基于用户行为的情感影响力和易感性学习

廖祥文[1),2)] 郑候东[1),2)] 刘盛华[3)] 沈华伟[3)] 程学旗[3)] 陈国龙[1),2)]

[1)](福州大学数学与计算机科学学院 福州 350116)

[2)](福建省网络计算与智能信息处理重点实验室(福州大学) 福州 350116)

[3)](中国科学院网络数据科学与技术重点实验室 北京 100190)

摘 要 在不同情感极性上建模用户间的影响力是观点形成和病毒式营销的一个关键问题.已有工作将用户间影响力直接定义在用户对上,无法刻画未观测到用户对之间的关联关系,造成用户影响力学习的过拟合问题.此外,目前尚无针对不同情感极性的用户间影响力建模的有效方法.因此,该文提出一种融合情感因素的用户分布式表达模型.该模型首先构建两个低维参数矩阵度量在不同情感极性上传播者的影响力和接受者的易感性,然后通过生存分析模型刻画级联的传播行为,最后利用负采样方法解决模型中存在正负例严重不平衡的问题.基于带有情感观点的微博转发所形成级联数据集的实验结果表明,与基准方法对比,该文方法在"预测动态级联"和"谁将会被转发"任务上MRR指标分别提高了273%和32.4%,在"级联大小预测"任务上MAPE指标下降了10.46%,很好地验证了该文模型的有效性.此外,该文分析用户的情感影响力和易感性分布并发现了一些重要的现象.

关键词 在线社交网络;观点传播;影响力;易感性;级联

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2017.00955

Learning Influences and Susceptibilities for Sentiments from Users' Behaviors

LIAO Xiang-Wen[1),2)] ZHENG Hou-Dong[1),2)] LIU Sheng-Hua[3)] SHEN Hua-Wei[3)]
CHENG Xue-Qi[3)] CHEN Guo-Long[1),2)]

[1)](*College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350116)

[2)](*Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fuzhou* 350116)

[3)](*Key Laboratory of Web Data Science and Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Modeling interpersonal influence on different sentiments is a key issue for opinion formation and viral marketing. Previous works directly define interpersonal influence on each pair of users. They fail to depict the unobserved relationships between user pairs and thus suffer from the overfitting problem of learning users' influences. Moreover, there are still not effective solutions to integrate users' sentiments to understand the interpersonal influence. Therefore, we propose a user's distributed representation model with sentimental factors. Firstly, two low-dimensional parameter matrices are applied to represent opinion propagators' influences and opinion recipients' susceptibility on different sentiments. And then, we describe cascade behaviors

收稿日期:2016-06-20;在线出版日期:2016-10-26.本课题得到国家"九七三"重点基础研究发展规划项目基金(2013CB329606,2013CB329602)、国家自然科学基金项目(61572467,61300105)、中国科学院网络数据科学与技术重点实验室开放基金课题(CASNDST20140X)资助.**廖祥文**,男,1980年生,博士,副教授,中国计算机学会(CCF)高级会员,研究方向为观点挖掘和倾向性分析.E-mail:liaoxw@fzu.edu.cn.**郑候东**,男,1990年生,硕士研究生,研究方向为观点挖掘与倾向性分析.**刘盛华**(通信作者),男,1982年生,博士,副研究员,研究方向为数据挖掘、社交网络和情感分析.E-mail:liushenghua@ict.ac.cn.**沈华伟**,男,1982年生,博士,副研究员,研究方向为社会网络分析和网络信息传播,数据挖掘和机器学习.**程学旗**,男,1971年生,博士,研究员,研究领域为大数据分析与挖掘、网络科学、网络与信息安全以及互联网搜索与数据挖掘.**陈国龙**,男,1965年生,博士,教授,研究领域为智能信息处理.

with the survival analysis model. Finally, the imbalance of positive and negative cases is solved by employing negative case sampling technique, according to the distribution of infected users' frequency. Experimental results conducted on Microblog database with different sentiments showed that, compared to the state-of-the-art models, our model improved 273% and 32.4% on MRR metrics on "Predicting Cascade Dynamics" and "Who will Be Retweeted" tasks respectively, and reduced 10.46% on MAPE metrics on "Cascade Size Predicting" task, which verified the validity of our model. Besides, analyzing the distribution of learned users' sentimental influences and susceptibilities resulted in some important discoveries.

Keywords online social networks; opinion propagation; influence; susceptibility; cascade

1 引 言

在线社交网络不仅给用户提供发表个人观点、意见及情感的平台,而且推动着各种网络信息的传播.用户可以通过发布、浏览、转发、点赞和分享消息等行为去影响周围的人,有影响力的用户能促进观点、行为、创新和产品在社交网络中的散播[1].在这种情况下,每对用户之间有一个特定的传播概率,可表示为用户间的影响力[2].因此,找到一种能够更好地刻画级联动态[2]和影响力最大化[3-4]的模型来学习用户间的影响力,对于研究观点形成[5]和病毒式营销[6]等具有重要意义.

目前,大多数工作将用户间影响力定义在用户对的边上.Goyal 等人[2]统计用户之间成功传播对的数目学习影响力,并通过 Bernoulli 模型和 Jaccard Index 模型估算用户间的传播概率作为影响力.但在很多应用场景下,该方法只是记录用户每次被感染的时间,却很少观测到用户间的传播路径,这就限制了基于该观测路径模型的应用.NetInf[7]在先验参数分布条件下,利用指数(Exponential)和幂律(Power-Law)模型来估算用户间影响力.文献[8-10]等通过最大化观测级联的似然值学习用户间影响力.然而这些模型直接使用标量参数定义每对用户之间的影响力,存在两点局限性:(1)参数是独立的,未能刻画由同一个用户产生,作用在不同用户的影响力间的关联关系;(2)如果级联中未能观测到用户对之间的传播或者具有传播的可能,则在这种用户对上的参数是不能被训练的.用户间传播概率将趋近于零或极小的先验值,这意味着在未来这些用户对不会或很少发生信息的传播.另一方面,Aral 和 Walker[11]将用户的影响划分为影响力和易感性两个属性维度,提出利用设计的特征和对应的线性系数对用户间影响力进行建模,这些系数可以通过学习单个用户而非用户对得到.不足的是,用户的属性在其他应用中可能无法或难以获取.总的来说,目前仍然缺乏在不同情感极性上有效地刻画用户间影响力的方法.

针对上述问题,本文提出一种融合情感因素的用户分布式表达模型.该模型假设影响用户观点传播的主要因素是传播者的影响力和接受者的易感性,定义两个低维参数矩阵对它们分别进行表示,并利用生存分析模型[12]和情感帖子被转发过程形成的级联对用户间影响力进行建模.该模型不仅可以有效地减少参数定义,即对于 n 个用户需要 $O(n)$ 参数,而不是 $O(n^2)$ 的用户对参数,有利于降低模型的复杂度,而且能够克服因未能观测到的用户对所导致参数学习过拟合问题.此外,针对观测到的级联中存在正负例严重不平衡问题,本文设计一种负例在数据集中出现的频率进行概率采样的方法.

本文采用新浪微博的数据集进行实验.结果表明,与 Bernoulli,Jaccard Index 和 NetRate 等基准方法对比,本文模型不仅在"预测级联动态"、"谁将会被转发"和"级联大小预测"任务上取得更好的效果,而且能够有效刻画用户在不同的情感极性上所表现出的不同影响力和易感性.更进一步地,通过分析用户的影响力和易感性可以有效挖掘两类重要用户:一类是"原始影响力"用户,具有创造力地发布有吸引力的原帖;另一类是"二次影响力"用户,通过捕获或转发系统中已存在的重要消息以提高自身的影响力得分.此外,通过分析用户活跃度与用户分布表达之间的关系,可以发现:影响力大的用户被他人转发的可能性越大,易感性大的用户转发他人的可能性越大.

本文第 2 节为相关工作;第 3 节为问题描述与动机;第 4 节提出本文的模型;第 5 节介绍实验数据

集；第 6 节为实验，通过与基准实验的对比验证本文方法的有效性，并对用户的情感影响力和易感性进行分析；第 7 节为结束语.

2　相关工作

在线社交网络中用户间的影响力已经成为当前研究的热点. 其中一项工作是提取与传播概率相关的特征，并从观测到的信息级联中学习. Crane 等人[13]利用社交系统中的内在因素和外在因素计算信息传播动态的响应函数. Artzi 等人[14]根据人口(demographic)和内容特征分类预测用户是否会回复或转发一条消息. 除了特征提取外，Tang 等人[15]提出话题因子图 (Topic Factor Graph)来寻找每个用户的话题分布，在大型网络中对话题级社会影响力的生成过程进行建模. 文献[16]提出一种概率因子图模型，对异构网络中相邻与不相邻用户之间的直接影响力和间接影响力进行刻画. Saito 等人[10]将用户受到感染的时间序列作为训练数据，通过独立级联模型来学习有向网络中邻居节点的传播概率从而刻画用户间的影响力. 此外，Goyal 等人[2]分别基于 Bernoulli 和 Jaccard Index 的假设，用计数方法估算用户间的影响力，并把传播概率表示为影响力. Gomez 等人[7]提出 NetInf 算法推理潜在网络，首先分析节点感染次序，接着提出一种融合时间因素的概率模型，最后将传播网络问题归结为最优化问题. Cao 等人[17]基于传播概率的随机游走排序算法 DiffRank，选择传播能力最强的 Top-k 个节点作为观察节点来检测网络中可能出现的信息传播.

