain. Therefore, non-topology social message prediction has good prospects for broader applications. In this paper, we propose a new method called NT-EP for predicting the propagation scope of social messages. NT-EP consists of four parts: 1) We construct a weighted propagation graph for each message based on the characteristics of message propagation decay over time, and then use a random walk strategy to obtain multiple propagation paths on the propagation graph; 2) We put multiple propagation paths of the target message into Bi-GRU, and combine the attention mechanism to obtain the propagation feature representation for the target message; 3) We use the gradient descent method to calculate the influence representation about other messages; 4) Combining the propagation feature representation for the target message with the influence representation about other events, we predict the propagation scope of the target message. The experimental results on Sina microblog and Flixster dataset show that our method is superior to existing social event prediction methods in terms of many indicators such as MSE and *F*1-*score*.

**Key words** social network; scope of propagation; topology structure; random walk; gradient descent

**摘 要** 准确预测社交网络中消息的传播范围是舆情分析的重要内容,该问题受到了数据挖掘领域的广泛关注.目前的大部分研究主要利用社交网络拓扑结构和用户的动作日志来预测社交消息的传播范围.在实际应用中用户的动作日志中通常容易获得,但是社交网络的拓扑结构(例如用户之间的朋友关系)并不容易获得,因此无拓扑结构的社交消息预测具有更广泛的应用前景.提出了一种新的社交消息传播

**收稿日期**:2019-08-23;**修回日期**:2019-12-09

**基金项目**:国家自然科学基金项目(61972135,61602159);黑龙江省自然科学基金项目(F201430);哈尔滨市科技局创新人才项目(2017RAQXJ094,2017RAQXJ131);黑龙江省属高等学校基本科研业务费基础研究项目(HDJCCX-201608,KJCX201815,KJCX201816)

This work was supported by the National Natural Science Foundation of China (61972135, 61602159), the Natural Science Foundation of Heilongjiang Province of China (F201430), the Innovation Talents Project of Science and Technology Bureau of Harbin (2017RAQXJ094, 2017RAQXJ131), and the Fundamental Research Funds of Universities in Heilongjiang Province (HDJCCX-201608, KJCX201815, KJCX201816).

**通信作者**:刘勇(liuyong123456@hlju.edu.cn)

范围预测方法 NT-EP,该方法由 4 部分构成:1)利用消息传播随时间衰减的特性为消息构造加权传播图,使用随机游走策略获取多条传播路径;2)把目标消息的传播路径输入到 Bi-GRU(bidirectional gated recurrent unite),结合注意力机制计算出目标消息的传播特征向量;3)使用梯度下降方法计算出其他消息的影响向量;4)将目标消息的传播特征向量和其他消息的影响向量结合在一起,预测目标消息的传播范围.在 Sina 微博和 Flixster 数据集上的实验结果表明:NT-EP 方法在均方误差(mean squared error, MSE),*F1-score* 等多个指标上都优于现有的社交消息预测方法.

**关键词** 社交网络;传播范围;拓扑结构;随机游走;梯度下降

**中图法分类号** TP391

近年来随着社交网络的快速发展,越来越多的用户使用新浪微博、Twitter、Facebook 等社交网站分享自己的生活.据统计 Facebook 截至 2018 年 12 月 31 日每月的活跃用户超过 23 亿[1].由此可见社交网已经成为许多人生活的一部分.与此同时各大社交平台也在促进着各种消息的快速传播.例如在新浪微博上平均每天有几亿条微博产生,在每天产生的微博中会包含很多重要信息.用户更新一条微博可能包含着用户对某消息的态度和观点[2],也可能是分享身边的新鲜事[3].预测消息的传播范围在病毒营销、舆情监控、商品推荐等诸多领域都有广泛的应用,因此受到了数据挖掘领域的广泛关注.

目前对消息传播范围进行预测所使用的方法主要有 2 种:1)根据消息特征或者消息的特定类型进行传播范围预测.例如:可以根据发布的 Twitter 是否带有标志性的图片从而预测它在 Facebook 上的传播范围[4];也可以通过分析发布的 Twitter 是否包含对消息传播有利的内容来预测它的传播范围[5].然而使用消息特征预测消息传播范围显然不能推广到不同的平台.2)使用社交网络中用户的拓扑结构[5-7]或消息的转发结构[8]来预测消息传播范围.然而在很多实际应用中,我们很难获得消息的传播结构以及用户的拓扑结构,通常只能获得消息的传播序列.例如在豆瓣网中,对于电影的影评只显示用户在什么时间评价了电影,而没有表明用户因为受到哪些用户影响才评价该电影.因此,只利用消息的传播序列而不考虑用户的拓扑结构来预测消息的传播范围具有更广泛的应用场景.

本文研究了无拓扑结构条件下的消息传播范围预测问题,提出了一种无拓扑结构的消息传播范围预测方法 NT-EP.该方法由 4 部分构成:1)利用消息传播随时间衰减的特性为每个消息构造一个加权的传播图,在传播图上使用随机游走策略获取多条传播路径,再使用 Word2vec 方法计算每个用户的特征向量;2)把目标消息的传播路径替换成用户的特征向量序列输入到双向门控制循环神经网络(bidirectional gated recurrent unite, Bi-GRU),结合注意力机制计算出目标消息的传播特征向量;3)考虑到不同消息传播可能存在的相互影响,利用目标消息发生前的其他消息,使用梯度下降方法计算出其他消息的影响向量;4)将目标消息的传播特征向量和其他消息的影响向量结合在一起,使用多层感知机(multilayer perceptron, MLP)拟合出目标消息的传播范围.与其他方法相比,NT-EP 方法具有 2 个明显的创新:1)首次考虑了消息之间的相互影响;2)利用消息传播随时间衰减特性为每条消息构造加权传播图,抽取传播路径.

NT-EP 方法充分考虑了消息之间的相互影响.这是因为在消息的传播过程中,消息与消息之间必然会产生影响.例如在公布个人所得税起征点改革消息之后的一段时间内,用户会增加对包含具体税率改革内容的消息的关注.因此个人所得税起征点改革消息对有相关内容的消息传播产生了影响.消息传播中的相互影响来自于 2 方面:1)来源于消息本身的内容,也就是消息本身是否为热点消息,是否会被普遍关注;2)来源于已经参与消息传播的用户对于其他消息传播的影响.在一段时间内,用户使用社交网络的时间上限是固定的.用户浏览某些消息的时间更多意味着用户浏览其他消息的时间会减少.因此本文方法 NT-EP 考虑了目标消息发生前后其他消息对目标消息可能存在的各种影响,构造了其他消息的影响向量,结合目标消息的传播特征向量来预测目标消息的传播范围.本文实验部分比较了利用消息影响与不利用消息影响 NT-EP 方法的 2 种变体,证明了消息影响对范围预测的重要性.

