

基于改进 HK 模型的社交用户观点演化预测研究

吴树芳¹ 郑依静¹ 朱杰^{2*}

(1. 河北大学管理学院, 河北 保定 071000; 2. 河北大学数学与信息科学学院, 河北 保定 071000)

摘要: [目的/意义] 准确预测舆情观点的演化态势, 对化解舆情危机具有重要意义。[方法/过程] 针对 HK 模型用于用户观点演化预测存在的不足, 进行如下改进: 首先, 从和当前事件相似度高于阈值的历史博文中获取用户的初始观点值; 其次, 利用用户间交互行为以及关注关系计算亲密度, 修订用户信任阈值; 最后, 融合用户全局影响力、用户间交互度以及观点接近度计算用户间的综合信任度, 突出用户间观点影响权重的差异性。最终利用改进的 HK 模型实现对社交用户观点演化预测。[结果/结论] 实验结果表明, 改进的 HK 模型用于观点演化预测具有较低 MSE 值和 MAE 值, 可为舆情发展提供有效预测。

关键词: HK 模型; 观点演化预测; 社交用户

DOI: 10.3969/j.issn.1008-0821.2024.02.012

[中图分类号] G203 [文献标识码] A [文章编号] 1008-0821 (2024) 02-0142-10

Research on Evolution Prediction of Social Users Opinion Based on the Improved HK Model

Wu Shufang¹ Zheng Yijing¹ Zhu Jie^{2*}

(1. School of Management, Hebei University, Baoding 071000, China;

2. College of Mathematics and Information Science, Hebei University, Baoding 071000, China)

Abstract: [Purpose/Significance] Accurate prediction of the evolution of public opinion is significance in resolving public opinion crises. [Method/Process] To solve the shortcomings of the existing HK model for predicting the evolution of social users' opinions, the following improvement were given: Firstly, the initial opinion values of users were obtained from historical blog posts with similarity to the current event above a threshold. Secondly, the user confidence threshold was revised by calculating the affinity using the interaction behavior between users and the following relationship. Then, the global influence of users, the interaction between users and the proximity of opinions were combined to calculate the comprehensive trust degree between users, highlighting the differences in the weight of opinion influence between users. Finally, the improved HK model was used to predict the evolution of social users' opinion. [Result/Conclusion] The experimental results show that the improved HK model for opinion evolution prediction has lower MSE and MAE values and can provide effective predictions for opinion development.

Key words: HK model; opinion evolution prediction; social users

由于社交网络具有用户门槛低、信息开放度高、传播容易等特点, 使得社交网络中的信息以“裂变”方式传播, 容易催生出恶性的网络舆情, 给社会安全稳定带来极大威胁。网络舆情形成的关键是用户观点

收稿日期: 2023-06-21

基金项目: 河北省人文社会科学研究重大课题攻关项目“基于大数据的河北省网络治理机制研究”(项目编号: ZD202102); 河北大学社会科学研究重大课题培育项目“网络生态驱动下谣言传播及阻断机制研究”(项目编号: 2021HPY004); 河北省自然科学基金项目“基于语义结构的自适应图卷积跨模态特征学习方法研究”(项目编号: F2022511001)。

作者简介: 吴树芳 (1979-), 女, 教授, 博士生导师, 研究方向: 信息处理、在线社交网络研究。郑依静 (1998-), 女, 硕士研究生, 研究方向: 在线社交网络研究。

通讯作者: 朱杰 (1982-), 男, 副教授, 研究方向: 人工智能。

的聚合,在观点聚合之前准确预测用户观点演化的趋势,可以在一定程度上控制舆情的恶意传播,具有较大的理论价值和应用价值,故该研究已引起国内外学者的广泛关注。社交用户观点演化预测研究主要基于观点动力学模型展开,已有研究主要包括两类:

一是基于离散型观点动力学模型的观点演化预测研究,离散型观点动力学模型中个体的观点值是离散的,一般利用-1(反对)、0(中立)、1(支持)表示观点的3种倾向,其中比较有代表性的有Ising模型^[1]、Sznajd模型^[2]、投票模型^[3]、多数决定模型^[4]。这些模型都基于“个体观点值离散”这一假设,但是在社交网络中由于用户间的多重交互性,并非严格离散,因此用离散值表示社交网络中的个体观点不够准确。

二是基于连续型观点动力学模型的观点演化预测研究,该类研究认为个体观点值是连续的,并设置个体观点值在区间波动,其中比较有代表性的连续型观点动力学模型主要包括DW(Deffuant-Weisbuch)模型^[5]和HK(Hegselmann-Krause)模型^[6],两者都基于有界信任的假设。如李根强等^[7]基于信息熵理论,借鉴有界信任模型中的HK模型,构建基于熵的观点演化模型并进行仿真实验,分析信任阈值、观点自信度和群体异质性对网络集群行为主体观点演化的影响;祁凯等^[8]基于现实背景扩展了DW模型,在是否存在智能推荐机制的信息推送机制的基础上,考虑平台影响力、话题类型等因素,运用仿真实验分析了不同参数对群体观点演化过程的影响;徐涵等^[9]受沉默的螺旋理论启发,引入“意见气候”这一参数,提出了一个新的基于意见气候的HK观点动力学模型,该模型以HK模型为基础,细化了观点更新主体,拓展了观点更新规则,增加了意见气候的影响;曹春萍等^[10]提出了一种基于用户推荐的HK模型(URHK),该模型不仅考虑了非邻居节点的影响,同时可以展示用户对某一话题的观点或态度,还改进了经典HK模型权重设置中所存在的缺陷。

上述研究中,离散型模型表征用户观点值的方式不够准确,而连续型模型使用连续值表示用户观点,应用性较强。在连续模型中,DW模型每个时刻选取两个个体进行观点交互,迭代过程时间复杂度较高。此外,在社交网络中,用户处于一定的社交群体,其观点的演化会受到群体内多个用户的影

响,如果选取两个个体进行观点交互则不符合实际情况,而HK模型则更多地考虑了周围用户对单个用户的影响。综上,本文将以HK模型为基础展开社交用户观点演化预测研究,提出基于改进HK模型的社交用户观点演化预测方法。

1 HK模型不足分析

HK模型模拟了用户对某一事件 E 的观点更新的动态过程,该模型假设,在一个用户数为 n 的群体中,用户集合表示为 $U=\{u_1, \dots, u_n\}$,初始时刻每个用户都对事件 E 持有一种观点,初始观点值被设置为 $[0,1]$ 间的随机分布。

在HK模型中,假设 t 时刻用户 $u_i(u_i \in U)$ 对某一事件 E 的观点值用实数 $S_{u_i}(t)(S_{u_i}(t) \in [0,1])$ 表示,观点值越接近1表明支持程度越强烈,越接近0表明反对程度越强烈。将群体中 n 个用户在 t 时刻的观点值表示为向量 $S(t)=(S_{u_1}(t), \dots, S_{u_n}(t))$,通过综合群体用户 U 的观点值即可预测某用户 u_i 在 $t+1$ 时刻对事件 E 的观点,预测过程如下:首先考虑 t 时刻与用户 u_i 的观点差值不超过信任阈值 ε 的用户,将这些用户定义为用户 u_i 在信任阈值内的用户集合 $I(u_i, S(t))$,如式(1)所示;然后将集合 $I(u_i, S(t))$ 中用户观点值加权求和预测用户 u_i 在 $t+1$ 时刻对事件 E 的观点值 $S_{u_i}(t+1)$,如式(2)和式(3)所示。

$$I(u_i, S(t)) = \{u_j | |S_{u_i}(t) - S_{u_j}(t)| \leq \varepsilon\} \quad (1)$$

$$S_{u_i}(t+1) = \sum_{u_j \in I(u_i, S(t))} \alpha_{ij} S_{u_j}(t) \quad (2)$$

$$\alpha_{ij} = |I(u_i, S(t))|^{-1} \quad (3)$$

其中, ε 是用户 u_i 的信任阈值, $|S_{u_i}(t) - S_{u_j}(t)|$ 表示用户 u_i 与 $u_j(u_j \in U)$ 的观点差值, u_j 表示满足观点差值小于信任阈值 ε 的用户。 α_{ij} 表示信任集合内用户 u_j 对 u_i 观点预测的影响权重,且 $\sum \alpha_{ij} = 1$ 。 $|I(u_i, S(t))|^{-1}$ 表示信任集合中用户数量的倒数。