与此同时，大量研究者利用生存分析模型及其变体学习用户对间的传播概率，然后用传播速率推断潜在网络. 文献[8]假设在均匀时间窗口和离散空间网络内的信息传播情况下，用户间传播发生的概率取决于节点被感染的时间和节点之间的传播速率，由此提出 NetRate 算法计算每对节点的传播速率. 但是该方法仅适用静态网络，而网络中信息传播的拓扑结构演变非常迅速. 为此，Gomez 等人[9]提出 InfoPath 方法对动态网络进行推理，通过学习随时间变化的用户对间传播速率作为隐藏动态网络的边权. 文献[12]分别引入加法风险和乘法风险(Additive and Multiplicative Risks)建模生存模型中的风险速率(Hazard Rate)以提高级联大小预测的准确性. 然而，这些方法针对的是用户对之间的传播概率，与本文提出的从历史级联中推断特定用户的影响力和易感性的方法截然不同.

在用户影响力的相关研究工作中也指出易感性(susceptibility)而非影响力(influence)才是推动传播现象的关键因素[18-19]. 文献[11]分析影响力和易感性表达的特点，表明传播概率由用户影响力和易感性所决定，通过学习用户间属性的相关度判断影响力用户和易感性用户. Wang 等人[20]基于用户被感染的顺序，提出一种序列化方法学习用户潜在影响力和易感性. 此外，情感传播作为信息传播的重要部分，文献[21-22]在 LiveJournal 和 Facebook 数据集上分别进行实验，结果表明用户的情绪会受到周围其他人的影响. 文献[23]利用格兰杰因果分析发现 Twitter 中观众的情感变化与流行用户的整体情感相关. 因此，本文提出一种学习融合情感因素的用户分布式表达模型，利用连续时间下生存分析模型刻画用户被感染的时间和用户间影响力随时间增大而衰减的规律.

3　问题描述与动机

3.1　问题描述

信息在网络中流动留下了“足迹”，我们称之为级联[24]. 每一条级联表示一个传播过程的时间片(Snapshot)，记录用户被感染后发生的一系列行为，比如用户在新浪微博中发表原帖，其邻居节点看到后将会做出分享、转发、点赞或评论该帖子等行为. 因此，本文定义每条级联 c 为一个时间序列.

$c=\{(v_1,t_{v_1}),(v_2,t_{v_2}),\cdots,(v_N,t_{v_N})|t_{v_1}\leq t_{v_2}\leq\cdots\leq t_{v_N}\}$，其中：$N$ 是级联 c 中感染用户的数量，即级联的大小；t_{v_i} 为级联 c 中用户 v_i 被感染时的时间戳. 并且定义 t_E 为观测到级联 c 的最大时间窗口；M 为用户总数，则未被感染的用户数为 $M-N$. 根据文献[11]，定义“影响力”表示用户影响他人的自身潜在影响力属性，“易感性”表示用户受到他人影响的自身潜在易感性属性. 基于该定义，本文假设用户转发消息的传播速率由用户活跃邻居节点的影响力和自身的易感性所决定的，然后引入生存分析模型对在线网络中一组带有情感极性的帖子被转发形成的级联进行建模.

3.2　模型动机

现有的大多数工作将传播概率定义在网络连边上，这样对于有 n 个用户的网络，需要 n^2 个独立参数来刻画用户间的影响力，即便影响力是由同一个用户产生的. 并且，对于未能观测到的连边用户对

的传播,会导致参数学习过拟合的问题.典型的例子如图1所示,$\{(a,t_a^1),(c,t_c^1),(e,t_e^1),(f,t_f^1)\}$和$\{(a,t_a^2),(b,t_b^2),(e,t_e^2),(f,t_f^2),(d,t_d^2)\}$为两条观测到的级联,实线表示社交关系,虚线表示信息传播路径.可以看出,尽管c,d和e形成了一种社交三角形,我们也很难观测到用户d被感染之前用户c是否被感染.在这种情况下,如果采用现有模型学习用户间的传播概率(Propagation Probability)或传播速率(Transmission Rate),则该值将趋近于零或极小的先验值[25],这也就意味着在未来用户c和d之间不会或很少发生信息的传播.但是,从级联1中可以观测到信息从用户c传播到用户e,从级联2中可以观测到信息从用户e传播到用户d,则用户c通过朋友的关系很可能影响到用户d.因此,本文定义每个用户在不同情感极性上的影响力表达和易感性表达,使得用户间的影响力能够关联到同一个用户的表达.在图1中,(c,d)之间影响力数值和(c,e)之间影响力数值都共同作用到用户c的影响力表达上,使用户c对用户d的影响力数值可以直观地由用户c的影响力和用户d的易感性共同表示,代替了一个极小的先验常数或零.

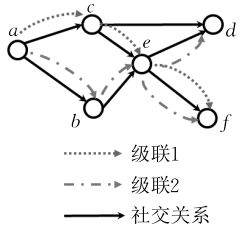


图1 模型基本动机

此外,在一条观测的级联中,未被感染的用户数量远大于感染的用户数量,即$M-N\gg N$,这类用户称为负例.如果考虑所有负例不仅会消耗更多的计算资源,尤其在无网络约束条件下,而且由于正例(Infected Cases)和负例(Uninfected Cases)用户数目严重不平衡,会导致负例的似然在目标函数的优化上占主导地位,如下所示:

$$\max\sum_c \ln\mathcal{L}=\sum_c\sum^{N^c}\ln\mathcal{L}_{pos}^c+\sum_c\sum^{M^c-N^c}\ln\mathcal{L}_{neg}^c,$$

其中:c表示相对应的级联的编号;$\mathcal{L}_{pos}^c$为级联c的正例似然;$\mathcal{L}_{neg}^c$为级联c的负例似然;为级联c的似然.由于M相对较大,即$M-N\gg N$,则求和项的右边更容易支配目标函数.因此,本文提出一种负采样算法来平衡目标函数.具体思路是,假定级联c中某个负例在其他级联中以正例形式出现的频率越大,那么它在该级联未来时间中越有可能被激活,对级联c的似然也会提供更多的信息.于是,根据一组级联数据中被感染用户出现的频率进行负采样是一个比较好的选择.

4 模型建立

4.1 生存分析模型介绍

本文通过引入生存分析模型[12]来建模用户间的影响力,因此简单介绍如下相关知识.

给定非负随机变量T表示事件发生的时刻,下列所有的函数被定义在区间$[0,\infty)$上.

定义1. $f(t)$为T的概率密度函数,则相应的累积分布函数(Cumulative Distribution Function):

$$F(t)=Pr(T\leq t)=\int_0^t f(x)\mathrm{d}x \tag{1}$$

定义2. $S(t)$表示t时刻事件未发生的概率,记为生存函数(Survivor Function),其式子为

$$S(t)=Pr(T\geq t)=\int_t^\infty f(x)\mathrm{d}x \tag{2}$$

定义3. 给定$f(t)$和$S(t)$,风险函数$h(t)$表示事件将发生在t时刻之后的一个极小的Δt区间内,即为瞬时发生率或风险速率,定义为

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t\leq T\leq t+\Delta t\mid T\geq t)}{\Delta t}=\frac{f(t)}{S(t)} \tag{3}$$

定义4. 由于$f(t)=S'(t)$,风险函数$h(t)$与生存函数$S(t)$之间可由对数求导法则关联如下:

$$h(t)=-\frac{\mathrm{d}}{\mathrm{d}t}\ln S(t) \tag{4}$$

定义5. 由于$S(0)=1$,则生存函数$S(t)$可由风险函数$h(t)$表示为

$$S(t)=\exp\left(-\int_0^t h(x)\mathrm{d}x\right) \tag{5}$$

定义6. 由于$f(t)=h(t)S(t)$,由此可得

$$f(t)=h(t)\exp\left(-\int_0^t h(x)\mathrm{d}x\right) \tag{6}$$

4.2 用户间影响力建模

通过上一节分析可知,本文所提出的融合情感因素的用户分布式表达模型中用户的传播属性由两个低维的参数矩阵表示,分别为观点传播者的影响力和观点接受者的易感性.

于是我们记$\boldsymbol{I}_v$为用户v的影响力矩阵,$\boldsymbol{S}_v$为用户v的易感性矩阵,$\boldsymbol{I}_v\in\mathbb{R}^{K\times D}$,$\boldsymbol{S}_v\in\mathbb{R}^{K\times D}$,其中$K$为情感类别数,$D$为每个情感类别上表示用户属性的维度.对于带有情感观点的帖子,定义了一个K维的one-hot向量$\boldsymbol{o}$,表示情感的隶属度.因此,在

带有情感 $\boldsymbol{o}$ 的级联中，用户 u 到用户 v 的传播速率函数 $\phi(\cdot)$ 如方程(7)所示：

$$\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o})=1-\exp\{-\boldsymbol{o}^{\mathrm{T}}\boldsymbol{I}_u\boldsymbol{S}_v^{\mathrm{T}}\boldsymbol{o}\} \tag{7}$$

为了简化式子，用 $\mathcal{H}_{uv}$ 表示 $\{\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o}\}$ 参数集合. 研究表明，用户间的传播概率或影响力会随着时间增加而衰减[2]. 文献[8]提出 3 种融合时间衰减因素的传播概率模型，本文选择一般条件下的幂律模型来刻画信息传播的过程. 假设用户 u 在时刻 t_u 被感染，用户 v 在时刻 t_v 受到用户 u 激活的概率密度函数形式如下：

$$f(t_v\mid t_u,\phi(\mathcal{H}_{uv}))=\begin{cases}\phi(\mathcal{H}_{uv})\cdot(t_v-t_u+1)^{-1-\phi(\mathcal{H}_{uv})}, & \text{若 } t_v>t_u\\ 0, & \text{其他}\end{cases} \tag{8}$$

t_u 到 t_v 时间段，用户 u 对用户 v 感染概率的累积密度函数为

$$\begin{aligned}F(t_v\mid t_u,\phi(\mathcal{H}_{uv}))&=\int_{t_u}^{t_v}\phi(\mathcal{H}_{uv})\cdot(t-t_u+1)^{-1-\phi(\mathcal{H}_{uv})}\,\mathrm{d}t\\&=1-(t_v-t_u+1)^{-\phi(\mathcal{H}_{uv})}\end{aligned} \tag{9}$$