NT-EP 方法根据传播序列构造加权传播图,来模拟接近真实传播轨迹的传播路径.在无拓扑结构的条件下,我们只有用户的动作序列.但是用户在接

受消息过程中必然会受到之前接受相同消息用户的影响,而且影响强度依赖于接受消息的时间差.假设在消息 $a$ 传播过程中用户 $V_1$ 接受了消息 $a$,在用户 $V_1$ 之前用户 $V_2$ 和用户 $V_3$ 也接受消息 $a$,并且用户 $V_2$ 接受的时间早于用户 $V_3$,那么直觉上用户 $V_3$ 对用户 $V_1$ 接受消息影响更大.根据消息传播随消息衰减的特性,我们构造有向边 $V_2 \rightarrow V_1$ 和 $V_3 \rightarrow V_1$,$V_3 \rightarrow V_1$ 边上的权值大于 $V_2 \rightarrow V_1$,边上的权值代表了影响强度,权值依赖于 2 个用户接受消息的时间差.NT-EP 方法按照这种方式为每个消息构造一个随时间衰减的传播图,然后使用随机游走策略抽取多条传播路径,这些传播路径更接近于真实的传播轨迹.本文实验部分构造了 NT-EP 方法的 2 种变体,一种利用时间衰减构造传播图,另一种不利用时间衰减构造传播图,比较了这 2 种变体的性能,再次证明了消息传播符合时间衰减特性.本文的贡献有 4 个方面:

1) 提出了一种新的无拓扑结构条件下的消息传播范围预测方法 NT-EP.

2) NT-EP 利用了消息之间的相互影响,提高了消息传播范围预测的准确性.

3) NT-EP 利用了消息传播随时间衰减的特性为消息构造加权传播图,使得抽取的随机游走路径更接近真实的传播轨迹,提高了消息传播范围预测的准确性.

4) 实验结果表明,NT-EP 能对无拓扑结构条件下的消息传播范围进行准确预测,并且预测效果明显优于现有的方法.

本文源码和数据可以从 https://github.com/Vimotus/NT-EP 下载.

## 1 相关工作

在社交网中消息的传播范围包括微博或 Twitter 在一定时间内的转发数[4-5,9-11]、照片的被浏览数[2]、视频被点赞的次数[12-14]、学术论文在一定时间内被引用的次数[15]等多种情况.相关工作大致可以分为 3 类:1)利用消息本身的特征进行预测;2)利用消息的传播序列和用户的社交关系进行预测;3)只利用消息的传播序列进行预测.

消息特征或者消息的特定类型可以帮助预测消息的传播范围.例如文献[4]根据发布的 Twitter 是否带有标志性的图片来预测它在 Facebook 上的转发次数;文献[12]分析视频在规定的时间段内观看人数的增长量来预测消息被观看的次数.然而消息的传播范围除了消息本身的特征,更多依赖发布消息或者转发消息用户的影响力,因此此类方法预测效果一般,而且不易推广到其他平台.

目前的绝大多数研究都是利用消息的传播序列和用户的社交关系进行预测.该类方法又可分为 2 种:1)将消息传播预测视为分类问题,通过预测传播范围是否会超过某个阈值来预测某个消息是否会变成流行消息[7,12,14];2)将消息传播范围预测看作回归问题,预测消息的最终传播范围或者截止到某一时刻的传播范围.此类研究通常使用确定的时间属性[12]、早期消息传播的拓扑结构[6,15-16]以及用户的拓扑结构[17],来进行传播范围预测.文献[18]学习多数消息传播过程中的普遍拓扑结构预测消息传播范围.此类方法需要消息的传播结构或者用户的拓扑结构,但实际应用中这些信息不易获得.

目前在无拓扑结构(用户社交关系)条件下,对消息传播范围预测的研究相对较少.2010 年 Gomez-Rodriguez 等人[19]利用用户被影响的时间特征推断消息传播的路径,然后累加路径上的用户数计算传播范围.2012 年 Simma 等人[20]提出了基于连续时间和霍克斯进程的随机过程范围预测模型.2014 年 Bourigault 等人[21]提出了基于学习映射观察动态时间对连续空间的影响,将参与扩散的节点投射到潜在的表达空间,然后计算用户向量的相似性判断用户是否会被另一个用户影响.2016 年 Bourigault 等人[22]使用用户表达空间,将用户的影响能力映射到一个多维空间中,通过计算多维空间中 2 个向量的距离来计算是否会被影响.

现有方法并没有考虑消息在传播过程中会存在相互影响的情况.本文利用了消息之间的相互影响,提出了一种无拓扑结构的传播范围预测方法 NT-EP,该方法具有 5 个优势:1)是一种端对端的学习框架;2)适用于无拓扑结构;3)考虑了消息传播过程中的相互作用;4)抽取的随机游走路径更接近真实的传播路径;5)结合目标消息的传播向量和其他消息的影响向量,同时利用注意力机制预测传播范围,使预测结果更准确.

## 2 问题定义

在无拓扑结构的消息传播预测中,我们没有社交网络的拓扑结构,只有用户的动作日志 $L=\{(u,$

$a,t)|u\in V,a\in A,t\in T\}$,其中 $V$ 为社交网中用户集合,$A$ 为社交网中发生的消息集合,$T$ 为用户参与消息传播的时间区间,$(u,a,t)$ 表示用户 $u$ 在时间 $t$ 参与了消息 $a$(例如用户 $u$ 对消息 $a$ 转发、点赞等).对于给定的目标消息 $a$,我们有消息 $a$ 从发生时刻 $t_1$ 到当前时刻 $t_2$ 的动作日志 $L_a^{t_c}=\{(u,a,t)|u\in V,t_1\leqslant t\leqslant t_2\}$.在一段时间 $\Delta t$ 之后,我们还可以搜集到消息 $i_a$ 截止到时间 $t_2+\Delta t$ 的动作日志 $L_a^{t_2+\Delta t}=\{(u,i_a,t)|u\in V,t_1\leqslant t\leqslant t_2+\Delta t\}$.在舆情监控中,对突发消息提前预警可以有助于政府和主管部门提前采取干预措施.因此本文研究问题定义为:对刚发生不久的消息 $a$,根据消息 $a$ 的当前动作日志 $L_a^{t_2}$,预测消息 $a$ 在未来一段时间 $\Delta t$ 之后的增量传播范围,即 $|L_a^{t_2+\Delta t}|-|L_a^{t_2}|$.