通过分析,发现HK模型存在以下不足:

1) 在HK模型中,用户 u_i 在初始时刻的观点值 $S_{u_i}(0)$ 被设置为 $[0,1]$ 间的随机分布,但是受用户 u_i 历史数据的影响,其对某一事件 E 的初始观点可以根据历史上该用户对待相似事件的态度来反映,故 $S_{u_i}(0)$ 设置为随机值是不合理的。

2) HK模型在选择 u_i 的信任用户集合 $I(u_i, S(t))$ 时,仅考虑了用户间的观点差值,实际上社交用户 u_i 在选择信任集合时,除了受观点差值影响外,也受其他因素的影响,例如由于社交行为引

起的用户间亲密度等。

3) 通过分析 HK 模型的观点更新公式, 发现信任集合内用户 u_j 对 u_i 观点预测的影响权重 α_{ij} 为 $|I(u_i, S(t))|^{-1}$, 即所有信任用户对 u_i 的观点更新的影响权重均相等, 忽略了社交网络中用户间的差异性。

为有效解决上述问题, 本文通过计算用户的初始观点值、信任阈值和用户间观点影响权重, 提出改进的 HK 模型, 实现对用户观点演化的预测, 该研究可助力相关部门实现对舆情演化的精准预测, 并进行及时有效的引导和控制。

2 研究框架

基于上述分析, 论文为解决 HK 模型存在的不足, 提出了改进的 HK 模型, 并将其用于社交网络中用户观点的演化预测, 具体研究框架如图 1 所示。观察图 1 可知, 本文的工作主要包括以下 4 个方面:

1) 社交用户初始观点值的获取。获取社交用户集合 $U = \{u_1, \dots, u_n\}$ 中每个用户在事件 E 发生时刻前的历史博文集 $Blog_{h1}, \dots, Blog_{hn}$, 分别计算每个用户的历史博文集与当前事件 E 的内容相似度, 选取每个用户历史博文中相似度大于阈值 ξ 的历

史博文构成事件 E 相似博文集 $Blog_{s1}, \dots, Blog_{sn}$, 量化每个用户 u_i 的相似博文集 $Blog_{si}$ 中每条博文观点值, 将这些观点值加权融合量化用户 u_i 对事件 E 的初始观点值 $S_{u_i}(0)$ 。

2) 社交用户 u_i 的信任集合 $I(u_i, S(t))$ 的获取。本文将基于亲密度修订信任阈值并进行二次判定: 首先筛选出与用户 u_i 的观点差值小于信任阈值 ε 的用户, 并将其构成信任用户集合 $I_1(u_i, S(t))$; 然后利用亲密度 F_{ik} 修订 ε , 筛选出观点差值小于修订后信任阈值的用户, 并将其构成信任用户集合 $I_2(u_i, S(t))$ 。综合 $I_1(u_i, S(t))$ 和 $I_2(u_i, S(t))$ 获得用户的信任集合。

3) 社交用户 u_j 对 u_i 的观点影响权重 α_{ij} 的计算。分别计算用户 u_j 的全局影响度、用户 u_i 和用户 u_j 的交互度、用户 u_i 与 u_j 的观点接近度, 将三者加权融合获得用户 u_i 对 u_j 的综合信任度, 并将其作为待预测用户 u_i 在 $t+1$ 时刻观点的影响权重 α_{ij} 。

4) 社交用户 u_i 对事件 E 的观点预测。通过上述 3 部分操作可以预测社交用户 u_i 在 $t+1$ 时刻对事件 E 的观点值 $S_{u_i}(t+1)$ 。重复迭代上述流程, 即可完成社交用户对事件 E 不同时刻的观点预测, 实现观点演化的预测。