通过生存分析模型可得，用户 v 在 t_v 时刻未被用户 u 感染的概率，即生存函数为

$$S(t_v\mid t_u;\phi(\mathcal{H}_{uv}))=(t_v-t_u+1)^{-\phi(\mathcal{H}_{uv})} \tag{10}$$

用户 v 在 t_v 到 $t_v+\varepsilon$ 被用户 u 感染的概率，即风险函数为

$$h(t_v\mid t_u;\phi(\mathcal{H}_{uv}))=\phi(\mathcal{H}_{uv})\frac{1}{t_v-t_u+1} \tag{11}$$

其中，ε 为无穷小的运行时间，风险概率随时间增大而单调衰减. 同时由于 t_v-t_u 可能为 0 或无穷小的值，加 1 是为了避免无界的风险概率. 这也与文献[8]刻画的幂律模型最小允许时间差(the Minimum Allowed Time Difference)选取为 1 相一致.

假设被“感染”的用户只能影响未被“感染”的用户节点，并且被“感染”的用户在一条级联中只能被感染一次[8]. 级联中已被感染用户都有可能对该用户进行激活. 因此，对于一条级联，非源节点用户 v 在时刻 t_v 被感染的似然为

$$\begin{aligned}f(t_v\mid t,\phi(\mathcal{H}))&=\sum_{v;t_v<t_u}f(t_v\mid t_u;\phi(\mathcal{H}_{uv}))\cdot\prod_{k\neq u,t_k<t_v}S(t_v\mid t_k;\phi(\mathcal{H}_{kv}))\\&=\sum_{v;t_v<t_u}h(t_v\mid t_u;\phi(\mathcal{H}_{uv}))\cdot\prod_{k;t_k<t_v}S(t_v\mid t_k;\phi(\mathcal{H}_{kv}))\end{aligned} \tag{12}$$

给定源节点用户在时刻 t_1 发布原创帖子，一条可观测的传播级联的联合似然为

$$f(t/t_1\mid t_1;\phi(\mathcal{H}))=\prod_{v>1}\sum_{u;t_u<t_v}h(t_v\mid t_u;\phi(\mathcal{H}_{uv}))\cdot\prod_{k;t_k<t_v}S(t_v\mid t_k;\phi(\mathcal{H}_{kv})) \tag{13}$$

如果用户 v_l 在时刻 t_E 后未被感染而成为负例，对应的生存概率为

$$S(t_E\mid t;\phi(\mathcal{H}))=\prod_{u;t_u\leq t_N}S(t_E\mid t_u;\phi(\mathcal{H}_{uv_l})) \tag{14}$$

考虑负例后，一条可观测到级联的 ln 似然式子如下：

$$\begin{aligned}\ln\mathcal{L}(\boldsymbol{I},\boldsymbol{S};\boldsymbol{o})=&\sum_{v>1}\ln\Big(\sum_{u;t_u<t_v}\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o})\frac{1}{t_u-t_v+1}\Big)-\\&\sum_{v>1}\sum_{k;t_k<t_v}\phi(\boldsymbol{I}_k,\boldsymbol{S}_v,\boldsymbol{o})\cdot\ln(t_v-t_k+1)-\\&\sum_{L}\mathbb{E}_{v_l\sim P(u)}\Big[\sum_{u=1}^{N}\phi(\boldsymbol{I}_u,\boldsymbol{S}_{v_l},\boldsymbol{o})\cdot\ln(t_E-t_u+1)\Big]\end{aligned} \tag{15}$$

通过观察可知，级联中负例数目远大于正例数目，一方面最大化所有负例似然限制了本文模型的可扩展性；另一方面，正例和负例数目的不平衡可能会误导优化方向. 因此，本文根据 $P(u)\propto R_u^{3/4}$ 分布[26]采样 L 个用户，其中 R_u 为整组观测数据集中用户 u 被感染的频率，并且在每次迭代优化过程中重复地对负例进行采样.

最后，在不同的情感极性上学习用户的影响力和易感性的优化目标函数为

$$\begin{aligned}&\min_{\boldsymbol{I},\boldsymbol{S}}-\sum_{c}\ln\mathcal{L}^c(\boldsymbol{I},\boldsymbol{S};o^c)\\&\text{s.t. }\boldsymbol{I}_{kd}\geq\boldsymbol{0},\ \boldsymbol{S}_{kd}\geq\boldsymbol{0},\ \forall k,d\end{aligned} \tag{16}$$

其中上标 c 表示级联的编号.

4.3　模型求解

对优化问题(16)的求解是学习用户间影响力的关键步骤. 首先，传播速率函数 $\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o})$ 对 $\boldsymbol{I}_u$ 和 $\boldsymbol{S}_v$ 求导的维度为 $K\times D$，其结果如下：

$$\frac{\partial\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o})}{\partial\boldsymbol{I}_u}=(1-\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o}))\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}\boldsymbol{S}_v \tag{17}$$

$$\frac{\partial\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o})}{\partial\boldsymbol{S}_v}=(1-\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o}))\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}\boldsymbol{I}_u \tag{18}$$

如果级联 c 的消息隶属于第 k 情感类别，即 $o_k=1$，则在 $\boldsymbol{I}_u$ 和 $\boldsymbol{S}_v$ 两个矩阵中仅第 k 行有非零的梯度. 更进一步地，如果用户 u 在级联 c 中被激活，即 $t_1\leq t_u\leq t_N$，则级联 c 的 ln 似然在矩阵 $\boldsymbol{I}_u$ 上的梯度是有效的. 如果非源节点用户 v 在级联中被激活，即 $t_1<t_v\leq t_N$ 或 v 为负例用户，则级联 c 的 ln 似然在矩阵

$\boldsymbol{S}_v$的梯度是有效的.此外,级联 c 中负例在每次迭代中被重复地采样.令$[\mathbb{V}_s^c]_{\mathcal{T}}$为级联 c 在第$\mathcal{T}$次迭代中采样的负例集合,即

$$[\mathbb{V}_s^c]_{\mathcal{T}}=\{v_l\sim P(u)\}_L,$$

其中 L 为集合的大小.

因此,目标函数(16)在矩阵 $\boldsymbol{I}_u$和 $\boldsymbol{S}_v$上的梯度分别如下

$$g_{\boldsymbol{I}_u}=-\sum_c \mathbf{1}(t_1^c\le t_u^c\le t_N^c)\cdot\frac{\partial\mathcal{L}^c(\boldsymbol{I},\boldsymbol{S};o^c)}{\partial\boldsymbol{I}_u} \tag{19}$$

$$g_{\boldsymbol{S}_v}=-\sum_c \mathbf{1}(t_1^c<t_v^c\le t_N^c)\cdot\frac{\partial\mathcal{L}^c(\boldsymbol{I},\boldsymbol{S};o^c)}{\partial\boldsymbol{S}_v}+\sum_c \mathbf{1}(v\in[\mathbb{V}_s^c]_{\mathcal{T}})\cdot\sum_{u=1}^{N^c}(1-\phi(\boldsymbol{I}_u,\boldsymbol{S}_v,\boldsymbol{o}))\cdot\ln(t_E^c-t_u^c+1)o^c\boldsymbol{o}^{c\mathrm{T}}\boldsymbol{I}_u \tag{20}$$

其中 $\mathbf{1}(\cdot)$为指示函数,当满足条件时输出 1,反之输出 0. $g_{\boldsymbol{I}_u}$,$g_{\boldsymbol{S}_v}$为 $K\times D$ 的矩阵,包含目标函数(16)中矩阵 $\boldsymbol{I}_u$和 $\boldsymbol{S}_v$上每个元素的偏导.

接着,采用批量随机梯度下降法(SGD)对目标函数进行求解,这里选取的批量大小为 12.通过投影梯度方法[27](PG)对参数矩阵进行非负约束,于是投影函数 $\phi(x)$表示将参数 x 投影到非负空间,即

$$\phi(x)=\begin{cases}0, & \text{若 } x<0\\ x, & \text{其他}\end{cases}.$$

不难看出,矩阵 $\boldsymbol{I},\boldsymbol{S}\in\mathbb{R}_+^{K\times DM}$表示所有用户 v 对应的 $\boldsymbol{I}_v$合并和 $\boldsymbol{S}_v$合并,M 为先前定义的用户总数.如果不等式(21)条件不满足,则以 $0<\beta<1$ 速率对每个用户 v 的 $\Delta\boldsymbol{I}_v$和 $\Delta\boldsymbol{S}_v$进行一次更新,为 $\beta\Delta\boldsymbol{I}_v$和 $\beta\Delta\boldsymbol{S}_v$,$\beta\in(0,1)$.

$$\mathcal{O}([\boldsymbol{E}]_{\mathcal{T}+1})-\mathcal{O}([\boldsymbol{E}]_{\mathcal{T}})\le\sigma\cdot Tr(\nabla\mathcal{O}([\boldsymbol{E}]_{\mathcal{T}})^{\mathrm{T}}([\boldsymbol{E}]_{\mathcal{T}+1}-[\boldsymbol{E}]_{\mathcal{T}})) \tag{21}$$

其中,$[\cdot]_{\mathcal{T}}$为第$\mathcal{T}$次迭代的参数集合,令 $\boldsymbol{E}=\{\boldsymbol{I},\boldsymbol{S}\}\in\mathbb{R}_+^{K\times 2DM}$,则 $\mathcal{O}(\boldsymbol{E})$是目标函数(16)简化表示. $Tr(\cdot)$表示矩阵的迹,$\sigma\in(0,1)$.

