基于时空特征的社交网络情绪传播分析与预测模型

熊 熙[1]　乔少杰[1]　吴 涛[2]　吴 越[3]
韩 楠[4]　张海清[5]

摘　要　社交网络用户情绪传播与用户的空间距离和时间跨度有关, 并且受到多种交互机制的影响. 从大规模社交网络数据中提取情绪传播的时空特征, 研究用户行为对情绪传播的影响, 对预测情绪传播趋势具有实际意义. 利用线性回归获取的各行为子层的情绪传输率之间存在差异. 提出一种基于多层社交网络的情绪传播模型, 被称为 ECM 模型 (Emotional contagion model). 该模型包括三个行为子层, 每层的拓扑结构各不相同, 由该行为的交互历史决定. 在真实数据上对 ECM 模型进行仿真分析, 可以获得社交网络中情绪传播的过程与规律: 1) 中性情绪用户所占比例随时间逐渐增大, 接近 57.1 %, 而正向情绪与负向情绪比例始终接近. 2) 情绪传输率越大, 用户情绪更容易受到其他用户的影响而发生变化; 初始情绪越中立的用户, 在演化过程中情绪波动越小, 而初始情绪极性越大的用户情绪波动越大. 此外, 通过实验对比该模型与其他情绪传播模型, 表明 ECM 模型更加接近真实数据, 对社交网络中情绪传播具有较好的预测效果, 预测准确率相比其他模型可以提高 1.8 %～7.8 %.

关键词　情绪传播, 多层网络, 行为分析, 社交网络

引用格式　熊熙, 乔少杰, 吴涛, 吴越, 韩楠, 张海清. 基于时空特征的社交网络情绪传播分析与预测模型. 自动化学报, 2018, **44**(12): 2290−2299

DOI　10.16383/j.aas.2018.c170480

Spatio-temporal Feature Based Emotional Contagion Analysis and Prediction Model for Online Social Networks

XIONG Xi[1]　QIAO Shao-Jie[1]　WU Tao[2]　WU Yue[3]
HAN Nan[4]　ZHANG Hai-Qing[5]

Abstract　Users′ emotion in social networks is related to spatial distance and time span, and affected by multiple interaction mechanisms. It has practical significance to extract the spatio-temporal features from large-scale social networks and study the influence of users′ behaviors on emotional contagion in order to predict the trend of emotional contagion. The transmisibility values on different behavioral layers are calculated by linear regression and the results show the differences between these values. An emotional contagion model called ECM on multilayer social networks is proposed. It consists of three behavioral layers with different topologies depending on users′ interaction history. By simulation on real dataset, it is discovered that, 1) the proportion of users with neutral emotion is gradually increased with time and reaches 57.1 % while the proportion of positive emotion is comparable to that of negative emotion from beginning to end; 2) users' emotion is more likely to be influenced by other users when transmissibility becomes larger and users with initial polar emotion fluctuate more drastically than users with initial neutral emotion. In order to show the advantages of the proposed model, it is compared with other emotional contagion models. The results demonstrate that the proposed model approximates to the real data of emotional contagion on social networks, and also shows better predictive performance of emotional contagion. The prediction accuracy is increased by 1.8 %～7.8 %.

Key words　Emotion contagion, multilayer networks, behavior analysis, social networks

Citation　Xiong Xi, Qiao Shao-Jie, Wu Tao, Wu Yue, Han Nan, Zhang Hai-Qing. Spatio-temporal feature based emotional contagion analysis and prediction model for online social networks. *Acta Automatica Sinica*, 2018, **44**(12): 2290−2299

情绪是一种复杂的心理体验. 个体可以通过模仿其他个体的肢体动作或面部表情来传播情绪[1], 同时情绪会受到各种非语言因素的影响. 对情绪的研究引起了多学科的广泛关注, 包括经济学、神经科学和心理学. 众多研究表明人们会受到其他人的情绪影响, 并且这种影响的持续时间或长或短[2]. 陌生人之间的短暂接触也能传播情绪, 例如服务员的 "微笑服务" 可以提升顾客满意度进而为自己带来小费[3]. 社交网络特别强调用户创造内容, 用户不但是信息接受者, 同时也是信息的制造者、发布者和传播者, 成为网络舆论形式中不可分割的一部分. 在线社交网络也成为人们交流信息与情绪的主要平台. 下面以一个直观的例子说明研究社交网络中情绪传播的重要性. 2015 年, 亚马逊网站创始人杰夫 · 贝佐斯 (Jeff Bezos) 曾在 Twitter 发布一条推文, 宣称自己刚刚实现了运载火箭的软着陆. 该条消息以极快的速度在网络上转发和扩散, 并且其关注者表现出极大的喜悦, 在 Twitter 上展开了热烈讨论. 于此同时, 嫉妒和抑郁的情绪在 SpaceX 公司 CEO 埃隆 · 马斯克 (Elon Musk) 的关注者中迅速蔓延. 随后马斯克发布推文表示三年前他的火箭已经完成了六次亚轨道飞行. 该条消息迅速为其关注者带来了积极的情绪. 从这个例子可以看出, 社交网络可以通过用户交互行为使情绪迅速扩散, 并充分放大个体的情绪影响力.

本文对多层社交网络中情绪传播的研究主要基于如下几点考虑: 1) 因为社交网络用户情绪与用户的空间距离和时间

收稿日期 2017-08-31　录用日期 2018-01-01
Manuscript received August 31, 2017; accepted Janaury 1, 2018
国家自然科学基金 (61772091, 61802035), 教育部人文社会科学研究青年基金 (17YJCZH202), 四川省科技计划项目 (2018GZ0253, 2018JY0448), 成都信息工程大学科研基金 (KYTZ201637, KYTZ201715, KYTZ201750), 成都市软科学研究项目 (2017-RK00-00125-ZF, 2017-RK00-00053-ZF), 成都信息工程大学中青年学术带头人科研基金 (J201701), 四川高校科研创新团队建设计划 (18TD0027), 广西自然科学基金项目 (2018GXNSFDA138005), 广东省重点实验室项目 (2017B030314073) 资助
Supported by National Natural Science Foundation of China (61772091, 61802035), Youth Foundation for Humanities and Social Sciences of Ministry of Education of China (17YJCZH202), Sichuan Science and Technology Program (2018GZ0253, 2018JY0448), Scientific Research Foundation for Advanced Talents of Chengdu University of Information Technology (KYTZ201637, KYTZ201715, KYTZ201750), Soft Science Foundation of Chengdu (2017-RK00-00125-ZF, 2017-RK00-00053-ZF), Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology (J201701), Innovative Research Team Construction Plan in Universities of Sichuan Province (18TD0027), Natural Science Foundation of Guangxi (2018GXNSFDA138005) and Guangdong Key Laboratory Project (2017B030314073)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 成都信息工程大学网络空间安全学院 成都 610225　2. 重庆邮电大学网络空间安全与信息法学院 重庆 400065　3. 西华大学计算机与软件工程学院 成都 610039　4. 成都信息工程大学管理学院 成都 610103　5. 成都信息工程大学软件工程学院 成都 610225
1. School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225　2. School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065　3. School of Computer and Software Engineering, Xihua University, Chengdu 610039　4. School of Management, Chengdu University of Information Technology, Chengdu 610103　5. School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225

跨度有关, 所以需要从大规模网络数据中提取时空特征, 进而预测情绪传播趋势; 2) 社交网络为用户提供了多种交互机制, 使信息和情绪的传播更加便捷, 同时也对情绪传播产生了多维度的影响, 因此有必要研究不同用户交互行为对情绪传播的影响. 3) 利用多层网络分析社交网络的结构和动力学特性, 可以突破传统单层网络分析的局限性. 多层网络的出现实质是为了突破传统单层网络中连边同质性的限制, 各层有不同的拓扑结构并且每层的节点之间不一定有对应关系.