## 3 NT-EP 传播范围预测框架

对于无拓扑结构下社交消息传播预测问题,本文提出了一种新的社交消息传播范围预测方法 NT-EP,其框架如图 1 所示.方法 NT-EP 首先根据消息动作日志中的传播时间差为每个消息构造一个加权的传播图,如图 1 的①②所示.传播图边上的数字代表用户之间的影响概率.在传播图构造完成之后,使用随机游走方式从传播图中提取若干条该消息可能的传播路径,如图 1 的③所示,然后使用 Word2vec 方法计算出每个用户的初始的特征向量.传播路径上的每个用户获得初始的特征向量后,再将消息的传播路径送入 Bi-GRU 中得到用户的最终向量表示,如图 1 的④所示.传播路径上的每个用户获得最终的特征向量后,再结合注意力机制计算出每个消息的传播特征向量,如图 1 的⑩所示.

在消息传播过程中,不同消息之间会存在相互影响.因此我们也必须计算目标消息发生前其他消息的可能影响.如图 1 的⑤~⑧所示,使用和目标消息类似的方式,构造加权传播图、随机游走、Word2vec 等方法计算其他消息参与用户的特征向量,然后构造其他消息的传播向量.

此后,使用梯度下降方法计算出其他消息的影响向量,如图 1 的⑨所示.最后 NT-EP 方法将目标消息的传播特征向量和其他消息的影响向量结合在一起,使用 MLP 拟合出目标消息的增量传播范围,如图 1 的⑫⑬所示.

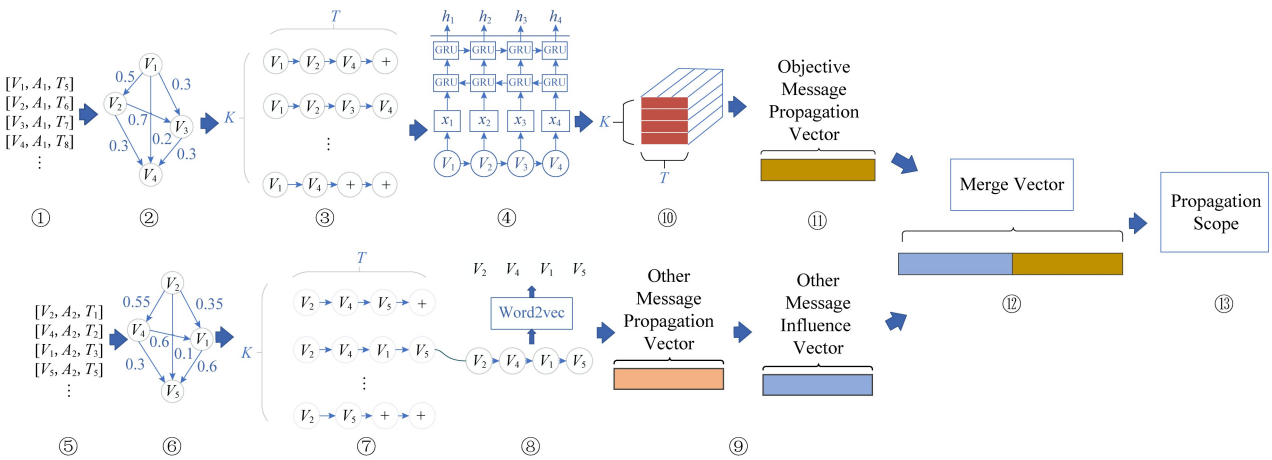


Fig. 1 NT-EP method framework

图 1 NT-EP 方法的框架

### 3.1 传播路径选取

给定的动作日志通常对每个消息的动作按照传播时间排序,如图 2 的①所示.用户 $V_1$ 在时间 1 接受了消息 $A_1$,用户 $V_2$ 在时间 2 接受了消息 $A_1$,….从给定的动作日志中我们无法获得消息真实的传播轨迹.因为真实的情况可能是:用户 $V_3$ 在时间 3 接受了消息 $A_1$ 可能是因为用户 $V_3$ 和用户 $V_1$ 是朋友,并且用户 $V_3$ 看到了用户 $V_1$ 接受了消息 $A_1$,从而影响用户 $V_3$ 也接受了消息 $A_1$.用户 $V_3$ 不认识用户 $V_2$,$V_3$ 接受了消息 $A_1$ 从来没有受到用户 $V_2$ 的影响.这样的真实传播轨迹在没有用户社交关系的条件下是无法获得的.

根据社交网上消息传播呈指数衰减特性[13],我们有理由认为当用户 $V_3$ 接受消息 $A_1$ 的时候,用户 $V_2$ 影响的概率大于用户 $V_1$ 影响的概率,因为 $V_2$ 接受消息 $A_1$ 的时间离 $V_3$ 接受消息 $A_1$ 的时间更近.

因此,我们根据2个用户接受消息的时间差来刻画2个用户的影响.假设用户 $V_i$ 和 $V_j$ 接受消息的时间分别为 $t_i$ 和 $t_j$,并且 $t_i<t_j$,则用户 $V_i$ 对 $V_j$ 的影响定义为

$$w(t_i,t_j)=e^{-\mu(t_j-t_i)}, \tag{1}$$

其中,$\mu$ 为调整时间差影响的超参数,实验中给出了该参数的选择过程.计算出每个消息中用户之间的影响概率后,我们可以根据影响概率为每个消息构造一个加权图来模拟该消息的真实传播轨迹.如图2的②所示,如果在消息 $A_1$ 中用户 $V_1$ 接受消息 $A_1$ 的时间早于用户 $V_2$ 接受消息 $A_1$ 的时间,则从用户 $V_1$ 引出一条边指向用户 $V_2$,边上的权值表示用户 $V_1$ 对用户 $V_2$ 的影响概率,由式(1)计算.在得到加权图后,我们再对加权图归一化,使得每个节点出边的概率和为1,如图2的③所示.为了模拟真实的传播路径,我们在归一化的加权图上根据边上的概率进行随机游走.每次游走的开始节点都是接受消息的第1个用户,例如图2的③中的 $V_1$.针对每个消息,我们采样 $K$ 条路径,并且每条路径的长度为 $T$.当游走过程中遇到某条路径长度小于 $T$ 的时候,在后面若干补充位,让每条路径长度都等于 $T$.在抽取了所有消息的传播路径后,我们使用词向量方法Word2vec计算每个用户的初始特征向量,细节见3.3节.