最后,在随机梯度下降法中,学习率(Learning Rate)是影响优化的重要因素,为此本文选择 Adadelta 算法[28]进行自适应地调整学习率.

算法 1. 学习用户分布式表达的算法.

输入:给定 $0<\rho,\beta<1$,常数 σ 和 $\in$;对每个用户 v 初始化参数 $\boldsymbol{I}_v$和 $\boldsymbol{S}_v$;级联集合 $\mathbb{C}$

输出:矩阵 $\boldsymbol{I}_v$和 $\boldsymbol{S}_v$

```
初始化 T:=0
REPEAT
随机洗牌 C 并进行分组;
    FOR 每一组 DO
        使用 Adadelta 方法更新[ΔI_v]_T,[ΔS_v]_T
        更新[I_v]_{T+1}=φ([I_v]_T+[ΔI_v]_T);
        更新[S_v]_{T+1}=φ([S_v]_T+[ΔS_v]_T);
        WHILE 不满足条件(21) DO
            [ΔI_v]_T=β[ΔI_v]_T, [ΔS_v]_T=β[ΔS_v]_T;
            更新[I_v]_{T+1}=φ([I_v]_T+[ΔI_v]_T)
            更新[S_v]_{T+1}=φ([S_v]_T+[ΔS_v]_T)
        END WHILE
        T:=T+1;
    END FOR
UNTIL 参数收敛或达到最大迭代次数
```

算法时间复杂度分析:对于给定的信息级联 c 考虑到 $t_u<t_v\le t_E$和 $t_v>t_E$的情况下,假设级联 c 由 N 个节点组成,用户总数为 M,数据集 $\mathbb{C}$ 分为 k 组,每组 b 个级联,则单个级联 c 通过 Adadelta 方法更新的时间为

$$1+2+\cdots+(N-1)+(M-N)=O\left(\frac{N^2}{2}\right),$$

则对应 b 个级联通过 Adadelta 方法更新时间为 $O(b\cdot N^2)$,同理对式子(21)计算的时间为 $O(b\cdot N^2)$.因此算法 1 的时间复杂度为

$$O(m\cdot b\cdot N^2\cdot\mathcal{T}),$$

其中 m 为不满足条件(21)操作的次数,$\mathcal{T}$为迭代的次数.

5 数据集描述

本文基于新浪微博开放的 API 接口①,采用宽度优先策略进行数据采集.首先选择部分用户作为初始节点,抓取微博信息,然后以他们所关注的用户,抓取相关内容,以滚雪球方式扩大采集范围.最后我们收集了大约 3.156 亿条微博记录,包括原始帖子,转发帖子和@帖子的消息,原始帖子的时间跨度为 2013-11-01 至 2014-02-28.由于表情字符在消息级联中通常作为情感指标,因此我们参考维基百科②上的表情字符列表,并为其标注情感极性.之后筛选出包含表情字符的高频率被转发的原帖记录,从这些帖子中抓取用户间的转发关系和转发时间作为实验数据集.同时,在不考虑中性情感帖子的情况下,本文定义,如果一条帖子中包含的正面表情字符个数大于负面表情字符个数,则为正面情感帖子,相应的该帖子被转发过程形成的级联为正面情感级联.反之为负面情感级联.

① http://www.weibo.com
② https://en.wikipedia.org/wiki/List_of_emoticons

通过对数据集统计，可以发现许多用户在转发关系中出现的频率较低，例如只出现过一次的用户数量为 886 039，这对模型的训练和预测产生了很大的干扰. 同时，为了使应用更广泛，本文在用户间转发关系和被转发关系的情况未知下，仅保留了用户被感染时的时间序列作为数据集. 接着，对数据集进行预处理，步骤如下：

(1) 由于帖子发布及被转发的时效性，因此若帖子转发过程中前后两次被转发的间隔超过一周，则将其之后的帖子转发链去掉.

(2) 定义用户 v 的活跃度 A_v 为数据集中用户 v 转发他人的次数 $A._v$ 和用户 v 被转发的次数 $A_{v.}$ 之和，即 $A_v = A._v + A_{v.}$.

(3) 选取数据集中用户活跃度超过 40 的用户作为种子用户(4853 人)，对于每条帖子转发关系链，按活跃用户所占的比例从大到小进行排序后删除小比例的级联，最终得到一组级联. 如表 1(a)所示，过滤后的数据集时间跨度从 2013-10-31 至 2014-03-03，其中共有 6219 个用户，所有级联的大小总和为 44 021. 数据集中有 325 个正面情感级联记录，412 个负面级联记录，两种情感极性的级联数正好保持均衡. 表 1(b)展示了级联中用户活跃度的中位数为 5，众数为 4，表明在数据集中可以较多地观测到用户的行为，保证了模型学习的有效.

表 1　数据的主要统计特征

(a)

时间跨度	用户总数	级联总大小	情感	
			正面	负面
10/31/13 至 03/03/14	6219	44 021	325	412

(b)

用户活跃度		级联大小	
中位数	众数	中位数	众数
5	4	37	10

图 2 为级联的时间跨度分布，可以看出，时间跨度为 2 d 的级联数目最多，有 257 条. 而且数据集中有 92.6%的级联生存周期在 8 d 之内. 图 3 给出了级联大小的累积分布情况，其中级联大小在 10～100 之间分布的比较密集. 图 4 展示了级联中包含正负两种表情字符高使用频率的分布情况.

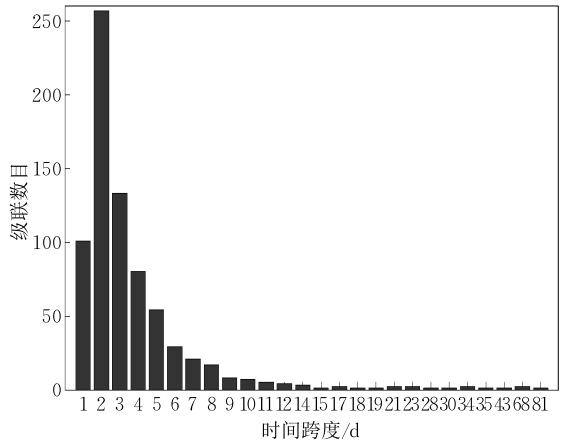


图 2　级联时间跨度分布

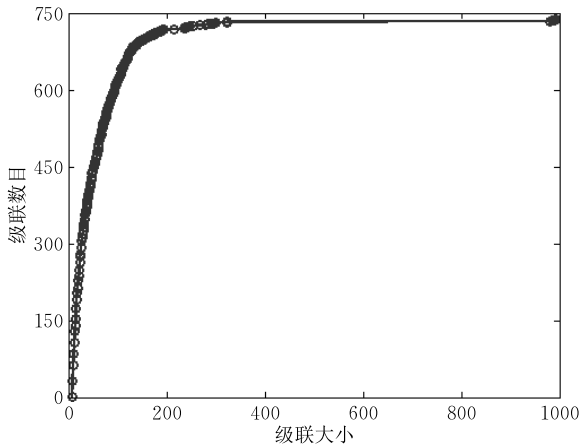


图 3　级联大小累积分布

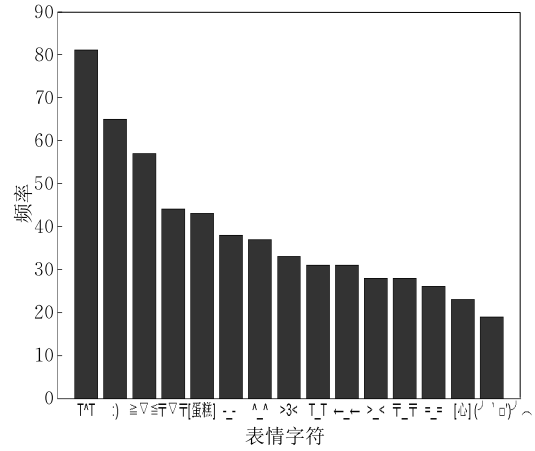


图 4　表情字符频率的分布

6　实　验

6.1　实验设计

实验环境为 Ubuntu 12.04.5 LTS，Java(TM) SE Runtime Environment (build 1.7.0_75-b13)，AMD Opteron(tm) Processor 6320，32 GB 内存. 将基准方法与本文的方法应用在相同的数据集，实验选取了以下对比模型：

(1) CT Bernoulli(Continuous Time Bernoulli)和 CT Jaccard(Continuous Time Jaccard)模型[2]. 两种模型都是连续时间模型，融合了时间衰减因素.