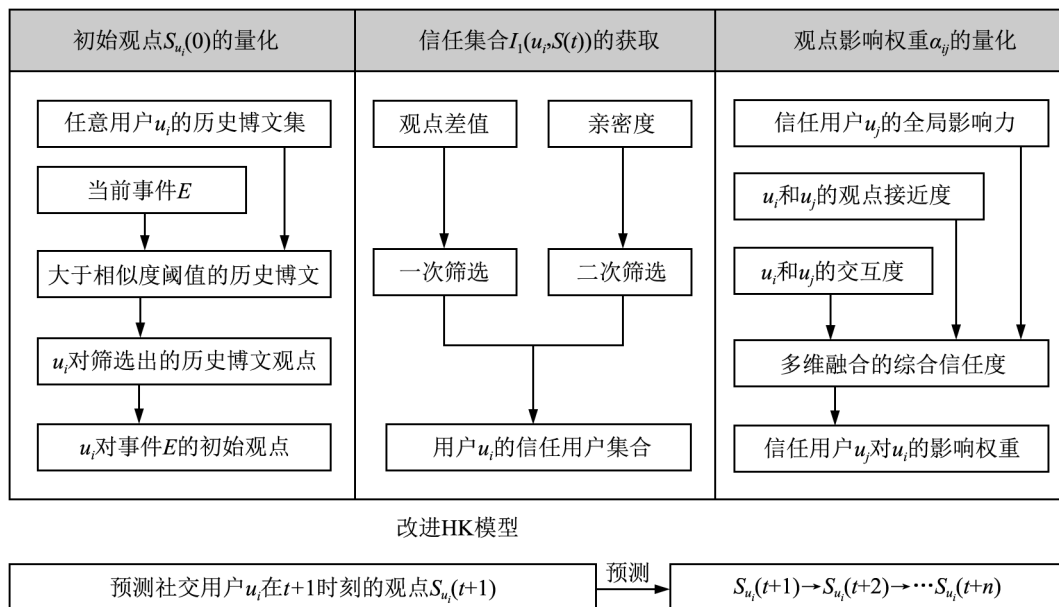


图 1 基于改进 HK 模型的社交用户观点演化预测研究框架

Fig. 1 Research Framework for Predicting the Evolution of Social User Opinions based on Improved HK Model

3 社交用户观点演化预测

本文通过修订社交用户的初始观点值、获取更为完备的信任用户集合、量化不同信任用户对观点预测的影响权重实现对 HK 模型的改进, 完成对社

交用户观点演化的预测。

3.1 社交用户初始观点值的获取

获取当前用户 u 在初始时刻 t_0 (事件 E 的发生时刻) 前的历史博文集 $Blog_{hi}$, 计算历史博文中

每条博文 d_{hi}^m 与当前事件 E 的文本相似度 $Sim_HO(d_{hi}^m, d_E)$ ($d_{hi}^m \in Blog_{hi}$), 考虑到距离初始时刻越近的博文参考性越强^[11], 本文引入时间衰减因子对文本相似度进行调整。文本相似度 $Sim_HO(d_{hi}^m, d_E)$ 的计算方法如式 (4) 所示:

$$Sim_HO(d_{hi}^m, d_E) = \frac{\sum_{l=1}^L \omega_{hi-l}^m \times \omega_{E-l}}{\sqrt{\sum_{l=1}^L (\omega_{hi-l}^m)^2} \times \sqrt{\sum_{l=1}^L \omega_{E-l}^2}} \times e^{-\gamma(t_0-t^m)} \quad (4)$$

其中, d_{hi}^m 表示用户 u_i 的某篇历史博文, d_E 表示当前事件 E , ω_{hi-l}^m 为用户 u_i 的历史博文 d_{hi}^m 中第 l 个关键词的权重, ω_{E-l} 为当前事件 E 的第 l 个关键词的权重, 关键词权重基于 TF-IDF 方法计算获得。 t_0 表示初始时刻, t^m 表示历史博文 d_{hi}^m 发表的时间, t_0-t^m 表示历史博文 d_{hi}^m 发表的时间与初始时刻的距离, γ 表示衰减率。本文根据文献[12]将衰减率 γ 设置为 0.02。