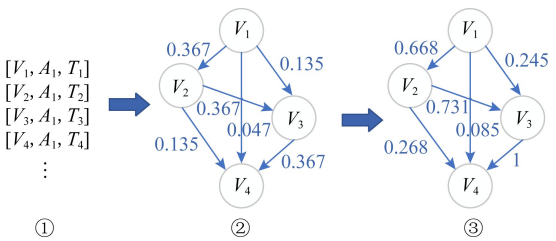


Fig. 2 Constructing a weighted graph of message propagation

图2 构造消息传播的加权图

### 3.2 其他消息影响向量

社交网中消息与消息之间存在着不同程度的联系,一个消息的传播可能促进或者抑制另一个消息的传播.例如国家个人所得税更改方案公布时,短时间内对税率信息查询有促进传播的作用.此外,用户上网浏览消息的时间有限,对于某些消息的关注增加,对其他消息的关注就会降低.

下面介绍其他消息对目标消息的影响能力.该影响能力通过一个影响向量来刻画.设当前的目标消息为 $a$,$a$ 发生时间为 $t_a$.如果消息 $A_1$ 在消息 $a$ 之前发生,并且消息 $A_1$ 的发生时间与消息 $a$ 的发生时间距离较近,那么消息 $A_1$ 的传播很可能会对消息 $a$ 的传播产生影响.基于这一思想,我们获取在 $t_a$ 之前很短的时间段 $\tau$ 内发生的消息集合 $S_a=\{A_1,A_2,\cdots,A_m\}$,该集合内每个消息对消息 $a$ 传播的影响都需要考虑.假设影响消息集合 $S_a=\{A_1,A_2,\cdots,A_m\}$ 中每个消息截至时刻 $t_a$ 的传播范围分别为 $n_1,n_2,\cdots,n_m$,发生时间分别为 $t_1,t_2,\cdots,t_m$,并且满足 $t_1<t_2<\cdots<t_m<t_a$,$t_a-t_1<\tau$.直觉上,对影响消息集合 $S_a=\{A_1,A_2,\cdots,A_m\}$ 中的某个消息 $A_1$ 来说,消息 $A_1$ 的影响能力与传播能力共同作用决定了消息 $A_1$ 在 $t_a$ 时的影响范围 $n_i$.消息 $A_1$ 的影响范围 $n_i$ 越大,消息 $A_1$ 的影响能力越强.这是因为在某段时间内若消息 $A_1$ 成为热点消息,在短时间内有巨大的浏览量,用户对当前目标消息的浏览量会有所减少.根据这一思想,我们使用 $d$ 维向量 $\boldsymbol{p}_i$ 表示消息 $A_i$ 的传播能力,使用 $d$ 维向量 $\boldsymbol{q}_i$ 表示消息 $e_i$ 的影响能力,构造目标函数:

$$\arg\min_{\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_m\}}\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{p}_i\cdot\boldsymbol{q}_i-\mathrm{lb}\,n_i)^2. \tag{2}$$

消息 $A_i$ 的传播能力 $\boldsymbol{p}_i$ 来自于参与消息 $A_i$ 的用户传播能力,因此消息 $A_i$ 的传播能力 $\boldsymbol{p}_i$ 可以表示为

$$\boldsymbol{p}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j, \tag{3}$$

其中,$\boldsymbol{p}_i$ 表示影响消息 $A_i$ 的传播向量,$\boldsymbol{x}_j$ 是使用Word2vec的skip-gram模型从传播路径上得到的用户向量,因为消息 $e_i$ 的传播范围为 $n_i$,所以有 $n_i$ 个不同的用户向量 $\boldsymbol{x}_j$.对于消息 $e_i$ 的影响向量 $\boldsymbol{q}_i$,本文使用梯度下降算法求解,使式(2)的目标函数最小化,具体算法如算法1所示.在得到影响消息集合 $S_a=\{A_1,A_2,\cdots,A_m\}$ 中每个消息的影响向量 $\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_k\}$ 后,我们计算整个消息集合 $S_a$ 对当前目标消息 $a$ 的影响向量 $\boldsymbol{q}_{S_a}$.

$$\boldsymbol{q}_{S_a}=\sum_{i=1}^{k}\frac{1}{t_a-t_i}\boldsymbol{q}_i. \tag{4}$$

**算法1.** 计算其他消息影响向量.

输入:其他消息的传播向量 $\boldsymbol{p}_i(i=1,2,\cdots,m)$、其他消息的当前传播范围 $n_i(i=1,2,\cdots,m)$、学习率 $\lambda$;

输出:其他消息的影响向量 $\boldsymbol{q}_i(i=1,2,\cdots,m)$.

① Init $\boldsymbol{q}_i(i=1,2,\cdots,m)$;

② repeat

③ $\Delta\boldsymbol{g}\leftarrow\frac{2}{m}\sum_{i=1}^{m}\boldsymbol{p}_i(\boldsymbol{p}_i\cdot\boldsymbol{q}_i-\mathrm{lb}\,n_i)$;

④ $\boldsymbol{q}_i \leftarrow \boldsymbol{q}_i - \lambda \Delta \boldsymbol{g}$;

⑤ until convergence.

### 3.3 目标消息传播向量

在抽取了所有消息的传播路径后,我们将每条传播路径当作一个句子,路径上的每个用户当作句子中的单词,输入到 Word2vec[16] 的 skip-gram 模型中,得到每个用户初始的特征向量.假设用户特征向量的维度为 $H$.

因为循环神经网络适合处理时序数据,我们采用门控循环单元(GRU)来捕捉目标消息的传播过程.对于每条传播路径,我们从前向后处理路径上的每个用户,第 $j$ 个用户的输入为用户 $V_j$ 的初始特征向量 $\boldsymbol{x}_j \in \mathbb{R}^H$,GRU 进行计算后更新隐藏状态 $\boldsymbol{h}_i = h_{\mathrm{GRU}}(\boldsymbol{x}_j, \boldsymbol{h}_{i-1})$,其中 $\boldsymbol{h}_{i-1} \in \mathbb{R}^H$ 表示 GRU 更新前的隐藏状态,$\boldsymbol{h}_i \in \mathbb{R}^H$ 表示 GRU 更新后的隐藏状态.