为了公平对比,采用相同的衰减函数来表示用户 u 感染用户 v 的传播概率 P_{uv},即 $P_{uv}=P^0_{uv}/(t_v-t_u+1)$. 假设被感染用户只能影响未被感染用户,并且被感染用户只能被感染一次. 此外,CT Bernoulli 模型初始传播概率 P^0_{uv} 服从 Bernoulli 分布,即 $P^0_{uv}=\mathcal{A}_{u2v}/\mathcal{A}_u$,CT Jaccard 模型初始传播概率 P^0_{uv} 服从 Jaccard Index,即 $P^0_{uv}=\mathcal{A}_{u2v}/\mathcal{A}_{u|v}$. 其中 $\mathcal{A}_{u2v}$ 表示在级联中用户 u 感染用户 v 次数,$\mathcal{A}_{u|v}$ 表示级联中用户 u 或用户 v 出现的次数,但要去掉用户 u 和 v 同时出现的次数 $\mathcal{A}_{u\&v}$,即 $\mathcal{A}_{u|v}=\mathcal{A}_u+\mathcal{A}_v-\mathcal{A}_{u\&v}$. 由于数据集中仅观测到用户被感染的时间,因此用户被感染的过程是一个时间序列过程,未被感染用户只能由先前已感染用户进行激活.

(2) NetRate[8] 算法. 该算法直接将参数定义在用户对的边上,在给定的时间窗口内,通过生存分析模型来学习用户间的传播速率. 由于 Jaccard Index 被认为是一种良好的传播概率估计方法[2]. 因此,我们采用 Jaccard Index 的结果作为 NetRate 算法的初始化参数. 上述 3 种模型都是基于 pair-wise 方法建模用户间的影响力.

(3) CT LIS 模型(Continue Latent Influence and Susceptibility). 该模型忽略了用户的潜在影响力和易感性分别在情感极性上的差异,通过定义两个向量 $\boldsymbol{I}_u$ 和 $\boldsymbol{S}_v$ 分别度量用户 u 的影响力和用户 v 的易感性,则用户 u 到用户 v 传播速率表示为 $\varphi_{uv}=\boldsymbol{I}_u^{\mathrm{T}}\cdot\boldsymbol{S}_v$. 文献[20]定义了类似的参数,采用一种静态方法刻画用户行为的过程. 而本文使用的"CT LIS"是连续时间模型的升级版本.

(4) Sent LIS 模型(Sentimental Latent Influence and Susceptibility). 本文考虑到不同情感观点的帖子对用户在对应情感上的影响力表达和易感性表达存在着差异性. 也就是说,用户的情感传播由用户所转发帖子的情感极性决定,并且设计了加入所有负例情况下学习融合情感因素的用户分布式表达模型.

(5) Sent LIS(neg sample)模型(Sent LIS with Negative Sample). 通过加入负采样算法对 Sent LIS 模型进行学习.

6.2 实验任务及评价指标

为了评估本文提出模型的有效性,采用了以下实验任务和相应的评价指标进行实验:

(1) PCD(Predicting Cascade Dynamics). 预测级联动态. 主要针对不同情感帖子被转发所形成的级联中,预测被感染用户和相应行为发生的时刻. 然而,为了使该任务简单且易于评估,我们仅在给定的时刻 t_v,预测用户 v 是否被感染,其中 t_v 为训练级联中用户真实转发帖子的时刻. 因此,给定级联真实发生时刻 t_v,可以通过函数 $f(t_v|t;\phi(\mathcal{H}_{\cdot v}))$ 计算出 t_v 之前未被感染用户在时刻 t_v 被感染的概率. 由于被感染概率排名越靠前的用户越有可能被激活,因此采用 MRR(Mean Reciprocal Rank)平均倒数排名[29] 评价指标来计算级联中每个真实时刻用户被感染的概率排名,其式子如下:

$$\mathrm{MRR}=\frac{1}{|N|}\sum_{v=1}^{|N|}\frac{1}{rank_{t_v}},$$

其中:$rank_{t_v}$ 为级联中真实发生在第 t_v 时刻条件下,真实被感染用户计算出来的被感染概率在整个未被感染用户集合的排名;N 为对应的级联大小. MRR 越大,对应的评价效果越好.

此外,该任务可以视为一组二元分类问题,在级联中用户真实被感染的情况下,采用已被感染用户作为正例,到时刻 t_E 后未感染用户作为负例. 那么正例 v 在给定的真实时刻 t_v 被感染概率为 $f(t_v|t;\phi(\mathcal{H}_{\cdot v}))$,负例 v_l 的激活概率为 $f(t_N+\varepsilon|t;\phi(\mathcal{H}_{\cdot l}))$,其中 ε 是一个非常小的常量. 因此,对所有用户感染的概率进行从大到小排序后,在给定的阈值下,可以通过 AUC[30] 指标(ROC 曲线下的面积)来评价,即由 ROC 横坐标 FPR(False Positive Rate)和纵坐标 TPR(True Positive Rate)所围成的面积,计算公式如下:

$$\mathrm{FPR}=\frac{\text{负例不正确分类个数}}{\text{负例总个数}},$$

$$\mathrm{TPR}=\frac{\text{正例正确分类个数}}{\text{正例总个数}}.$$

该面积范围一般在(0.5,1)之间. 其物理意义为任取一对正例和负例,正例得分大于负例得分的概率. AUC 越大,表明正例被激活的概率就越大,对应的评价效果也越好.

(2) WBR(Who will Be Retweeted). 谁将会被转发. 如果微博用户被感染,则该用户所发生的行为就是对周围朋友发表的帖子进行转发,评论和点赞等. 因此,对于"谁将会被转发"进行预测是一种定量地评估用户间影响力的方法. 在线社交网络中,高影响力的用户所发表的情感帖子有更大的概率被转发. 令 (v,t_v) 表示用户 v 在 t_v 时刻发生行为,用户 v 所转发的感染用户为

$$\arg\max_{u:t_u<t_v} f(t_v\,|\,t_u;\phi(\mathcal{H}_{uv})).$$

从而我们把预测任务看作影响力排名的问题,具有较高排名的用户更有可能被转发. 本文采用排名第一的平均精度(Acc)[20] 和 MRR 作为评估预测

的指标，MRR 和 Acc 越大表示有更好的预测效果.

(3) CSP(Cascade Size Predicting). 级联大小预测是评价社交网络中用户影响力的一个重要部分，对于信息传播和病毒式营销具有指导意义. 为了提高预测的有效性，本文选取了每条真实级联中前 P 个感染用户为已知级联的长度，对剩余的时间段 t_N-t_P 进行均匀等分 R 个 Δt 时间段后，分别预测每个时刻被感染的用户数量. 假设已被感染的用户集为 $\mathcal{P}(u)$，给定时刻 t_u 感染用户 $u\in\mathcal{P}(u)$ 和时刻 t_v 未被感染用户，当 $t_v=t_{P+1}$，则在 t_u 到 t_v 时间段内，用户 u 对用户 v 感染的概率为

$$P(t_v\mid t_u,\phi(\mathcal{H}_{uv}))=F(t_v\mid t_u,\phi(\mathcal{H}_{uv})).$$

当 $t_{P+1}<t_v\leq t_N$，则在 t_u 到 t_v 时间段内，用户 u 对用户 v 感染的概率为

$$P(t_v\mid t_u,\phi(\mathcal{H}_{uv}))=\frac{F(t_v\mid t_u,\phi(\mathcal{H}_{uv}))-F(t_v-\Delta t\mid t_u,\phi(\mathcal{H}_{uv}))}{1-F(t_v-\Delta t\mid t_u,\phi(\mathcal{H}_{uv}))}.$$

接着，我们每次取 Top-k 的用户间感染概率 $P(t_v\mid t_u,\phi(\mathcal{H}_{uv}))$ 进行抽样来确定用户 v 是否被感染. 实验采用平均绝对百分误差(Mean Absolute Percentage Error(MAPE))对预测精度进行检验，其式子如下：

$$\mathrm{MAPE}=\frac{1}{C}\sum_{c=1}^{C}\left|\frac{M_c-F_c}{M_c}\right|,$$

其中：M_c 为级联 c 大小的真实值；F_c 为级联 c 大小的预测值；C 为测试级联的总数. MAPE 的值越小，表示预测结果越好.

6.3　实验结果分析

对数据进行预处理后，我们将整个数据集在不考虑区分用户情况下均匀分成 10 组，在每个任务下进行十倍交叉验证，并记录相应评价指标的平均值和标准偏差. 同时，考虑到实验环境与机器的性能和评价任务计算的有效性，本文设置用户分布式表达在情感极性上的维度 $D=8$，对于矩阵 $\boldsymbol{I}_v$ 和 $\boldsymbol{S}_v$ 参数的初始化，每一维度上的元素通过函数 $f(x)=\sqrt{x}$，$x\sim U(0,0.1)$ 采样[20]得到.

6.3.1　评价结果分析

(1) PCD. 将本文方法得到的 10 组评价指标平均值和标准差(SD)与基准方法进行对比，结果如表 2 所示. 从实验结果得知：本文提出的模型“Sent LIS”和“Sent LIS(neg sample)”在 MRR 指标上分别达到 0.0216 和 0.0265，以 p-value$<$0.01 显著性优势压倒其他模型，说明加入负采样在平衡正负例数量上起到了效果. 此外，在 pair-wise 模型中，NetRate 通过调整 Jaccard Index 学习出来的参数 MRR 有了较大的提升，并且“CT Jaccard”比“CT Bernoulli”效果好，与文献[2]所述在传播概率估计上 Jaccard Index 模型优于 Bernoulli 模型说法一致.