选取文本相似度 $Sim_HO(d_{hi}^m, d_E)$ 高于阈值 ξ 的历史博文作为当前事件 E 的相似微博集 $Blog_{si}$, $Blog_{si} = \{blog_{si}^1, \dots, blog_{si}^h\}$ 。

通过综合相似博文集 $Blog_{si}$ 中微博的观点值, 可得到用户 u_i 的初始观点值 $S_{u_i}(0)$, 如式 (5) 所示:

$$S_{u_i}(0) = \frac{1}{n} \sum_{h=1}^n Senti_{si}^h \quad (5)$$

其中, $Senti_{si}^h$ 为相似博文集 $Blog_{si} = \{blog_{si}^1, \dots, blog_{si}^h, \dots, blog_{si}^n\}$ 中每条微博的情感值。考虑到用户对舆情事件的情感是用户对事件观点的直接体现^[13], 本文基于情感分析获取用户的情感值, 作为其对舆情事件的观点值。在后续实证研究中也采用情感值量化用户的观点。

3.2 社交用户 u_i 信任集合 $I(u_i, S(t))$ 的获取

在选择社交用户 u_i 的信任用户集合 $I(u_i, S(t))$ 时, 本文不仅考虑观点差值, 还考虑了用户间的亲密度, 因为亲密度高的用户一般兴趣观点接近。但是, 若直接将用户间的观点差值和亲密度融合, 则会减弱观点差值在选择信任用户时的重要性, 影响信任用户选择的准确性, 例如 A 、 B 、 C 3 个用户在针对某个事件 E 的观点时, 假设 A 与 B 的观点差值为 0.1 (观点接近), A 与 C 的观点差值为 0.9 (观点几乎不同), A 与 B 的亲密度为 0.01 (几乎不是朋友), A 与 C 的亲密度为 0.99 (非常亲

密), 若直接将观点差值和亲密度融合计算, 则结果可能会倾向于 C 的观点更新用户 A 的观点, 而实际上二者对事件 E 的观点几乎不同, 显然不符合情况。为解决上述问题, 将采用二次筛选的手段选择出更为准确、完备的信任用户, 其过程为: 首先选取与用户 u_i 的观点差值小于信任阈值 ε 的社交用户, 构成信任集合 $I_1(u_i, S(t))$, 然后通过计算用户间亲密度, 修订 ε , 二次判定小于修订后 ε 的社交用户, 构成信任集合 $I_2(u_i, S(t))$, 综合 $I_1(u_i, S(t))$ 和 $I_2(u_i, S(t))$ 获取 $I(u_i, S(t))$ 。上述过程的关键是利用亲密度修订阈值 ε , 社交用户 u_i 对 u_k 的亲密度 F_{ik} 的计算方法如式 (6) 所示:

$$F_{ik} = \begin{cases} \phi \times d_{ik} + (1-\phi) \times INT_{ik}, & d_{ik} = 1 \\ (1-\phi) INT_{ik}, & d_{ik} = 0 \end{cases} \quad (6)$$

其中, d_{ik} 用于量化社交用户 u_i 与用户 u_k 的邻居关系, 其取值为 0 或 1, 如果两个用户间有关注关系, 则认为二者之间存在邻居关系, d_{ik} 取值为 1, 否则取值为 0。 ϕ 用来调和邻居关系和社交关系的贡献程度。 INT_{ik} 为社交用户 u_i 和 u_k 的交互度, 计算方法如式 (7) 所示:

$$INT_{ik} = \mu_1 \times \frac{R(i, k)}{\sum_{p \in I(i)} R(i, p)} + \mu_2 \times \frac{C(i, k)}{\sum_{p \in I(i)} C(i, p)} + \mu_3 \times \frac{M(i, k)}{\sum_{p \in I(i)} M(i, p)} \quad (7)$$