社交网络的结构和动力学特性比随机网络、小世界网络和无标度网络等典型网络更加复杂, 而且多种用户行为对情绪传播会产生重要影响. 在此基础上, 本文的研究主要实现以下目标:

1) 在考虑多种用户行为等复杂要素的基础上构建一种社交网络中的情绪传播模型.

2) 利用该模型研究社交网络中情绪传播规律, 并预测其传播趋势.

本文主要贡献包括:

1) 提出一种基于社交网络多种交互行为的情绪传播模型, 被称为 ECM 模型 (Emotional contagion model). 利用该模型可以分析社交网络中情绪传播的过程与规律. 研究发现: 多层社交网络中中性情绪用户所占比例随时间逐渐增大, 并且正向情绪与负向情绪比例始终接近. 情绪传输率越大, 用户情绪更容易受到其他用户的影响而发生变化. 初始情绪越中立的用户, 在演化过程中情绪波动越小, 而初始情绪极性越大的用户情绪波动越大.

2) 通过实验对比了本文所提模型与其他情绪传播模型, 包括: 基于情绪的 Spreader-ignorant-stifler (ESIS) 模型[4] 和独立级联模型[5], 实验结果表明 ECM 模型对社交网络中情绪传播具有较好的预测效果.

1 国内外研究现状

情绪可以看作是由许多的关键成分所组成的复杂心理现象, 通常包括主观情绪体验、面部表情以及躯体行为等, 同时可以利用 “效价–唤醒度” 的划分方法[6] 将情绪分为不同类型: 依据效价 (Valence) 将情绪分为正、负两极, 位于正极的称积极情绪, 通常带来愉悦感受, 如快乐、爱、愉快、幸福等; 位于负极的称消极情绪, 通常产生不愉悦感受, 如忧愁、悲伤、愤怒、紧张、焦虑、痛苦、恐惧、憎恨等; 同时依据唤醒度 (Arousal) 区分情绪的强弱, 唤醒度越大, 所产生的情绪就越强烈.

不同类型情绪的传播各有特点, 利用弗雷明汉心脏研究 (Framingham heart study, FHS)[7] 的参与者数据可以分别研究高兴、抑郁和孤独等多种情绪在社交网络中的传播过程[8], 进而通过广义估计公式分析好友间情绪的关联度, 最终发现各种情绪都会在社交网络中传播, 并且都能产生长时间的影响. Coviello 等[8−9] 研究了在线交互行为对传播用户情绪的作用, 以阴雨天气为例, 发现下雨不仅可以直接造成人们的情绪低落, 还可以通过社交网络影响另一个天气晴朗的城市的用户情绪. 上述研究主要针对消息的内在特征, 但未考虑用户多种行为对情绪传播的影响.

信息传播为情绪传播提供了必要的条件. 现有的信息传播模型可以分为两类: 图模型和传染病模型[10]. 图模型以网络结构为基础, 主要包括独立级联模型 (Independent cascade model, IC model)[5] 和线性阈值模型 (Linear threshold model, LT model)[11], 其中独立级联模型中的用户以一定概率在节点间传递信息, 线性阈值模型的每个节点受到邻点的影响力超过自身阈值就会被激活. 传染病模型主要通过模拟传染病的传播过程来对信息传播过程建模, 其中常见的传染病模型包括 SIR (Susceptible-infected-recovered) 模型[12] 和 SIS (Spreader-ignorant-stifler) 模型[13] 等. 这些模型将用户分为几类, 各类型用户在某些条件下可以相互转化. 近年来, 一些不同场景下的信息传播模型陆续被提出. Xiong 等[14] 提出一种信息扩散模型, 该模型在 SIR 模型的基础上增加了一种保留状态, 用于表示用户收到信息但未做出决策的状态. Wang 等[4] 提出了基于情绪的 SIS (ESIS) 模型, 将情绪划分为若干细粒度类型, 边权值等于用户间带有某种情绪的消息的转发数, 而接收消息的概率由传播概率和转发强度共同决定. 虽然上述模型总结了信息和情绪传播过程中的部分特征, 但是却忽略了情绪传播的多维度时空特性.

Boccaletti 等[15] 将多层网络视为类似于一个由多个单层网络组成的网络集, 每个单层网络构成一个网络层, 以 (G, C) 表示整个多层网络, 其中, G 是由一组单层网络组成的集合, C 是包含所有不同层间连边的集合, 进而形成网络层内的邻接矩阵和网络层间的邻接矩阵. Kivela[16] 进一步考虑多层网络中同一层的网络节点之间存在多重类型连边的情况, 即同一层中网络又可进一步分为 “亚层”, 提出用张量分析的形式来表示这类多层网络整体的邻接矩阵. 社交网络多种交互机制所构成的多层网络结构具有其特殊性, 例如转发关系网是关注关系网的子网, 上述抽象的多层网络分析方法无法获得满意的结论.

社交网络的多层结构使信息和情绪可以同时在多个拓扑结构中传播, 增加了研究的复杂性. Yagan 等[17−18] 研究了在线和真实社会网络中信息的传播规律, 通过数学解析与模拟仿真的方法, 发现获得信息的用户比例存在阈值, 当该比例大于阈值时, 信息将会大范围传播, 并且倾向于在同一个社区中传播. Kim 等[19] 研究了信息跨多个异质社交网络的扩散动力学. 网络用户通过 RSS 订阅器或社交网络聚合器等工具, 跨平台浏览各种类型的新闻, 使不同社交媒体发生耦合. 上述研究的不足在于: 跨平台采集数据具有较大难度, 即使利用社交网络聚合器等工具取得数据, 仍然难以将同一个用户在不同平台中的数据对应起来.

社交网络的用户情绪更多地受用户行为的影响, 例如 “转发” 和 “提及” 这两种动作会为情绪传播带来不同的影响: “提及” 对单个用户的影响力较大, 但影响范围不及 “转发”. 本文正是综合考虑不同用户行为对情绪传播的影响, 构建社交网络中的情绪传播模型来分析情绪传播的特征.

2 情绪传播模型工作原理

2.1 建模过程

如图 1 所示, 构建基于多层社交网络的情绪传播模型包括四个主要步骤:

1) 从在线社交网络 Twitter 和新浪微博中采集一段时间的用户信息及其行为关系信息, 以及在这段时间内发送的文本消息. 将这些数据进行预处理以供分析使用.

2) 用户的多种交互行为构成多层网络, 并且用户对其好友随后的信息会产生影响. 利用统计方法分析不同时间点和不同网络位置的用户情绪及其交互行为数据, 以提取情绪在空间和时间上的多维度传播特征.

3) 构建社交网络中的情绪传播模型, 其中包含若干行为子层. 每个子层根据该行为的交互历史形成不同拓扑结构, 并且每个子层中拥有不同的情绪传输率.