表 2　10 倍交叉验证下 PCD 任务的平均 MRRs 和 AUCs

	CT Bernoulli	CT Jaccard	NetRate (Jaccard)	CT LIS	Sent LIS	Sent LIS (neg sample)
MRR	0.0062±0.0029	0.0064±0.0036	0.0071±0.0038	0.0196±0.0039	**0.0216**±0.0033	**0.0265**±0.0044
AUC	0.8739±0.0658	0.8621±0.0802	0.8718±0.0730	0.8793±0.0207	**0.8992**±0.0152	**0.8983**±0.0156

在二分类测试上，“Sent LIS”和“Sent LIS(neg sample)”的 AUC 指标优于其他模型，分别为 0.8992 和 0.8983，前者结果较好. 此外，机器学习模型NetRate 的 AUC 在 3 个 pair-wise 模型中是最好的. 图 5 显示 10 倍交叉验证中的一组 ROC 曲线，进一步验证了本文提出的模型“Sent LIS”，“Sent LIS(neg sample)”以及“CT LIS”在 AUC 上有更好的表现. 综上分析，在预测级联动态的排名和二分类问题上，本文提出的学习用户分布式表达模型比 pair-wise 模型在缓解过拟合问题以及降低模型的复杂性上具有更大的优势.

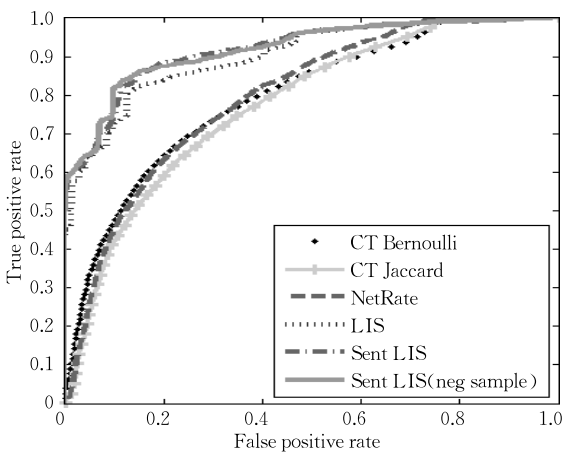


图 5　PCD 评价任务的 ROC 曲线

(2) WBR. 这里仍采用 10 倍交叉验证方法来计算排名为 1 的 Acc 和 MRR 指标的平均值与标准差(SD). 表 3 给出了实验结果，可以得出，“CT LIS”、“Sent LIS”和“Sent LIS(neg sample)”结果优于 pair-wise 模型，这是由于 pair-wise 模型对于未能观测到的传播用户对会导致参数学习的过拟合问题. 与 NetRate 相比，3 种 LIS 模型在预测“谁将会被转发”的排名为 1 的 Acc 指标上分别提升了 37.2%，27.8% 和 32.4%，而在 MRR 指标上分别提升了

表 3 10 倍交叉验证下 WBR 任务的平均 Acc and MRRs

	CT Bernoulli	CT Jaccard	NetRate (Jaccard)	CT LIS	Sent LIS	Sent LIS (neg sample)
Acc	0.1221±0.0365	0.3000±0.0964	0.3005±0.0961	**0.4123**±0.0874	0.3840±0.1255	**0.3980**±0.1392
MRR	0.2592±0.0703	0.4349±0.1275	0.4354±0.1273	0.4696±0.0876	**0.4822**±0.1269	**0.4920**±0.1348

7.9%,10.7%和13.0%.此外,"CT Jaccard"在两个评价指标中结果都高于"CT Bernoulli",而且"NetRate(Jaccard Index)"是 pair-wise 模型中最好的结果.综上分析,通过负采样方案,"Sent LIS(neg sample)"一方面可以平衡正例和负例的数目,另一方面使得目标函数向更好的梯度方向优化.而且,相比于"Sent LIS"模型,它在 Acc 和 MRR 上表现出更好的结果也证明负采样方法的优势.

(3)CSP.首先对参数进行设置,这里取级联中初始已被感染的用户数为 10,$R=10$, Top-$k=5$ 进行概率抽样.通过 10 倍交叉验证方法对每组测试集进行 100 轮抽样来计算 MAPE 指标,如表 4 所示,结果表明"CT LIS"、"Sent LIS"和"Sent LIS(neg sample)"在 MAPE 指标上明显优于 pair-wise 模型的 3 种方法,分别为 0.6259,0.6259 和 0.6362.而且,前两者比负采样算法结果会好点,可能是概率抽样随机性所导致的.此外与 pair-wise 模型中"CT Jaccard"对比,MAPE 指标至少下降了 10.46%.因此,该方法通过用户的分布式表达在级联大小的预测上表现出更好的效果.

表 4 10 倍交叉验证下 CSP 任务的平均 MAPE

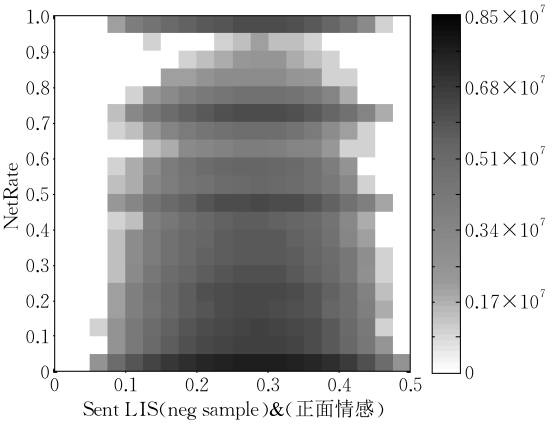
	CT Bernoulli	CT Jaccard	NetRate (Jaccard)	CT LIS	Sent LIS	Sent LIS (neg sample)
MAPE	0.7199±0.0270	0.7105±0.0333	0.7109±0.0350	**0.6259**±0.0883	**0.6259**±0.1458	**0.6362**±0.2252

6.3.2 传播速率分析

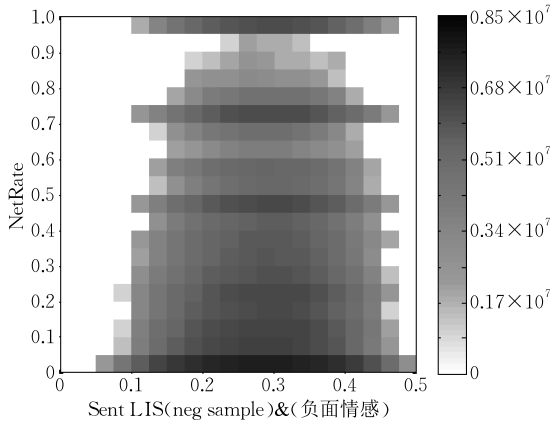
为了说明本文提出的"Sent LIS(neg sample)"模型和 NetRate 算法学习出的传播速率的差异.首先,针对用户情感影响力矩阵 $\boldsymbol{I}_u$ 和易感性矩阵 $\boldsymbol{S}_v$,利用式(7)计算 $\boldsymbol{I}_u$ 和 $\boldsymbol{S}_v$ 矩阵中对应每一行上影响力向量和易感性向量的内积,表示不同情感极性上的用户 u 到用户 v 的传播速率.然后取两种模型在对应情感上的传播速率构成一个坐标点,其中横坐标为本文提出的模型,纵坐标为 NetRate 算法,并统计落入到每个单元格中坐标点的数量.如图 6 所示,根据颜色棒可以看出单元格中落入该区域坐标数量的多少.图 6(a)和(b)表示本文模型和 NetRate 算法下两种情感极性的热度图.可以看出,在正面情感或负面情感上都有一条深颜色的长网格落在 X 轴 0.15 至 0.4 区间上,说明通过 NetRate 算法学习出很多过拟合的传播速率,这些值通常为零或很小的常数,而通过本文提出的用户分布式表达学习出来的传播速率可以较为明显地将其区分开来.另外,在与 X 轴平行的深色调单元格中,NetRate 算法学习出的高传播速率在本文的模型下也有明显的区分.此外,图 6(c)和(d)的 Jaccard 模型与图 6(e)和(f)的 Bernoulli 模型也得出相同结论.由此表明,本文模型更加能够区分影响力和易感性使得在评价结果中表现得更好.

6.3.3 用户情感影响力和易感性分析

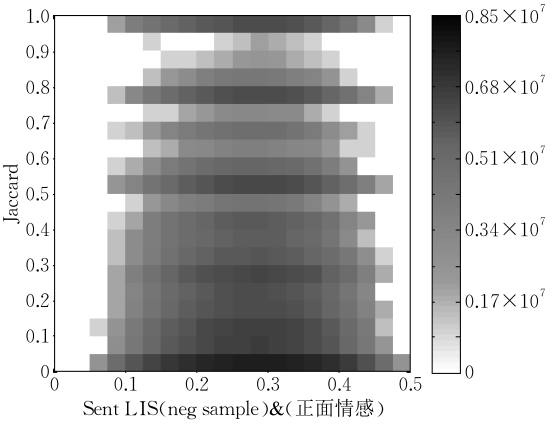
除了评价模型的比较外,本文对用户 v 在不同情感上的低维参数矩阵 $\boldsymbol{I}_v$ 和 $\boldsymbol{S}_v$ 也进行了分析.矩阵 $\boldsymbol{I}_v$ 和 $\boldsymbol{S}_v$ 中每一行分别表示用户 v 在相应的情感极性上的影响力和易感性表达,分别记为"正面影响力"、"负面影响力"、"正面易感性"和"负面易感性",并通过计算这些行向量的 L1 范数来衡量用户在不同情感上的影响力和易感性大小.然后使用这些 L1 范数的值来表示用户的坐标点,统计落入到每个单元格的数量.图 7(a)和图 7(b)分别为在正面情感与负面情感下用户影响力与易感性的三维等高线地图.可以看出,两张地图中都出现两个峰值,其中一个峰值出现在"正面易感性"和"负面易感性"的 L1 范数为零的坐标轴上,这些用户往往具有较大的影响力而不易受到别人的感染[11],记为在正面或负面情感上的"原始影响力"(Original Influentials).另外一个峰值出现在等高线地图的右上方,可以看出这些用户自身拥有较高影响力的同时也很活跃地去转发别人的消息,记为在正面或负面情感上的"二次影响力"(Secondary Influentials).换句话说,拥有"二次影响力"的用户可能通过转发有吸引力的消息来提高自身的关注度和影响力.而"原始影响力"的用户更倾向于发表有吸引力的原创帖子.因此,"原始影响力"在社交网络中一般为大 V 用户,而"二次影响力"一般为让他人获取消息的广告商或倾向转发热门微博的用户.更进一步,我们分析了用户的影响力和易感性在不同情感上的分布情况.图 7(c)和图 7(d)以二维视图展示了等高线地图的一个主峰.