为了知道消息在传播过程中会被哪些用户影响,我们使用相同的方法再从后向前处理路径上的每个用户.因此本文使用的是双向 GRU(Bi-GRU),拼接隐藏状态的输出得到对应用户的最终向量表示.具体通过式(5)对用户向量进行更新:

$$\begin{aligned}
&\boldsymbol{r}_i = \sigma(\boldsymbol{W}^r \boldsymbol{x}_j + \boldsymbol{U}^r \boldsymbol{h}_{i-1} + \boldsymbol{b}^r),\\
&\boldsymbol{u}_i = \sigma(\boldsymbol{W}^u \boldsymbol{x}_j + \boldsymbol{U}^u \boldsymbol{h}_{i-1} + \boldsymbol{b}^u),\\
&\bar{\boldsymbol{h}}_i = \tanh(\boldsymbol{W}^h \boldsymbol{x}_j + \boldsymbol{r}_i \boldsymbol{U}^h \boldsymbol{h}_{i-1} + \boldsymbol{b}^h),\\
&\boldsymbol{h}_i = \boldsymbol{u}_i \bar{\boldsymbol{h}}_{i-1} + (\boldsymbol{1} - \boldsymbol{u}_i)\boldsymbol{h}_{i-1},\\
&\boldsymbol{y}_i = \sigma(\boldsymbol{W} \cdot \boldsymbol{h}_i),
\end{aligned} \tag{5}$$

其中 $j$ 表示传播路径中第 $j$ 个节点,模型输入的用户节点向量 $\boldsymbol{x}_j$ 和隐藏状态 $\boldsymbol{h}_{i-1}$ 一起作为输入,并通过 GRU 的公式计算更新.其中 $\boldsymbol{W}$ 和 $\boldsymbol{U}$ 作为训练期间学习的 GRU 参数.

如图 1 的④所示,每条传播路径经过 GRU 处理后,会得到该路径上每个用户的最终向量表示 $\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots$;然后我们使用注意力机制合并这些用户的最终向量表示 $\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots$,计算该传播路径的向量表示;最后对所有传播路径的向量表示累加求和,得到目标消息传播向量 $\boldsymbol{p}_a$:

$$\boldsymbol{p}_a = \sum_{k=1}^{K} \sum_{i=1}^{T} \eta_i \boldsymbol{h}_{ki}, \tag{6}$$

其中,$K$ 代表消息 $a$ 上抽取的传播路径数,$T$ 表示每条传播路径的长度,$\boldsymbol{h}_{ki}$ 是通过 GRU 得到的第 $k$ 条路径上第 $i$ 个用户的最终用户向量.为了区分同一条路径上不同节点对消息传播向量的作用,我们使用注意力机制,设 $T$ 个节点的权重分别为 $\eta_1, \eta_2, \cdots, \eta_T$,并且满足 $\sum_{i=1}^{T} \eta_i = 1$. 参数 $\eta_1, \eta_2, \cdots, \eta_T$ 通过端对端的学习获得.

### 3.4 传播范围预测

目标消息 $a$ 的传播范围依赖于目标消息 $a$ 的传播向量 $\boldsymbol{p}_a$ 和其他消息的影响向量 $\boldsymbol{q}_{S_a}$,因此我们将目标消息 $a$ 的传播向量 $\boldsymbol{p}_a$ 和其他消息的影响向量 $\boldsymbol{q}_{S_a}$ 融合为一条向量 $\boldsymbol{l}_a$,即:

$$\boldsymbol{l}_a = \boldsymbol{p}_a + \alpha \boldsymbol{q}_{S_a}, \tag{7}$$

其中,$\alpha$ 为其他消息影响向量的权重,实验部分给出了参数 $\alpha$ 的选择过程.将融合后的向量 $\boldsymbol{l}_a$ 作为一个多层感知机的输入,输出预测的传播范围 $f(a) = f_{\mathrm{MLP}}(\boldsymbol{l}_a)$,其中 MLP 为一个多层感知机,$f(a)$ 为消息 $a$ 的预测传播范围.

## 4 实验结果及分析

### 4.1 实验数据

本文中我们使用 2 套无拓扑结构的传播数据进行实验并对结果评估.数据描述如表 1 所示:

**Table1 Dataset Statistics**

**表 1 实验数据描述**

| Dataset | # User | # Message | # Record |
|---|---|---|---|
| Weibo | 261 839 | 1 280 | 933 683 |
| Flixster | 109 816 | 1 000 | 581 202 |

1) 微博[18].微博是基于用户关系的信息分享、传播的社交媒体.我们从论文中提供的数据中选取在 2012-09-28—2012-10-29 之间发生的 1 280 个消息的动作日志.截取的数据中包含 261 839 个用户、1 280 个消息以及 933 683 条传播记录.

2) Flixster[11].Flixster 是一个电影社交网站,可以让用户分享电影的评分,讨论新的电影.我们使用 1 000 个消息的动作日志.其中包含 109 816 个用户和 581 202 条传播记录.

实验中预测消息传播范围时通过调整时间长度 $t$ 和 $\Delta t$ 来选择预测的时间区间.$t$ 表示消息从发生开始到当前时刻所经过的时间,也就是消息已经发生了多久;$\Delta t$ 表示在时间 $t$ 之后的时间长度.实验中我们选择 $t$ 与 $\Delta t$ 的大小分别为 12 h,24 h,36 h 来对消息的传播范围进行预测.实验中需要进行其他消息影响向量和用户特征向量占用空间存储.实验中将数据按 7∶1∶2 的比例分为训练集、验证集、测试集.数据集中每个消息的全部动作日志只出现在训练集、测试集、验证集中的一个.我们在训练集中训练模型,在验证集中调整超参数,在测试集中测试方法的性能.

### 4.2 评估指标

本文使用均方误差(mean squared error, MSE)来评估传播范围预测效果.这是回归任务中常见的评估指标.它是由预测值与真实值差的平方和求平均得到,定义为

$$E_{\mathrm{MSE}}=\frac{1}{n}\sum_{i=1}^{n}(m_i-\hat{m}_i)^2, \tag{8}$$

其中,$m_i$ 为实际传播用户数,$\hat{m}_i$ 为预测值,Deepcas[7]中为了避免误差的数值过大导致 MSE 值过大,将 $m_i$ 取对数后再计算均方误差的大小,即 $m_i=\mathrm{lb}(\Delta n_i+1)$,其中 $\Delta n_i$ 为实际传播用户数.本文采用与 Deepcas 相同的处理方式.

本文使用精确率、召回率、$F1\_score$ 来评估消息热点预测效果.在热点消息预测时,我们只进行采样 12 h 的传播并预测消息的最终传播范围.实验中设置一个阈值,超过阈值会被认为是热点消息,实验中选择的阈值为 1 000.具体如下:

1) $TP$(真正例).$TP$ 表示预测传播范围大于阈值,并且实际传播范围大于阈值的消息数.

2) $FN$(假负例).$FN$ 表示预测传播范围大于阈值,但实际传播范围小于阈值的消息数.