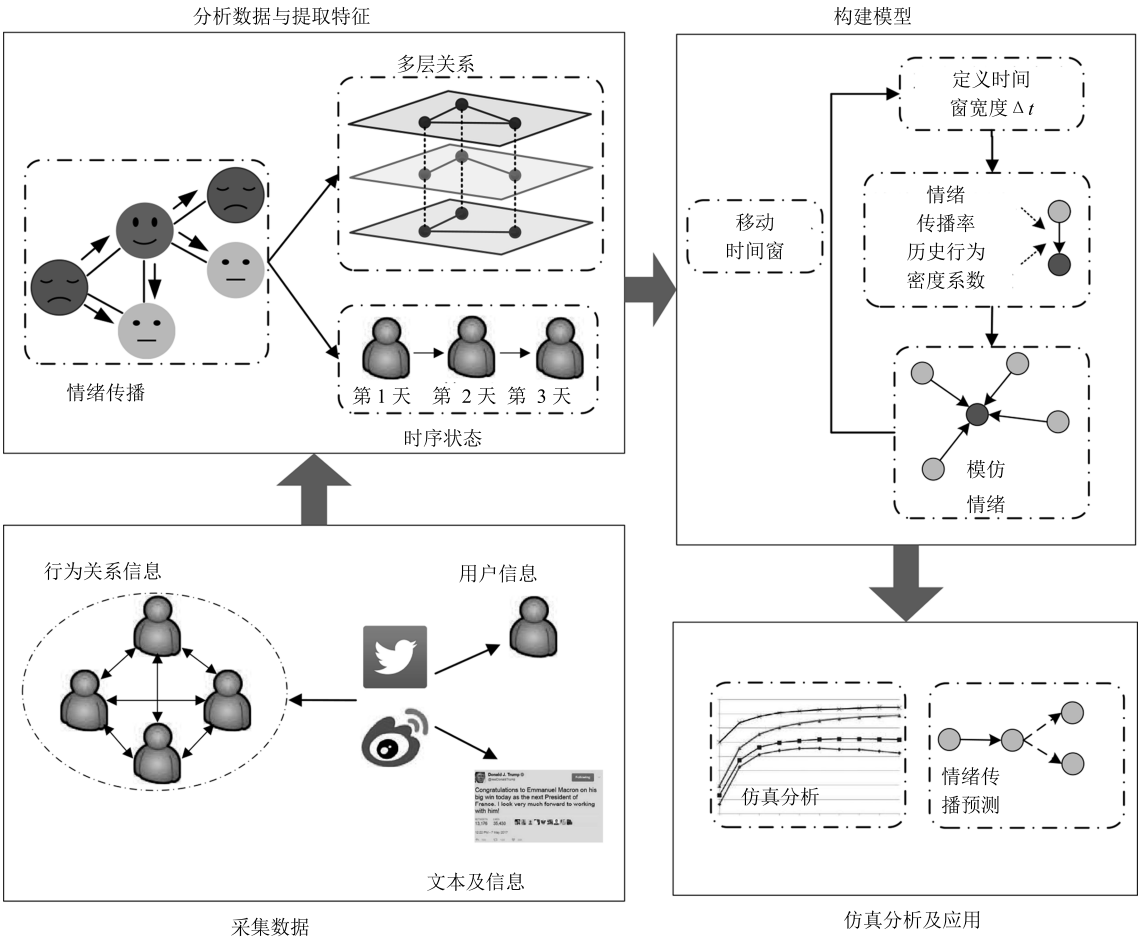


图 1 社交网络中情绪传播分析及模型构建示意图

Fig. 1 Analysis and modeling of emotion contagion in social networks

4) 基于采集的数据对该模型进行仿真实验, 分析情绪的传播规律, 并利用该模型预测情绪的传播趋势.

2.2 多种交互与多层网络的映射

利用文献 [20] 中提到的方法可以将用户不同交互机制形成的多层社交网络用 $G = \cup G_\alpha$ 来表示, 其中, α 表示不同子层. 四个子层分别为关注子层 ($\alpha = F$)、转发子层 ($\alpha = R$)、提及子层 ($\alpha = M$) 和回复子层 ($\alpha = S$). 每个子层中的用户都可以表示为节点. 这些子层存在以下特征:

1) 各子层内部的连边分别具有不同的含义: 关注子层的每条边表示两个用户间存在好友关系; 转发子层的每条边则表示用户转发了其他用户的消息; 提及子层的每条边表示用户在自己发布的消息中提到了其他用户, 该机制可以用于专门构建用户间的对话关系, 或者仅仅是为了提醒某人查看该消息[21], 从而使被提及用户阅读该消息的几率大大增加; 回复子层中每条边表示用户回复其他用户的消息.

2) 关注子层是其他子层的基础, 提供了信息和情绪传播的通道, 而其他每个子层的节点集合和连边集合都是关注子层相应集合的子集, 因而其他子层的节点分布比关注子层稀疏, 这表明用户只会主动挑选部分消息进行转发、提及或回复, 而不像查看消息那样是一个被动接受的过程. 其他交互行为都受到关注子层的非规则拓扑结构的影响.

3) 用户的关注行为在一段时间内相对稳定, 不容易发生变化, 因此用户在较短时间内 (1 小时至 10 天) 的交互只需要考虑转发、提及与回复这三种行为.

这三个行为子层中情绪传输效果存在较大差异, 因此采用情绪传输率[20] 来衡量一对用户间传播情绪的能力. 情绪传输率受到用户行为的影响, 即不同的行为子层拥有不同的情绪传输率.

3 基于多层社交网络的情绪传播模型

3.1 模型描述

社交网络用户间的不同交互机制构成了具有不同拓扑结构的用户关系网络, 这些网络之间相互依存并相互影响. 利用多层网络分析社交网络的结构和动力学特性, 可以突破传统单层网络分析的局限性, 多维度挖掘情绪传播的特征. 多层网络的出现实质是为了突破传统单层网络中连边同质性的限制, 各层有不同的拓扑结构并且每层的节点之间不一定有对应关系.

Kramer 通过发现社交网络用户可以影响其好友情绪, 并且影响距离最大为 3 (用户与其直接好友之间的距离为 1), 持续时间最多为 3 天[22]. 这一事实说明用户间情绪具有时间关联性和空间关联性. 同时, 社交网络中用户行为的多样性使情绪传播又具有特殊性. 为有效分析情绪传播规律, 并预测其传播趋势, 本文提出基于多层社交网络的情绪传播模型

(Emotional contagion model, ECM 模型).

为简化模型构建, 本模型基于以下假设: 为方便表示情绪的传播过程, 可以将连续时间轴划分为若干细小时间段, 其中每个时间段称为一个时步. 在一个时步中, 两个节点最多完成每种交互行为各一次, 并且该行为子层上的所有节点(用户) 依次更新情绪状态.

如果用 ρ 表示关注子层的节点密度, 它在整个模型演化过程中保持不变. α 表示某一个行为子层, 则该子层的节点密度 $\rho_\alpha < \rho$, 可以表示为 $\rho_\alpha = \rho\gamma_\alpha$, 其中, γ_α 称为密度系数, 由 $[t-\Delta t, t]$ 内该层中发生交互行为的用户分布决定.

α 子层这两个用户之间在时步 t 新出现连边的概率为 γ_α, 即 α 子层的密度系数. 假设 α 子层中用户 i 与 j 之间存在连边, 而用户 k 与 j 之间不存在连边, 则 i 和 k 分别对 j 采取 α 行为的概率为:

$$\begin{aligned}\frac{\mathrm{d}E_j(t)}{\mathrm{d}t} =& (1-\lambda)\sum_{i\in\delta_\alpha(j)} p_{ij}\Delta E_{ij}(t) + \\ & \lambda\sum_{k\in\delta_F(j)} p_{kj}\Delta E_{kj}(t)\end{aligned} \tag{1}$$