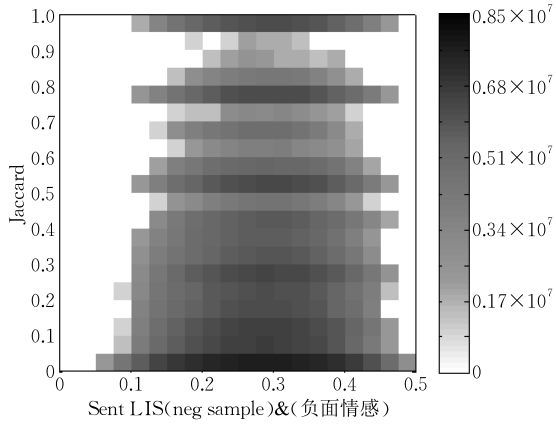
(a)正面情感Ours v.s. NetRate的传播速率



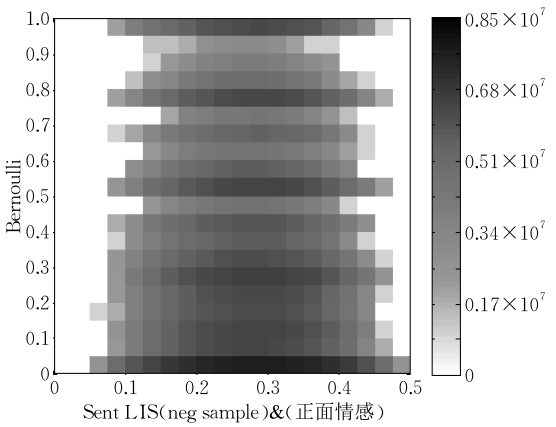
(b)负面情感Ours v.s. NetRate的传播速率



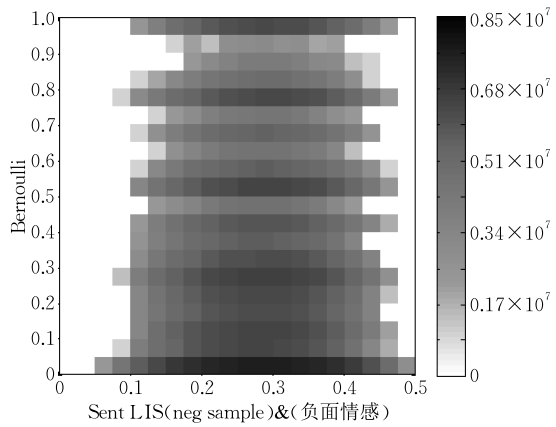
(c)正面情感Ours v.s. Jaccard的传播速率



(d)负面情感Ours v.s. Jaccard的传播速率



(e)正面情感Ours v.s. Bernoulli的传播速率



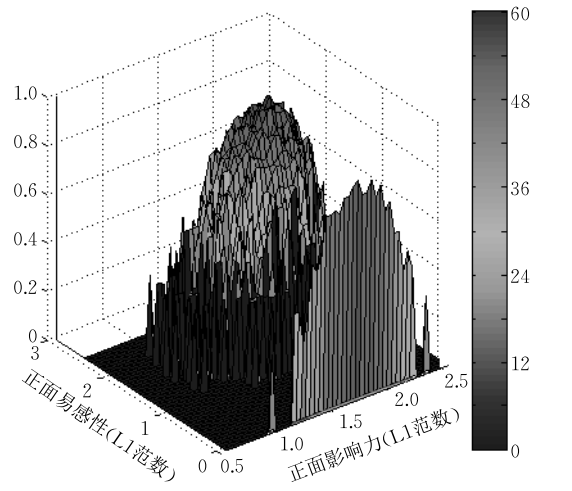
(f)负面情感Ours v.s. Bernoulli的传播速率

图 6　传播速率的分析

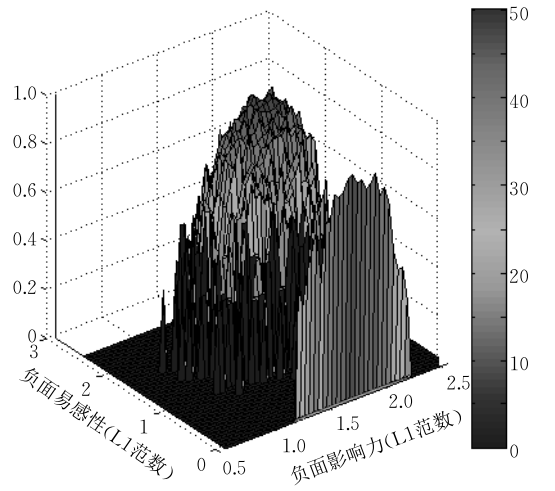
由图 7(c)可以获知，在正面情感上拥有高影响力的用户往往在负面情感上拥有较低的影响力，则该类用户比较倾向于发布正面帖子吸引更多粉丝. 反之亦然. 同时可以看出还有一部分比例的用户在两种情感极性上的影响力几乎相当. 由图 7(d)在用户易感性上可以得出相同的结论，即一些用户对正面情感更敏感些，而其他用户对负面情感更敏感些. 因此，用户在不同的情感极性上可能会表现出一些不同的行为.

6.3.4　用户活跃度与用户分布式表达的关系分析

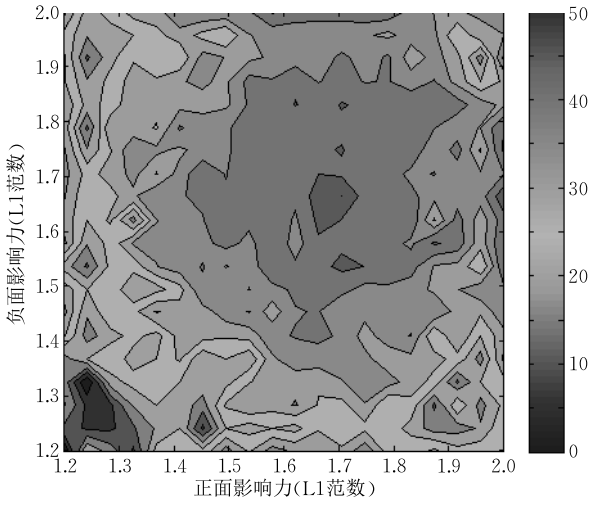
在线社交网络中用户的活跃度往往也是影响用户间观点传播的一个重要因素. 为此，本文将活跃度分为转发用户活跃度与被转发用户活跃度并对其进行分析. 一般而言，转发用户活跃度表示用户转发他人消息的活跃程度，与用户的易感性相关. 而被转发用户活跃度表示影响他人的活跃程度，与用户的影响力相关. 因此，本文采用 L1 范数衡量用户在不同



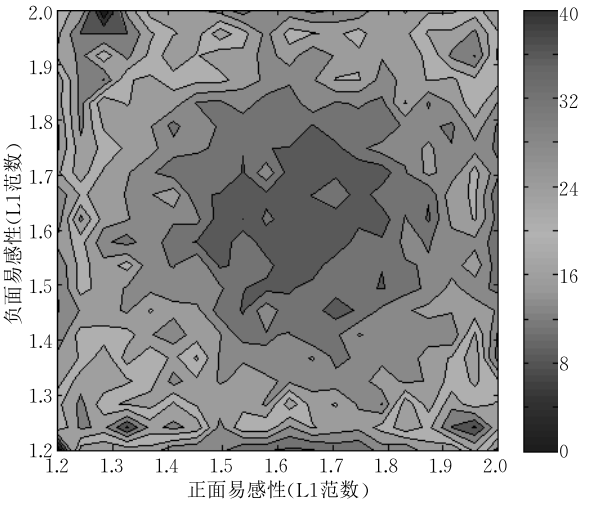
(a) 正面影响力v.s.正面易感性



(b) 负面影响力 v.s.负面易感性



(c) 正面影响力v.s.负面影响力



(d) 正面易感性v.s. 负面易感性

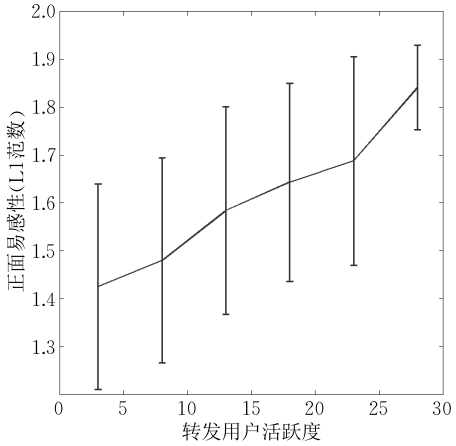
图 7 情感影响力 L1 范数和情感易感性 L1 范数分析

情感上的影响力和易感性大小,分析与用户活跃度之间的关系.结果如图 8 所示的误差图,图中的曲线表示相同用户活跃度下用户分布表达 L1 范数的平均值,误差线为偏离平均值的程度,即标准偏差.在图 8(a)和(b)中,转发用户的易感性分别在正负情感上随着用户活跃度增大总体呈现上升趋势,尤其在正面情感上更为显著.同理,图 8(c)和(d)展示了被转发用户活跃度与正负情感之间的关系,可以看出,两者之间也呈现出正相关的趋势.由此表明影响力大的用户被他人转发的可能性越大,易感性大的用户转发他人的可能性越大.