3) $FP$(假正例).$FP$ 表示预测传播范围小于阈值,但实际传播范围大于阈值的消息数.

4) $TN$(真负例).$TN$ 表示预测传播范围小于阈值,并且实际传播范围小于阈值的消息数.

5) 精确率 $P$(precision).$P$ 表示在所有被预测为热点的消息中,实际为热点的消息所占的百分比,即:

$$P=\frac{TP}{TP+FP}. \tag{9}$$

6) 召回率 $R$(recall).$R$ 表示在所有实际为热点的消息中,被预测为热点的消息所占的百分比,即:

$$R=\frac{TP}{TP+FN}. \tag{10}$$

7) $F1$ 分数($F1\_score$).$F1\_score$ 是统计学中同时兼顾精确率和召回率的一种指标,即:

$$F1\_score=\frac{2\times P\times R}{P+R}. \tag{11}$$

### 4.3 对比方法

1) Embedding-IC[19].它是一种嵌入版本的独立级联模型,充分考虑用户之间的相互影响,把用户嵌入到隐藏投影空间中,借助 EM(expectation-maximization)算法求发送方和接收方的嵌入向量,推测传播概率.根据计算出的传播概率计算最终消息传播范围.

2) Deepcas[6].它是一种消息传播范围预测方法.通过随机采样获得消息扩散的路径,使用 GRU 网络将路径转换为路径的表达向量.最后通过注意力机制来预测消息的传播范围.

3) NT-EP-T.它是 NT-EP 的一种变体.通过时间衰减游走采样传播路径,不利用消息的相互影响.

4) NT-EP-R.它是 NT-EP 的一种变体.不使用时间衰减游走(使用传统的随机游走)采样传播路径,但是利用消息的相互影响.

实验中,传播序列的选择算法使用 C 语言编写,在 VS 环境下编译运行,对比算法也使用 C 语言编写.NT-EP 中神经网络部分使用 python 语言和 tensorflow 框架编写,在 Anaconda3 环境下编译运行.评价标准也使用 python 语言进行处理.所使用的台式机环境为 Intel® Core i7-7700K 4.2 GHz CPU,16 GB RAM,操作系统为 Windows10.

### 4.4 参数选择

我们在验证集中调整模型的超参数,包括用户向量维度 $d$、其他消息影响向量的权重 $\alpha$、时间差的影响参数 $\mu$、学习率 $\lambda$、消息抽取的路径数 $K$、路径长度 $T$ 等.实验中设置算法 1 中计算消息影响向量的学习率 $\lambda=0.000\,5$.

参数选择均使用微博数据进行实验,采样 12 h 传播序列,预测未来 12 h 传播范围,图 3~7 为不同参数下 NT-EP 方法的 MSE 值.我们先随机固定其他参数来考察用户向量维度 $d$ 对 MSE 的影响,实验结果如图 3 所示.随着维度 $d$ 的增加,MSE 的值在逐渐减小,表明预测效果越来越好;但当维度超过 50 时,预测效果改善并不明显.为了平衡预测效果和运行时间,本文后面的所有实验都采用用户向量维度 $d=50$.

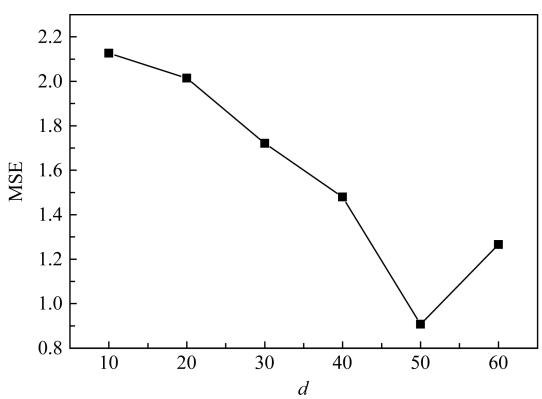


Fig. 3 Influence of different $d$ of user vector on MSE

图 3 用户向量不同维度 $d$ 对 MSE 的影响

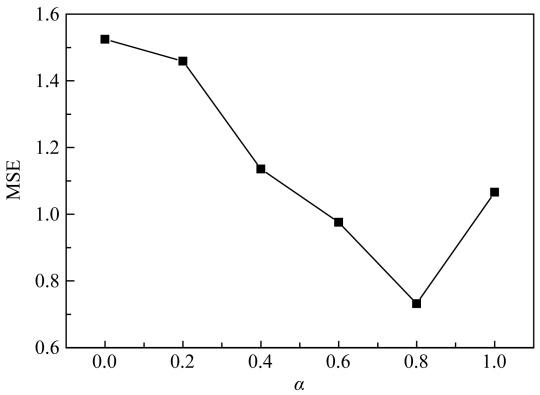


Fig. 4 Influence of different $\alpha$ of selection on MSE

图 4 $\alpha$ 选取对 MSE 的影响

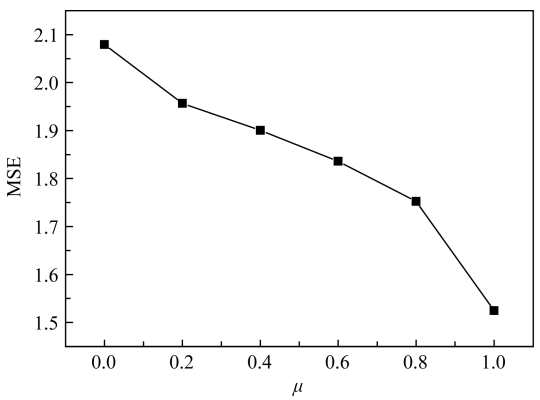


Fig. 5 Influence of $\mu$ value selection on MSE

图 5 $\mu$ 值选取对 MSE 的影响

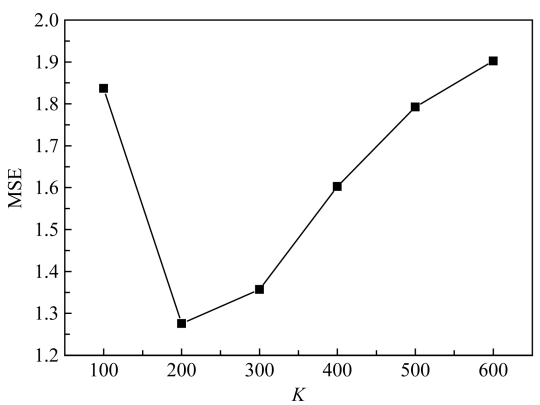


Fig. 6 Influence of $K$ value selection on MSE

图 6 $K$ 值的选择对 MSE 的影响

其他参数的选取也采用上述类似的处理方式.在选取其他消息影响向量的权重 $\alpha$ 时,我们固定用户向量维度 $d=50$.图 4 为 $\alpha$ 选取过程,我们选取 $\alpha$ 时在 0 到 1 之间每 0.2 取值,其中 $\alpha=0.8$ 时 MSE 的取值最优,因此本文后面的所有实验都采用 $\alpha=0.8$.