其中, $\Delta E_{ij}(t)$ 和 $\Delta E_{kj}(t)$ 表示节点 i 和 k 与节点 j 在时步 t 的情绪差, 即 $\Delta E_{ij}(t) = E_i(t) - E_j(t)$, $\Delta E_{kj}(t) = E_k(t) - E_j(t)$; $\delta_\alpha(j)$ 和 $\delta_F(j)$ 分别表示 j 在 α 层和关注子层的邻点集合; p_{kj} 表示 k 与 j 之间新产生连边的概率, 该值约等于 γ_α, 而 p_{ij} 则表示用户 i 与 j 之间在时步 t 将发生交互的概率, 可以表示为下面的公式:

$$p_{ij} = \varepsilon_\alpha h_{ij}^\alpha(t) \tag{2}$$

如果用户 i 对用户 j 在 $[t-\Delta t, t]$ 内采取了 α 行为, 则 h_{ij}^α 表示 α 子层中用户 i 和用户 j 之间的连边权重, 可以按以下公式计算:

$$\begin{aligned}h_{ij}^\alpha(t) =& \frac{\int_{t-\Delta t}^t k_i^\alpha r_{ij}^\alpha(\tau)\mathrm{d}\tau}{\int_{t-\Delta t}^t \left[\sum_i k_i^\alpha r_{ij}^\alpha(\tau)\right]\mathrm{d}\tau} = \\ & \frac{\int_{t-\Delta t}^t k_i^\alpha r_{ij}^\alpha(\tau)\mathrm{d}\tau}{\sum_i \int_{t-\Delta t}^t k_i^\alpha r_{ij}^\alpha(\tau)\mathrm{d}\tau} = \\ & \frac{q_{ij}^\alpha(t)}{\sum_i q_{ij}^\alpha(t)} \in [0,1]\end{aligned} \tag{3}$$

其中, k_i^α 表示节点 i 在 α 子层的度, 可以发现 $h_{ij}^\alpha(t)$ 满足 $\sum_i h_{ij}^\alpha(t) = 1$. 而 $r_i^\alpha(t)$ 可以表示为:

$$r_{ij}^\alpha(t) = \begin{cases} 0, & t \text{ 时步未产生行为 } \alpha \\ 1, & t \text{ 时步产生行为 } \alpha \end{cases} \tag{4}$$

在式 (3) 中, 分子与分母分别表示在时间区间 $[t-\Delta t, t]$ 内, j 与 i 之间以及 j 与其在 α 层所有邻点之间发生该行为的次数. 因此, $h_{ij}^\alpha(t)$ 是一个基于历史行为数据的时变参数, 随时间窗的移动而改变. 式 (1) 表示用户 j 模仿相邻用户的情绪, 即情绪从相邻用户向 j 扩散, 因此该式可以转换为:

$$\begin{aligned}\frac{\mathrm{d}E_j(t)}{\mathrm{d}t} =& (1-\lambda)\sum_{i\in\delta_\alpha(j)} \varepsilon_\alpha h_{ij}^\alpha(t)\Delta E_{ij}(t) + \\ & \lambda\sum_{k\in\delta_F(j)} \gamma_\alpha \Delta E_{kj}(t) = \\ & (1-\lambda)\varepsilon_\alpha \sum_{i\in\delta_\alpha(j)} h_{ij}^\alpha(t)\Delta E_{ij}(t) + \\ & \lambda\gamma_\alpha \sum_{k\in\delta_F(j)} \Delta E_{kj}(t) = \\ & (1-\lambda)\varepsilon_\alpha \langle\langle \Delta E_j(t)\rangle\rangle_\alpha + \\ & \lambda\gamma_\alpha \langle \Delta E_j(t)\rangle_F\end{aligned} \tag{5}$$

其中, $\langle\langle \Delta E_j(t)\rangle\rangle_\alpha$ 表示用户 j 在 α 子层与邻居情绪差的加权平均值, $\langle \Delta E_j(t)\rangle_F$ 表示用户 j 在关注子层与邻居情绪差的算术平均值. 将上式两侧分别在初始时步 t_0 到时步 t 上积分, 可以得到:

$$\begin{aligned}E_j(t) =& E_j(t_0) + (1-\lambda)\varepsilon_\alpha \times \\ & \sum_{\tau=t_0}^t \langle\langle \Delta E_j(\tau)\rangle\rangle_\alpha + \lambda\gamma_\alpha \sum_{\tau=t_0}^t \langle \Delta E_j(\tau)\rangle_F\end{aligned} \tag{6}$$

最后, 同时考虑 3 个行为子层, 可以得到:

$$\begin{aligned}E_j(t) =& E_j(t_0) + \\ & (1-\lambda)\sum_\alpha \varepsilon_\alpha \sum_{\tau=t_0}^t \langle\langle \Delta E_j(\tau)\rangle\rangle_\alpha + \\ & \lambda\sum_\alpha \gamma_\alpha \sum_{\tau=t_0}^t \langle \Delta E_j(\tau)\rangle_F\end{aligned} \tag{7}$$

式 (7) 表示用户 j 在时步 t 的情绪表达式, 其等于该用户与相邻用户情绪差异的时间累积和行为累积.

3.2　模型实现

本文提出一种基于多层社交网络的情绪传播模型–ECM 模型. 该模型包括三个行为子层, 并且每层的拓扑结构各不相同, 分别由用户的交互历史决定. 算法过程简单描述如下:

算法 1. 基于多层社交网络的情绪传播模型–ECM 模型

1) for each $t \in$ [1, 循环次数] do
2) for each $\alpha \in$ {三个行为子层} do
3) 计算行为 α 在 $[t-\Delta t, t]$ 内发生的次数;
4) end for
5) for each $\alpha \in$ {三个行为子层} do
6) for each $j \in$ {关注层的节点} do
7) 按式 (7) 更新 j 的情绪;
8) end for
9) end for
10) end for

算法共执行 sn 个时步 (第 1 行), 在每次循环结束时需要更新时步; 每个时步的处理过程可以分为两个部分, 分别用于计算 $[t-\Delta t, t]$ 的时间段中每种行为发生的次数 (第 2～4 行), 以及更新每个用户的情绪值 (第 5～9 行).

表 1 数据集统计信息

Table 1 The statistical information of the datasets

	Higgs 数据集	数据堂数据集	新采集 Twitter	新采集新浪微博
数据来源	Twitter	新浪微博	Twitter	新浪微博
用户 (节点) 数	456 626	63 641	33 070	6 344
好友关系数	14 855 842	1 391 718	185 393	54 093
转发次数	328 132	27 759	88 677	24 027
提及次数	150 818	未采集	41 245	10 428
回复次数	32 523	未采集	12 174	4 207
是否包含文本	否	是	是	是

3.3 算法时间复杂度分析

为了说明 ECM 模型具有较好的时间性能, 可用于预测情绪传播趋势, 需要分析 ECM 模型的时间复杂性. 每个时步的流程都分为两个部分, 第一部分用于计算每种行为的发生次数, 其时间复杂度为 $O(n^2)$, 第二部分用于更新用户情绪, 其时间复杂度也为 $O(n^2)$. 综合上述步骤获取整个 ECM 模型的时间复杂度为 $O(m \times n^2)$, 其中, m 和 n 分别表示时步数和用户总数.

4 实验与分析

4.1 数据集描述

Twitter 是一种基于互联网的社交网络, 在世界范围受到用户的广泛欢迎. 据统计, 2015 年 Twitter 的月均活跃用户量达到 2.71 亿, 成为传播信息和情绪的有力工具. 与此同时, 作为国内最大的微博网站, 新浪微博每天也有超过 1 亿条微博内容产生. 目前常用的社交网络数据集主要有以下两个:

1) 斯坦福大学 SNAP 实验室提供的 Higgs 网络数据集[23]. 欧洲核子研究组织 (CERN) 于 2012 年 7 月 4 日宣布发现 Higgs 玻色子, 该消息引起社交网络上的广泛议论. 该数据集包含 7 月 1 日 ∼7 月 4 日该消息在 Twitter 传播过程中的相关信息, 其中包括好友、转发、提及和回复这四种关系分别构成的网络, 以及每次行为发生的时间点. 由于该数据集不包括任何文本信息, 因此无法提取用户行为发生时的情绪状况, 需要人为指定被传播消息的情绪值.

2) 数据堂提供新浪微博数据集. 其中包含用户好友关系和他们对 12 个主题相关信息的转发关系, 但是未包含提及与回复这两种行为数据.