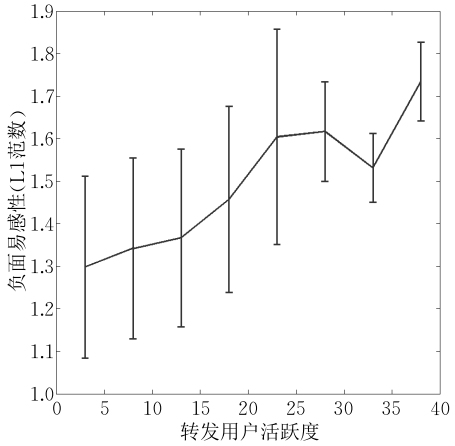
7 结束语

本文提出了一种融合情感因素的用户分布式表达模型学习用户间的影响力.首先,定义两个低维参数矩阵分别表示在不同情感极性上观点传播者的影响力和观点接受者的易感性,该方法不仅降低了模型参数的复杂度,而且缓解了参数学习过拟合的问题.其次,设计了一种负采样算法对引入生存分析模型的级联最大化似然进行求解,该算法有效地克服了模型中正负例严重不平衡的现象,并且使模型适用于更大规模的级联数据集.最后,与基准方法对比,本文模型在“级联的动态”、“谁将会被转发”、“级联大小预测”等任务上表现出较好的性能.

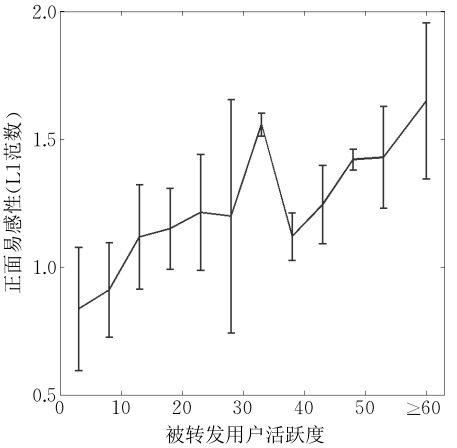
此外,本文对用户情感影响力和易感性进行分析,挖掘出“原始影响力”和“二次影响力”两类重要用户.前者通过发布高质量的原创消息影响他人,后者通过转发热门或有趣的消息引起他人转发.同时,发现用户在不同的情感消息上可能表现出不同行为.通过分析用户活跃度与用户分布式表达之间的关系,我们也发现活跃度高的转发用户可能有高易



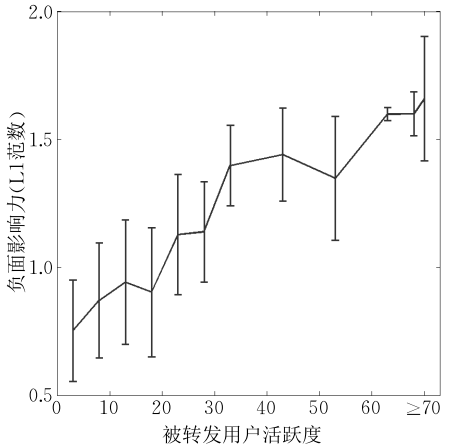
(a) 转发活跃度与正面易感性



(b) 转发活跃度与负面易感性



(c) 被转发活跃度与正面影响力



(d) 被转发活跃度与负面影响力

图 8　用户活跃度与用户潜在属性 L1 范数关系

感性，活跃度高的被转发用户可能有高影响力. 在接下来的工作中，我们将继续对用户的影响力和易感性的潜在信息进行挖掘并应用于用户间的影响力分析.

参　考　文　献

[1] Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512

[2] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Web Data Mining. New York, USA, 2010: 241-250

[3] Gionis A, Terzi E, Tsaparas P. Opinion maximization in social networks//Proceedings of the 13th SIAM International Conference on Data Mining. Austin, USA, 2013:387-395

[4] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146

[5] Bindel D, Kleinberg J, Oren S. How bad is forming your own opinion? Games and Economic Behavior, 2015, 92: 248-265

[6] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing. ACM Transactions on the Web, 2007, 1(1): 5

[7] Gomez Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1019-1028

[8] Rodriguez M G, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 561-568

[9] Gomez Rodriguez M, Leskovec J, Schölkopf B. Structure and dynamics of information pathways in online media//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 23-32

[10] Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model//Proceedings of the International Conference on Knowledge-

Based and Intelligent Information and Engineering Systems. Zagreb, Croatia, 2008: 67-75

[11] Aral S, Walker D. Identifying influential and susceptible members of social networks. Science, 2012, 337(6092): 337-341

[12] Gomez-Rodriguez M, Leskovec J, Schölkopf B. Modeling information propagation with survival theory//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 666-674

[13] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 2008, 105(41): 15-649-15-653

[14] Artzi Y, Pantel P, Gamon M. Predicting responses to microblog posts//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 602-606

[15] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816

[16] Liu L, Tang J, Han J, et al. Mining topic-level influence in heterogeneous networks//Proceedings of the 19th ACM Conference on Information and Knowledge Management. Toronto, Canada, 2010: 199-208

[17] Cao Jiu-Xin, Dong Dan, Xu Shun, et al. A k-Core based algorithm for influence maximization in social networks. Chinese Journal of Computers, 2015, 38(2): 238-248(in Chinese)
(曹玖新,董丹,徐顺等. 一种基于 k-核的社会网络影响最大化算法. 计算机学报, 2015, 38(2): 238-248)

[18] Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data. SIAM Review, 2009, 51(4): 661-703

[19] Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: Model and implications//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 6-14

[20] Wang Y, Shen H W, Liu S, et al. Learning user-specific latent influence and susceptibility from information cascades//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015

[21] Kramer A D I, Guillory J E, Hancock J T. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 2014, 111(24): 8788-8790

[22] Zafarani R, Cole W D, Liu H. Sentiment propagation in social networks: A case study in livejournal//Proceedings of the Advances in Social Computing. Berlin Heidelberg: Springer, 2010: 413-420

[23] Bae Y, Lee H. Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the American Society for Information Science and Technology, 2012, 63(12): 2521-2535

[24] Guille A, Hacid H, Favre C, et al. Information diffusion in online social networks: A survey. ACM SIGMOD Record, 2013, 42(2): 17-28

[25] Goyal A, Bonchi F, Lakshmanan L V S. A data-based approach to social influence maximization. Proceedings of the VLDB Endowment, 2011, 5(1): 73-84

[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111-3119

[27] Lin C J. Projected gradient methods for nonnegative matrix factorization. Neural Computation, 2007, 19(10): 2756-2779

[28] Zeiler M D. ADADELTA: An adaptive learning rate method. arXiv preprint arXiv: 1212.5701, 2012

[29] Voorhees E M. The TREC-8 question answering track report//Proceedings of the 8th Text REtrieval Conference. Gaithersburg, USA, 1999, 99: 77-82

[30] Fawcett T. An introduction to roc analysis. Pattern Recognition Letters, 2006, 27(8): 861-874

LIAO Xiang-Wen, born in 1980, Ph. D., associate professor. His research interests include opinion mining and sentiment analysis.

ZHENG Hou-Dong, born in 1990, M. S. candidate. His research interests include opinion mining and sentiment analysis.

LIU Sheng-Hua, born in 1982, Ph. D., associate professor. His research interests include data mining, social network, and sentiment analysis.

SHEN Hua-Wei, born in 1982, Ph. D., associate professor. His main research interests include social network analysis, network information dissemination, data mining and machine learning.

CHENG Xue-Qi, born in 1971, Ph. D., professor. His research interests include big data analysis and mining, network science, network and information security, web search and data mining.

CHEN Guo-Long, born in 1965, Ph. D., professor. His research interest is intelligent information processing.

Background

Online social network provides possibilities for information sharing and propagation from peer to peer, resulting in temporal sequences of happening times when users disseminate messages. And the temporal sequences forms cascades through the diffusion network, reflecting interpersonal influences. In turn, interpersonal influences can be modeled and analyzed from the observed temporal sequences of cascades in history. Personal influence analysis as user profiling is fundamental to influence maximization, social recommendation and viral marketing. Moreover, sentiment propagation is an important part of information diffusion on social network. Users can not only express their sentiments by publishing posts, but also communicate with each other in the community. Thus, sentiment is also an important factor to depict user influence.

Interpersonal influence is usually defined as propagation probability between users. Most existing works intuitively model the interpersonal influence in a pair-wise manner with n^2 independent variables to learn, assuming that propagation probability between different pairs of users is independent of each other, even if there exists one common user among different pairs. However, these methods require too many parameters and may suffer from overfitting problem. Moreover, there are seldom methods for estimating sentimental influence between pairs of users. Thus this paper proposes to model the interpersonal influence with two low-dimensional user-specific matrices, capturing their influence and susceptibility on different sentimental polarities respectively.

This work is supported by the National Key Technology R&D Program (Nos. 2013CB329606 and 2013CB329602), the National Natural Science Foundation of China (Nos. 61300105 and 61572467), the Key Laboratory of Network Data Science & Technology, Chinese Science and Technology Foundation (No. CASNDST20140X).