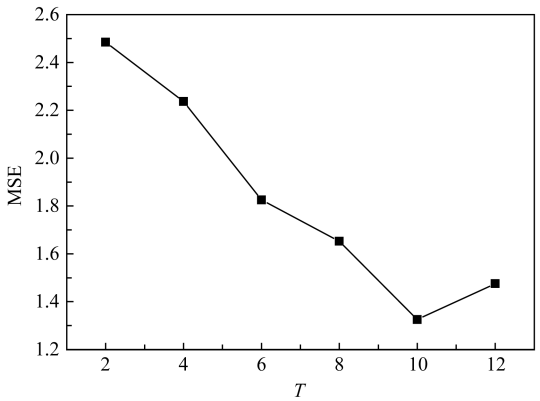


Fig. 7 Influence of $T$ value selection on MSE

图 7 $T$ 值的选择对 MSE 的影响

图 5 给出了参数 $\mu$ 的选择过程,先固定用户向量维度 $d=50$ 和 $\alpha=0.8$,其他参数随机选择,观察参数 $\mu$ 对 MSE 的影响,在 $\mu=1$ 时 MSE 值最小.因此后续实验选择 $\mu=1$ 时作为时间衰减游走采样的参数值.传播路径数量 $K$ 与传播路径长度 $T$ 的选择过程如图 6 和图 7 所示,因此后续实验中我们固定 $K=200$ 和 $T=10$ 作为传播路径数量和传播路径长度.

### 4.5 实验结果

不同方法对传播范围的预测效果如表 2 和表 3 所示.其中表 2 为微博数据上的实验结果,实验中分别采样 12 h,24 h,36 h,然后对未来 12 h,24 h,36 h 的传播范围预测.表 3 为 Flixster 数据集的实验结果,实验中分别采样 10 d,20 d,30 d,然后对未来10 d,20 d,30 d 的传播范围预测.从表 2 和表 3 可以看出,本文方法 NT-EP 及其变体 NT-EP-R 和 NT-EP-T 的预测效果均优于对比方法 Deepcas 和 Embedding-IC.表 3 中的实验结果好于表 2 中的结果,其原因在于 Flixster 数据比微博数据的消息数更多、每个消息的传播范围更广,能让各种模型学习得更充分.

Embedding-IC 方法所在行只有一个实验结果,因为 Embedding-IC 方法和时间长度 $\Delta t$ 无关.Embedding-IC 把所有用户映射到一个向量空间中,通过距离计算用户之间的影响概率.该方法考虑所有用户对当前用户的影响,导致许多无关的用户也进行计算,但实际上激活时间上相近的用户才可能产生影响,所以 Embedding-IC 方法很容易导致过拟合,预测效果较差.Deepcas 使用传统随机游走采样传播路径,没有考虑时间差对传播消息的影响,而且 Deepcas 也没有考虑消息之间的相互影响,因此预测效果并不理想.

**Table 2 MSE Result of Weibo Dataset**

**表 2 微博数据传播范围预测 MSE 结果**

| Method | $t=12$ h | | | $t=24$ h | | | $t=36$ h | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\Delta t=12$ h | $\Delta t=24$ h | $\Delta t=36$ h | $\Delta t=12$ h | $\Delta t=24$ h | $\Delta t=36$ h | $\Delta t=12$ h | $\Delta t=24$ h | $\Delta t=36$ h |
| Embedding-IC | 3.215 | 3.215 | 3.215 | 3.156 | 3.156 | 3.156 | 2.874 | 2.874 | 2.874 |
| Deepcas | 2.070 | 2.292 | 2.079 | 1.864 | 1.682 | 1.727 | 1.739 | 1.575 | 2.024 |
| NT-EP-T | 2.482 | 1.606 | 1.524 | 1.88 | 1.54 | 1.274 | 1.328 | 1.112 | 1.274 |
| NT-EP-R | 0.949 | 1.716 | 1.581 | 1.713 | 1.581 | 1.508 | 1.613 | 1.466 | 1.361 |
| NT-EP | **0.674** | **0.628** | **0.731** | **0.901** | **0.687** | **0.853** | **0.878** | **0.877** | **0.877** |

Notes: $t$ is the sampling time, $\Delta t$ is the future sampling time, the best results are in bold.

**Table 3 MSE Result of Flixster Dataset**

**表 3 Flixster 数据集传播范围预测结果**

| Method | $t=10$ d | | | $t=20$ d | | | $t=30$ d | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\Delta t=10$ d | $\Delta t=20$ d | $\Delta t=30$ d | $\Delta t=10$ d | $\Delta t=20$ d | $\Delta t=30$ d | $\Delta t=10$ d | $\Delta t=20$ d | $\Delta t=30$ d |
| Embedding-IC | 2.134 89 | 2.134 89 | 2.134 89 | 1.978 4 | 1.978 4 | 1.978 4 | 1.967 85 | 1.967 85 | 1.967 85 |
| Deepcas | 0.553 | 0.619 | 0.535 | 0.609 | 0.641 | 0.63 | 0.632 | 0.555 | 0.606 |
| NT-EP-T | 0.509 | 0.473 | 0.413 | 0.448 | 0.379 | 0.311 | 0.44 | 0.514 | 0.513 |
| NT-EP-R | 0.461 | 0.428 | 0.453 | 0.526 | 0.551 | 0.55 | 0.465 | 0.401 | 0.431 |
| NT-EP | **0.429** | **0.388** | **0.341** | **0.23** | **0.205** | **0.166** | **0.381** | **0.375** | **0.257** |

Notes: $t$ is the sampling time, $\Delta t$ is the future sampling time, the best results are in bold.

NT-EP 方法的变体 NT-EP-T 是通过时间衰减游走采样传播路径,不利用消息的相互影响预测时间传播范围.从表 2 和表 3 可以看出,NT-EP-T 优于 Deepcas,说明时间差对消息的传播起重要作用,消息之间确实存在相互影响.NT-EP 的变体 NT-EP-R 利用消息的相互影响,但不使用时间衰减游走采样传播路径.从表 2 和 3 可以看出,NT-EP 方法同时考虑消息的相互影响与时间差对消息传播的作用,预测效果明显优于 Deepcas, NT-EP-T 和 NT-EP-R.