现有数据集具有一定局限性, 无法全面分析本文模型. 因此本文利用爬虫工具从 Twitter 和新浪微博网站重新采集了大量数据. 其中 Twitter 数据集包括 33 070 个用户及其关系信息, 以及 2016 年 3 月间 5 起热门话题的相关文本内容; 新采集的新浪微博数据集包括 6 344 个用户及其关系信息, 以及 2017 年 5 月间的 9 起热门事件的相关文本内容. 表 1 对比了本文新采集的数据集与现有数据集的主要统计信息.

4.2 文本情绪量化

本文采用情感分析工具 SentiStrength[24] 对情绪传播过程进行量化分析. 每条消息都同时包含正向情绪或负向情绪, 因此每条消息都被同时赋予一个正向情感值 $S^+(t)$ 与一个负向情感值 $S^-(t)$. 这两个值分别取 1 (中性) 到 5 (强正向和强负向) 之间的一个整数值. 为使用统一的度量方法来衡量消息文本的情绪, 可以将情绪极化值定义为正向情绪值和负向情绪值之和, 即极化值 $S(t)$ 取值范围为 $-4(S^+(t)=1, S^-(t)=5)$ 到 $+4$ $(S^+(t)=5, S^-(t)=1)$. 当正向和负向情绪值相同时 $(S^+(t)=S^-(t))$ 则为中性情绪 $(S(t)=0)$. 当情绪较弱时, 极化值接近 0, 可以近似看作中性情绪.

此外, 可以利用情绪极化值来定义情绪倾向. $S(t)$ 取值为 -4 到 -2 表示负向情绪倾向; $S(t)$ 取值为 -1 到 1 表示中性情绪倾向; $S(t)$ 取值为 2 到 4 则表示正向情绪倾向. 如果需要在时变模型中表示情绪值, 则可以使用连续情绪值, 即采用 θ_1 表示正向情绪和中性情绪的界线, θ_2 表示负向情绪和中性情绪的界线. 如果连续情绪极性值服从 $[-4, 4]$ 的均匀分布, 即 $a=-4$, $b=4$, 并且三种情绪取值区间宽度相同, 则有下面公式:

$$|b-\theta_1| = |\theta_1-\theta_2| = |\theta_2-a| \tag{8}$$

利用式 (8) 可以求得 $\theta_1 \approx 1.33, \theta_2 \approx -1.33$.

为对比不同两个数据集的文本情感, 本文仿照 SentiStrength 对新浪微博数据集的中文文本进行分词和情感分析, 主要 IKAnalyzer 分词工具[25] 和 BosonNLP 情感词典[26] 对新浪微博的文本进行情感标注.

每个用户通过三种行为影响其邻居的情绪. 通过对本文采集的数据进行分析, 可以发现一系列特征. 统计数据来自非连续三天的平均值并且每个时步定义为 2 个小时. 在每个时步中, 用户情绪的变化可以表示为各种行为出现频率以及不同行为情绪传输率的线性函数[20]. 利用线性回归方法分析两个数据集, 可以得到置信度为 95 % 时三个子层的情绪传输率. 如表 2 所示, 在两个数据集中, 提及子层的情绪传输率都最大, 表明该行为更利于情绪在网络中的扩散. 并且新浪微博中情绪传播更加迅速, 主要是由于新浪微博中公共消息更多, 更容易受到用户的关注并形成情绪聚集.

表 2 两个数据集不同子层的情绪传输率

Table 2 The transimisibilities on different layers in the two datasets

	Twitter 数据集	新浪微博数据集
转发	0.27	0.31
提及	0.95	1.07
回复	0.44	0.45

4.3 情绪传播特征分析

社交网络的结构和动力学特性比随机网络、小世界网络和无标度网络等典型网络更加复杂, 而且各种因素都会对社交网络中的情绪传播产生重要影响. 本小节将利用 ECM 模型分析社交网络中的情绪传播过程及其特征. 由于两个数据集中的实验结果近似, 因此本小节仅展示在 Twitter 数据集上的结果.

用户之间存在某些特殊关系, 例如亲戚、朋友或拥有相同的爱好. 用户通过这些现实世界的关系产生在线关注关系, 情绪也会因为这些关系而在网络中传播. 如图 2 所示, 可以发现: 具有某种情绪的用户在一段时间内发布的消息中都会带有该情绪倾向, 并且该情绪会影响该用户的直接或间接好友. 同时, 情绪传播过程具有明显的局部性, 例如用户一般只能影响距离在 3 以内的用户, 并且距离越近关联度越大, 而对距离大于 3 的用户几乎没有影响. 此外还可以从数据中发现抑郁、孤独和愤怒等负向情绪比愉快、兴奋等正向情绪更容易传播.

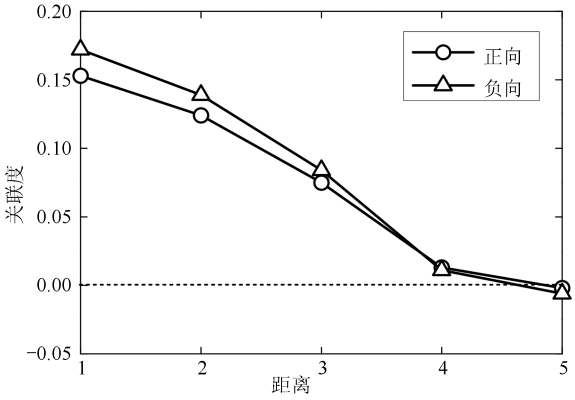


图 2　用户间情绪关联度与距离之间的关系图

Fig. 2　Relation between emotional correlation and distances

利用 ECM 模型可以定量展示社交网络用户情绪的动态传播过程. 如图 3 所示, 三种情绪具有相近的初始比例, 比例之差不超过 4.0 %. 初始阶段, 三种情绪同时在网络中传播, 中性情绪减少, 其他两种极化情绪增多. 这主要是因为网络的非均衡性会产生一些中心用户, 他们的极化情绪会对周围用户产生较大影响, 使他们也同样 “情绪化”, 中性情绪与极性情绪的比例差达到 10.1 %. 用户在与多个邻居的交互中获得了更多的信息, 极化情绪用户逐渐减少, 而中性情绪用户所占比例随时间逐渐增大, 并且正向情绪与负向情绪比例始终接近, 比例差最大仅为 2.5 %. 在演化趋于稳定时, 中性情绪处于主导地位, 约占 57.1 % 的比例, 同时存在一部分用户仍然持有极性情绪. 通过分析网络情绪分布, 可以发现这些极化用户之间形成了多个社区, 每个社区内部用户相互影响, 情绪趋同, 却不易随其他社区的情绪而改变.

为分析不同行为对情绪传播的影响, 需要研究情绪在单一行为子层的传播过程, 同时忽略其他子层的影响. 图 4 表示情绪转换数 (即情绪从一种倾向转换为另一种倾向的次数) 与参数的关系, 其中横坐标表示用户初始情绪与节点度的乘积的平均值. 不同的子层具有不同的情绪传输率, 其中提及行为的传输率最高, 而转发行为的传输率最小. 图 4 中三条曲线的关系表明情绪传输率越大, 用户情绪更容易受到其他用户的影响而发生变化. 对同一条曲线, 初始情绪越中立, 则用户情绪波动越小, 例如初始平均情绪值为 0 时, 则用户在演化过程中仅平均改变 2 次情绪倾向. 而初始情绪极性越大, 则用户情绪波动越大. 横坐标为 150 时, 平均每个用户约改变 24 次情绪倾向. 尤其是具有较大节点度的中心用户, 其极性情绪更能影响其他用户.

4.4　模型对比

为了展示 ECM 模型的预测效果, 可以将 ECM 模型、ESIS 模型和 IC 模型与真实数据进行对比实验. 鉴于这些模型之间略有差异, 因此需要对参数进行一定的调整, 使它们在同一基准上进行比较, 具体参数调节过程如下:

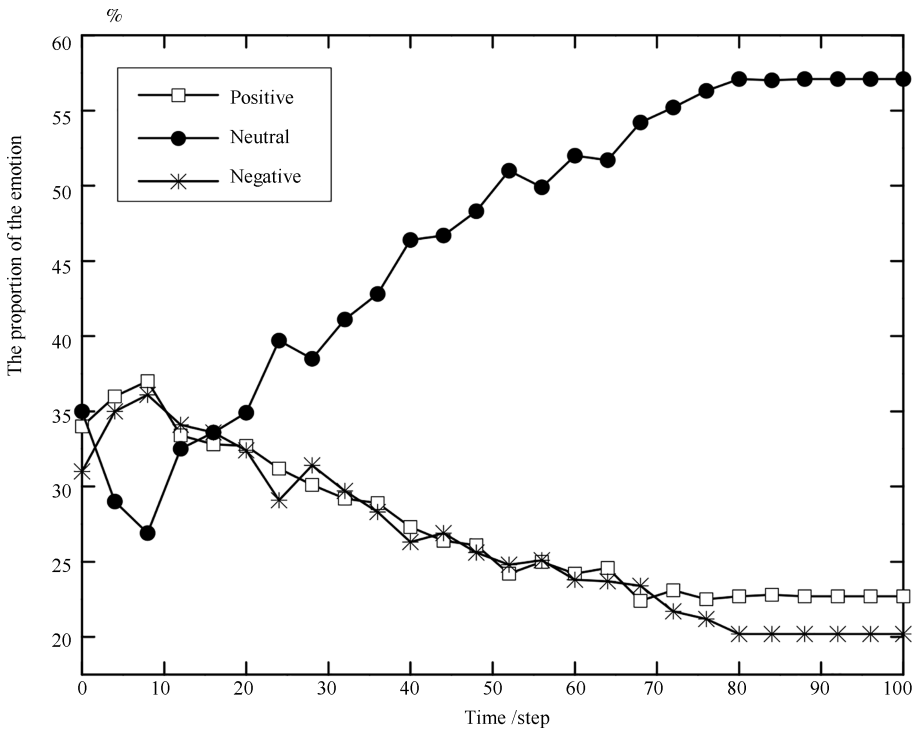


图 3　ECM 模型的演化规律

Fig. 3　Evolutionary process of ECM model

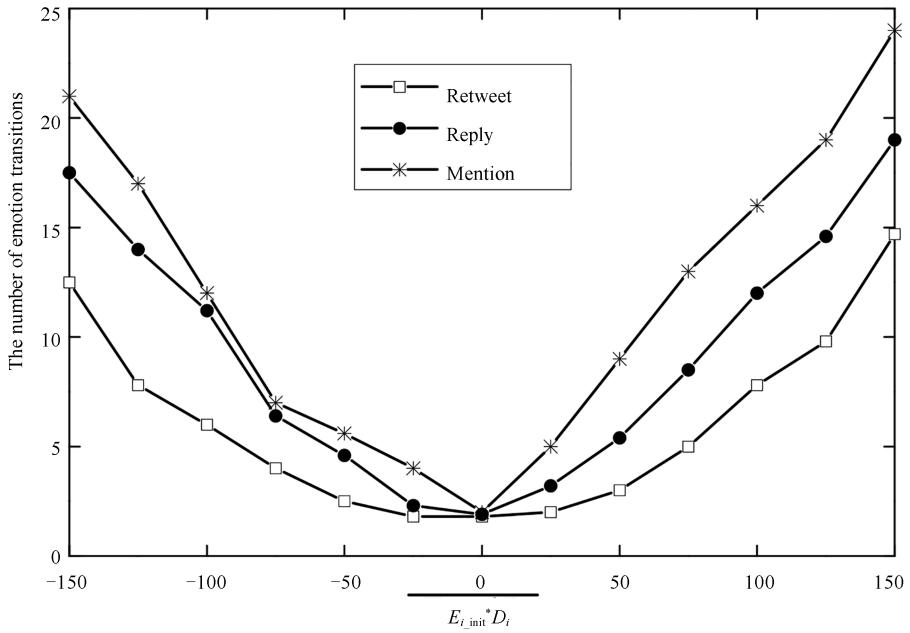


图 4 情绪转换数随用户初始情绪与节点度乘积的变化

Fig. 4 The relation between the number of individual emotional tendency changes, the degree and the initial emotion

1) ESIS 模型将情绪细分为六种. 首先根据某用户转发的含有该情绪的消息数来计算该用户的该种情绪值. 然后将这六种情绪归为正向、负向和中性三类: 高兴是正向, 惊讶是中性, 而愤怒、伤心、害怕和厌恶则是负向. 最后某用户在某时步内的情绪值即为他在该时步内所有消息的各情绪值之和.

2) 修改 IC 模型, 使边的权重表示用户间的影响力, 而不仅仅是表示获得信息的概率, 因此节点即使受到该情绪影响也不会停止演化. 另外, 该模型使用与 ECM 模型相同的情绪值计算方法: 当用户收到一个消息, 用户当前情绪值为该消息的情绪值与之前用户情绪值之和.

3) ESIS 模型和 IC 模型也被看作是多层模型, 只不过每层的拓扑结构相同.

4) 所有演化时步都被固定为 2 个小时.

通过对 ESIS 模型和 IC 模型的时间分析, 可以发现它们的时间复杂度均为 $\mathrm{O}(m \times n^2)$, 其中 m 和 n 分别表示时步数和节点数. 这与 ECM 一致, 表明三种模型拥有近似的执行时间. 此外, 图 5 展示了三种模型与真实数据在不同演化时步下的接近程度, 其中纵坐标 $\overline{E \times k}$ 表示节点情绪值与节点度的乘积平均值.

从图 5 可以看出, ESIS 模型比其他模型拥有更好的数据拟合性. IC 模型最简单, 而 ESIS 模型由 SIS 模型演化而来, 可用于解释信息传播的过程. 但是这两种模型偏离真实数据较多, 因为它们都只考虑了情绪本身的因素, 而未考虑多种网络行为对情绪传播的影响. 对比实验表明, ECM 模型与其他两种模型具有相同的时间复杂度, 但是与真实数据的拟合度更好. 此外, 图 5 中几种曲线都具有类似的走向, 先是快速上升, 然后缓慢下降. 这是因为热门事件通常可以在短时间内激起人们的广泛关注并出现极化情绪, 随着时间的推移, 人们的情绪会慢慢趋于理性和稳定.

分类算法的分类效果可以通过混淆矩阵中的准确率 (Precision)、查全率 (Recall) 和 F 值 (F-measure) 等三个指标[27] 来衡量. 本文将情绪传播中正向、中性和负向三种情绪分别归属到两个分类: 正向情绪为一类, 中性和负向情绪为一类, 则两个分类之间的界线就是 θ_1. θ_1 为典型值 1.3 时三种模型的分类效果如图 6 所示, 显然 ECM 模型拥有更好的分类准确率. 新浪微博的公众信息较多, 用户易受到中心用户的影响, 不易随着其他个人用户情绪而发生改变, 因此分类准确率较高; 而 Twitter 的用户通常关注了较多的个人好友, 其情绪也容易受到这些好友的影响, 导致分类准确率降低.

三种模型中 F-1 值随参数 θ_1 的变化曲线如图 7 所示, 可以看出 ECM 模型的 F-1 值比其他两种模型提高了 2.7 %～7.8 %, 说明其拥有更高的分类准确率. 三种模型的 F-1 值在 $\theta_1 = 1.5$ 附近达到最大值, 这是因为在情绪值均匀分布的条件下, 此时三种情绪都拥有近似的用户数量. ECM 模型的曲线波动较大, 并且与其他两种模型的 F-1 之差也在 $\theta_1 = 1.5$ 附近达到最大, 说明 ECM 模型对参数 θ_1 最为敏感. 随着 θ_1 的增大或减小, 情绪分布都会发生变化, 从而导致情绪预测准确率的下降.