为了进一步验证本文方法的有效性,我们也对热点消息进行了预测,实验结果如表 4 所示.实验中我们使用微博数据采样 12 h,对未来 12 h,24 h,36 h 是否会成为热点消息进行预测,我们在实验中设置的阈值为 1 000.如果传播范围预测值大于阈值,则预测为热点消息.实验结果再次表明本文方法 NT-EP 优于现有方法,也再次证明了消息之间的相互影响确实存在以及消息传播具有时间衰减等特性.

**Table 4 Precision, Recall, *F1_score* Results of Weibo Data**

**表 4 微博数据精确率、召回率、*F1_score* 结果**

| Method | $t=12$ h | | | $t=24$ h | | | $t=36$ h | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | *F1_score* | Precision | Recall | *F1_score* | Precision | Recall | *F1_score* |
| Deepcas | 0.554 | 0.29 | 0.381 | 0.536 | 0.262 | 0.352 | 0.545 | 0.212 | 0.306 |
| NT-EP-T | 0.515 | 0.354 | 0.420 | 0.570 | 0.801 | 0.666 | 0.474 | 0.333 | 0.391 |
| NT-EP-R | 0.541 | 0.418 | 0.472 | 0.541 | 0.290 | 0.383 | 0.518 | 0.312 | 0.405 |
| NT-EP | **0.557** | **0.751** | **0.638** | **0.594** | **0.851** | **0.699** | **0.518** | **0.765** | **0.618** |

Notes: $t$ is the sampling time, the best results are in bold.

## 5 结 论

本文研究了无拓扑结构条件下的消息传播范围预测问题,提出一种社交消息传播范围预测方法 NT-EP.NT-EP 首次利用消息之间的相互影响来提高范围预测的准确性.实验结果表明:NT-EP 在多个评价指标上优于现有的方法 Deepcas 和 Embedding-IC.

未来研究我们准备加入用户兴趣向量和用户基本属性进行范围预测,以及增加多层注意力机制尝试改善预测性能.

## 参 考 文 献


[1] Zephoria, Inc. The top 20 valuable Facebook statistics-updated April 2018 [OL]. [2019-01-01]. https://zephoria.com/top-15-valuable-facebook-statistics/

[2] Yu Honglin, Xie Lexing, Sanner S. The lifecyle of a Youtube video: Phases, content and popularity [C] //Proc of the 9th Int AAAI Conf on Web and Social Media. Menlo Park: AAAI, 2015: 533-542

[3] Althoff T, Jindal P, Leskovec J. Online actions with offline impact: How online social networks influence online and offline user behavior [C] //Proc of the 10th ACM Int Conf on Web Search & Data Mining. New York: ACM, 2017: 537-546

[4] Cheng J, Adamic L, Dow P A, et al. Can cascades be predicted? [C] //Proc of the 23rd Int Conf on World Wide Web. New York: ACM, 2014: 925-936

[5] Tan Chenhao, Lee L, Pang Bo. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 175-185

[6] Li Cheng, Ma Jiaqi, Guo Xiaoxiao, et al. Deepcas: An end-to-end predictor of information cascades [C] //Proc of the 26th Int Conf on World Wide Web. New York: ACM, 2017: 577-586

[7] Islam M R, Muthiah S, Adhikari B, et al. DeepDiffuse: Predicting the 'Who' and 'When' in cascades [C] //Proc of 2018 IEEE Int Conf on Data Mining (ICDM). Piscataway, NJ: IEEE, 2018: 1055-1060

[8] Cheng Li, Guo Xiaoxiao, Mei Qiaozhu. Joint modeling of text and networks for cascade prediction [C] //Proc of the 20th Int AAAI Conf on Web and Social Media. Menlo Park: AAAI, 2018: 640-643

[9] Chen Guandan, Kong Qingchao, Mao Wenji, et al. A partition and interaction combined model for social event popularity prediction [C] //Proc of the 2018 IEEE Int Conf on Intelligence and Security Informatics (ISI). Piscataway, NJ: IEEE, 2018: 232-237

[10] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets [C] //Proc of the 22nd Int Conf on World Wide Web Companion. New York: ACM, 2013: 657-664

[11] Cheung M, She J, Junus A, et al. Prediction of virality timing using cascades in social media [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2017, 13(1): 24-46

[12] Wu Siqi, Rizoiu M A, Xie Lexing. Beyond views: Measuring and predicting engagement in online videos [C] //Proc of the 20th Int AAAI Conf on Web and Social Media. Menlo Park: AAAI, 2018: 434-443

[13] Ren Zhuoming, Shi Yuqiang, Hao Liao. Characterizing popularity dynamics of online videos [J]. Physica A: Statistical Mechanics and Its Applications 2016, 453: 236-241

[14] Shen Huawei, Wang Dashun, Song Chaoming, et al. Modeling and predicting popularity dynamics via reinforced poisson processes [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2014: 291-297

[15] Le Q, Mikolov T. Distributed representations of sentences and documents [C] //Proc of the 12th Int Conf on Machine Learning. New York: ACM, 2014: 1188-1196

[16] Zhang Jing, Liu Biao, Tang Jian, et al. Social influence locality for modeling retweeting behaviors [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2013: 2761-2767

[17] Zhang Jing, Tang Jie, Zhong Yuanyi, et al. Structinf: Mining structural influence from social streams [C] //Proc of the 21st AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 73-79

[18] Yu Linyun, Cui Peng, Wang Fei, et al. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics [C] //Proc of 2015 IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2015: 559-568

[19] Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 1019-1028

[20] Simma A, Jordan M I. Modeling events with cascades of Poisson processes [C] //Proc of the 26th Conf on Uncertainty in Artificial Intelligence. Madrid, Spain: AUAI, 2012: 546-555

[21] Bourigault S, Lagnier C, Lamprier S, et al. Learning social network embeddings for predicting information diffusion [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 393-402

[22] Bourigault S, Lamprier S, Gallinari P. Representation learning for information diffusion through social networks: An embedded cascade model [C] //Proc of the 9th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2016: 573-582


**Liu Zitu**, born in 1994. Master candidate at Heilongjiang University. His main research interests include data mining and social network.

**Quan Ziwei**, born in 1999. Bachelor at Heilongjiang University. Her main research interests include data mining and social network.

**Liu Yong**, born in 1975. Associate professor at Heilongjiang University. His main research interests include data mining and social network.

**Mao Rubai**, born in 1995. Master candidate at Heilongjiang University. His main research interests include data mining and social network.

**Zhu Jinghua**, born in 1976. Professor at Heilongjiang University. Her main research interests include social network, machine learning, recommendation system.