三种模型中 F-1 值随用户数量的变化曲线如图 8 所示. 可以看出, ECM 模型将分类准确率提高了 1.8 %～6.2 %. 三种模型的 F-1 值都会随用户数的增大而增大, 这是因为大规模的训练集将会提高分类准确率. ECM 模型描述情绪传播特征更加充分, 因此 F-1 值上升更加迅速, 在用户数为 1 900 时达到最大值 70.5 %.

5 结束语

本文提出一种基于社交网络多种交互行为的情绪传播模型, 利用该模型分析社交网络中情绪传播的过程与规律. 在集的社交网络数据基础上进行仿真分析, 发现中性情绪用户所占比例随时间逐渐增大, 并且正向情绪与负向情绪比例始终接近. 情绪传输率越大, 用户情绪更容易受到其他用户的影响而发生变化. 初始情绪越中立的用户, 在演化过程中情

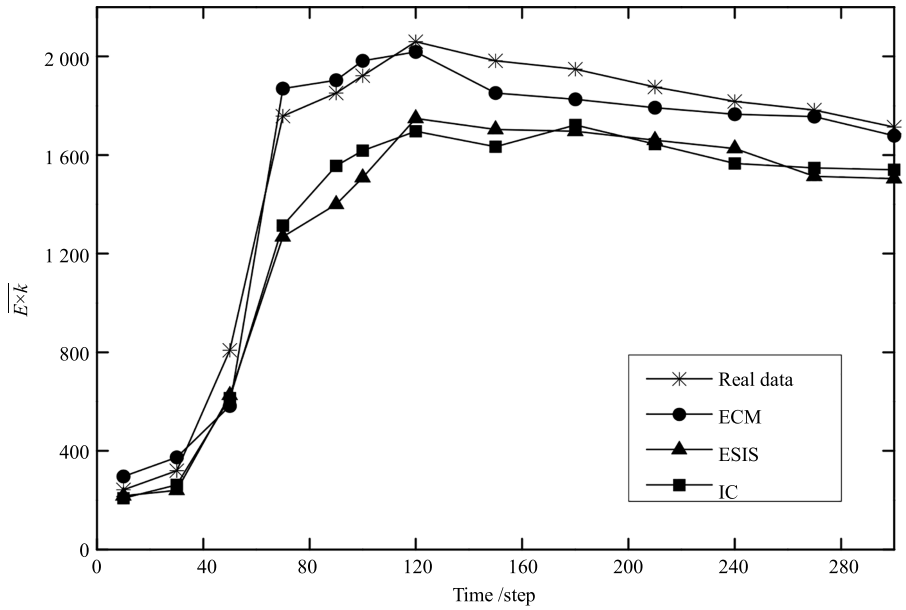
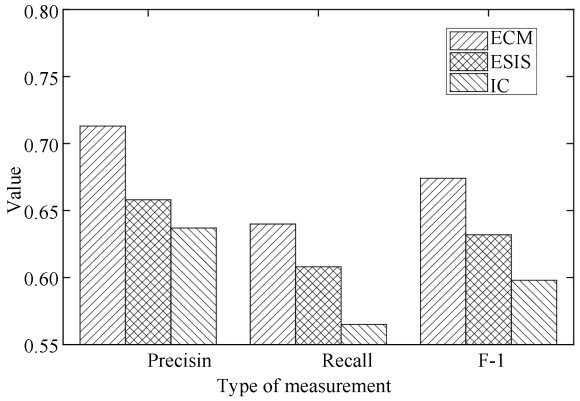


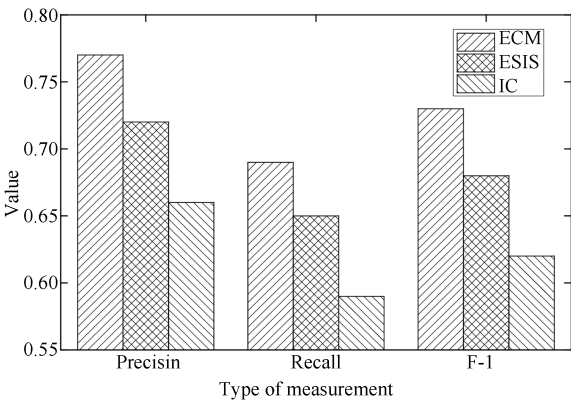
图 5　三种模型与真实数据的对比 (Twitter 数据集)

Fig. 5　The comparison of the three models and the real data (Twitter dataset)



(a) Twitter 数据集

(a) Twitter dataset



(b) 新浪微博数据集

(b) Weibo dataset

图 6　三种模型分类度量值的对比

Fig. 6　The comparison of classification measurements of the three models

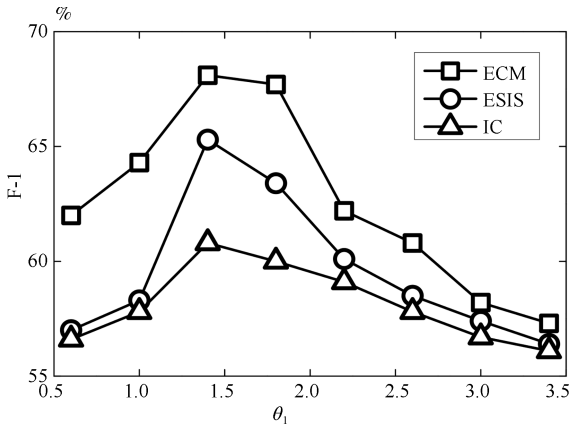


图 7　三种模型中的 F-1 值随 θ_1 的变化规律 (Twitter 数据集)

Fig. 7　F-1 changes with θ_1 for the three models (Twitter dataset)

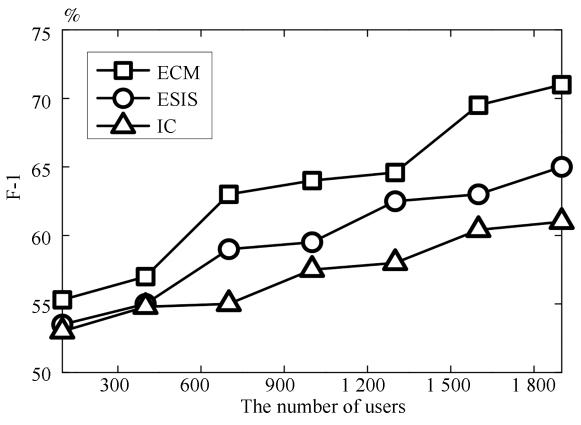


图 8　三种模型中的 F-1 值随用户数的变化规律 (Twitter 数据集)

Fig. 8　F-1 changes with the number of users for the three models (Twitter dataset)

绪波动越小, 而初始情绪极性越大的用户情绪波动越大. 最后, 本文还对比了该模型与其他情绪传播模型, 如: 基于情绪的 SIS 模型和独立级联模型, 实验表明 ECM 模型对社交网络中情绪传播具有较好的预测效果, 预测准确率比其他两种模型提高 1.8 %∼7.8 %.

本文工作仍然存在一些需要改进的地方, 例如:

1) 社交网络中的情绪传播是一个复杂的过程, 目前很难考虑所有网络因素的影响, 例如网络结构的动态演化. 也就是说, 用户倾向与拥有相近情绪的用户建立新的连接, 而与相反情绪的用户断开连接. 分析多种因素对情绪传播的影响将是未来一项有价值的工作.

2) 本文的工作基于情绪分析算法, 并采用了 SentiStrength 等工具和手段对消息文本进行分析. 虽然比之前的分析方法准确, 但仍然无法解析人类语言表达中的微妙情绪, 例如挖苦和嘲讽, 也无法很好地识别一句话中的多种情绪. 情绪的这些特点都给其识别带来了困难, 需要在未来进行深入研究.

References

1 Gabriel A S, Cheshin A, Moran C M, van Kleef G A. Enhancing emotional performance and customer service through human resources practices: a systems perspective. *Human Resource Management Review*, 2016, **26**(1): 14−24

2 Xiong X B, Zhou G, Huang Y Z, Chen H Y, Xu K. Dynamic evolution of collective emotions in social networks: a case study of Sina weibo. *Science China Information Sciences*, 2013, **56**(7): 1−18

3 Lo S C, Huang K P. The smiling mask in service encounters: the impact of surface and deep acting. *International Journal of Management, Economics and Social Sciences*, 2017, **6**(1): 40−55

4 Wang Q Y, Lin Z, Jin Y H, Cheng S D, Yang T. ESIS: emotion-based spreader-ignorant-stifler model for information diffusion. *Knowledge-Based Systems*, 2015, **81**: 46−55

5 Wang Q Y, Jin Y H, Yang T, Cheng S D. An emotion-based independent cascade model for sentiment spreading. *Knowledge-Based Systems*, 2017, **116**: 86−93

6 Dignath D, Janczyk M, Eder A B. Phasic valence and arousal do not influence post-conflict adjustments in the Simon task. *Acta Psychologica*, 2017, **174**: 31−39

7 Lloyd-Jones D M, Larson M G, Leip E P, Beiser A, D'Agostino R B, Kannel W B, et al. Lifetime risk for developing congestive heart failure: the Framingham Heart Study. *Circulation*, 2002, **106**(24): 3068−3072

8 Coviello L, Sohn Y, Kramer A D I, Marlow C, Franceschetti M, Christakis N A, et al. Detecting emotional contagion in massive social networks. *PLoS One*, 2014, **9**(3): Article No. e90315

9 Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, et al. Echo chambers: emotional contagion and group polarization on Facebook. *Scientific Reports*, 2016, **6**: Article No. 37825

10 Guille A, Hacid H, Favre C, Zighed D A. Information diffusion in online social networks: a survey. *ACM SIGMOD Record*, 2013, **42**(2): 17−28

11 Bozorgi A, Samet S, Kwisthout J, Wareham T. Community-based influence maximization in social networks under a competitive linear threshold model. *Knowledge-Based Systems*, 2017, **134**: 149−158

12 Liu L J. A delayed SIR model with general nonlinear incidence rate. *Advances in Difference Equations*, 2015, **2015**: 329

13 Zhang X H, Jiang D Q, Alsaedi A, Hayat T. Stationary distribution of stochastic SIS epidemic model with vaccination under regime switching. *Applied Mathematics Letters*, 2016, **59**: 87−93

14 Xiong F, Liu Y, Zhang Z J, Zhu J, Zhang Y. An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A*, 2012, **376**(30): 2103−2108

15 Boccaletti S, Bianconi G, Criado R, del Genio C I, Gómez-Gardeñes J, Romance M, et al. The structure and dynamics of multilayer networks. *Physics Reports*, 2014, **544**(1): 1−122

16 KivelaM, Arenas A, Barthelemy M, Gleeson J P, Moreno Y, Porter M A. Multilayer networks. *Journal of Complex Networks*, 2013, **2**(3): 261−268

17 Yağan O, Qian D J, Zhang J S, Cochran D. Information diffusion in overlaying social-physical networks. In: Proceedings of the 2012 46th Annual Conference on Information Sciences and Systems (CISS). Princeton, NJ, USA: IEEE, 2012. 1038−1048

18 Zhuang Y, Yagan O. Information propagation in clustered multilayer networks. *IEEE Transactions on Network Science and Engineering*, 2016, **3**(4): 211−224

19 Kim M, Newth D, Christen P. Modeling dynamics of diffusion across heterogeneous social networks: news diffusion in social media. *Entropy*, 2013, **15**(15): 4215−4242

20 Xiong X, Li Y Y, Qiao S J, Han N, Wu Y, Peng J, et al. An emotional contagion model for heterogeneous social media with multiple behaviors. *Physica A: Statistical Mechanics and its Applications*, 2018, **490**: 185−202

21 Ferrara E, Yang Z Y. Measuring emotional contagion in social media. *PLoS One*, 2015, **10**(11): Article No. e0142390

22 Kramer A D I. The spread of emotion via Facebook. in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, Texas, USA: ACM, 2012. 767−770

23 De Domenico M, Lima A, Mougel P, Musolesi M. The anatomy of a scientific rumor. *Scientific Reports*, 2013, **3**: Article No. 2980

24 Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology*, 2012, **63**(1): 163−173

25 Zhang Yun. The practice on the development of software on the Chinese academic bibliometrics based on the open source software. *New Technology of Library and Information Service*, 2010, **26**(4): 87−91
(张云. 基于开源软件的中文学术文献计量软件的开发实践. 现代图书情报技术, 2010, **26**(4): 87−91)

26 The Boson Data. Sentiment analysis [Online], available: http://docs.bosonnlp.com/sentiment.html, 2017.

27 Powers D M W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2011, **2**(1): 37−63

熊 熙 成都信息工程大学网络空间安全学院讲师. 2013 年获得四川大学信息安全专业博士学位. 主要研究方向为 web 挖掘, 社会计算, 机器学习.
E-mail: xiongxi@cuit.edu.cn
(**XIONG Xi** Lecturer at the School of Cybersecurity, Chengdu University of Information Technology. He received his Ph. D. degree in information security from Sichuan University in 2013. His research interest covers web mining, social computing, and machine learning.)

乔少杰 成都信息工程大学网络空间安全学院教授. 2009 年获得四川大学计算机学院工学博士学位. 主要研究方向为轨迹预测, 移动对象数据库, 大数据. 本文通信作者.
E-mail: sjqiao@cuit.edu.cn
(**QIAO Shao-Jie** Professor at the School of Cybersecurity, Chengdu University of Information Technology. He received his Ph. D. degree from the College of Computer Science, Sichuan University in 2009. His research interest covers trajectory prediction, moving objects databases, and big data. Corresponding author of this paper.)

吴 涛 重庆邮电大学网络空间安全与信息法学院讲师. 2017 年获得电子科技大学计算机科学与工程学院博士学位. 主要研究方向为数据挖掘.
E-mail: wutaoadeny@gmail.com
(**WU Tao** Lecturer at the School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications. He received his Ph. D. degree from the School of Computer Science and Technology, University of Electronic Science and Technology of China in 2017. His main research interest is data mining.)

吴 越 西华大学计算机与软件工程学院副教授. 2014 年获得四川大学信息安全专业博士学位. 主要研究方向为数据挖掘, 复杂网络.
E-mail: wuyue_xh@sina.com
(**WU Yue** Associate professor at the School of Computer and Software Engineering, Xihua University. She received her Ph. D. degree in information security from Sichuan University in 2014. Her research interest covers data mining and complex networks.)

韩 楠 成都信息工程大学管理学院讲师. 2012 年获得成都中医药大学博士学位. 主要研究方向为数据挖掘.
E-mail: hannan@cuit.edu.cn
(**HAN Nan** Lecturer at the School of Management, Chengdu University of Information Technology. She received her Ph. D. degree from Chengdu University of Traditional Chinese Medicine in 2012. Her main research interest is data mining.)

张海清 成都信息工程大学软件工程学院副研究员. 2015 年获得法国里昂第二大学博士学位. 主要研究方向为智能信息处理与知识工程.
E-mail: zhanghq@cuit.edu.cn
(**ZHANG Hai-Qing** Associate researcher at the School of Software Engineering, Chengdu University of Information Technology. She received his Ph. D. degree from Lumi ère University Lyon 2 in 2015. Her research interest covers intelligent information processing and knowledge engineering.)