其中, $I(i)$ 表示用户 u_i 对其交互群体 U 中施加过交互行为的用户集合, p 表示 $I(i)$ 中的某个用户, $R(i, k)$ 表示用户 u_i 对 u_k 发布的博文施加转发行为的次数, $C(i, k)$ 表示用户 u_i 对 u_k 的博文施加评论行为的次数, $M(i, k)$ 表示用户 u_i 对 u_k 的博文施加点赞行为的次数。 μ_1 、 μ_2 、 μ_3 分别表示转发、评论、点赞行为的权重, 且 $\mu_1 + \mu_2 + \mu_3 = 1$ 。

基于式 (6) 和式 (7), 给出了用于二次筛选信任用户的修订后的信任阈值 $\varepsilon_{F_{ik}}$ 的计算方法, 如式 (8) 所示:

$$\varepsilon_{F_{ik}} = (1 + e^{\frac{1}{F_{ik}}}) \varepsilon \quad (8)$$

综合以上分析, 可得到用户 u_i 的信任集合 $I(u_i, S(t))$, 如式 (9)~(11) 所示:

$$I_1(u_i, S(t)) = \{u_j \mid |S_{u_i}(t) - S_{u_k}(t)| \leq \varepsilon\} \quad (9)$$

$$I_2(u_i, S(t)) = \{u_j \mid |S_{u_i}(t) - S_{u_k}(t)| \leq \varepsilon_{F_{ik}}\} \quad (10)$$

$$I(u_i, S(t)) = I_1(u_i, S(t)) \cup I_2(u_i, S(t)) \quad (11)$$

3.3 用户 u_j 对 u_i 观点影响权重 α_{ij} 的计算

本文将通过综合用户 u_j 的全局影响力、用户

u_i 与用户 u_j 的交互度 INT_{ij} ，以及用户 u_i 与用户 u_j 的观点接近度这 3 个方面因素，量化信任集合中的用户 u_j 对用户 u_i 的观点影响权重 α_{ij} ，其中交互度 INT_{ij} 依据式 (7) 计算。

1) 用户 u_j 的全局影响力

在社交网络中，越多的用户对 u_j 的历史博文进行转发、评论、点赞，说明用户 u_j 在社交网络中影响力越高。此外，社交网络中影响力较高的用户具有高粉丝数、低关注数特征^[14]。综上，本文将基于用户 u_j 的博文影响力和粉丝关注比来综合度量其全局影响力。博文影响力 $BI(j)$ 和粉丝关注比 $FR(j)$ 分别采用式 (12) 和式 (13) 计算。

$$BI(j) = \frac{\mu_1 \times retw_j + \mu_2 \times com_j + \mu_3 \times pra_j}{N_j} \quad (12)$$

$$FR(j) = \frac{followers(j)}{followees(j)} \quad (13)$$

其中， μ_1 、 μ_2 、 μ_3 表示转发、评论、点赞的权重系数，且 $\mu_1 + \mu_2 + \mu_3 = 1$ 。 N_j 代表用户 u_j 历史发表的博文总数量， $retw_j$ 、 com_j 、 pra_j 分别表示用户 u_j 历史博文被转发、评论、点赞的总次数。 $followers(j)$ 表示用户 u_j 的粉丝数， $followees(j)$ 表示用户 u_j 的关注数。

综合博文影响力和粉丝关注比可计算出用户 u_j 的全局影响力 $GI(j)$ ，如式 (14) 所示：

$$GI(j) = \beta \times \frac{BI(j) - \min(BI)}{\max(BI) - \min(BI)} + (1 - \beta) \times \frac{FR(j) - \min(FR)}{\max(FR) - \min(FR)} \quad (14)$$

其中， $\max(BI)$ 、 $\min(BI)$ 分别表示所有社交用户中博文影响力的最大值和最小值， $\max(FR)$ 、 $\min(FR)$ 表示所有用户粉丝关注比的最大值和最小值。由于在衡量用户的影响力时，用户的博文内容和其粉丝关注比两方面因素同等重要，因此，本文将 β 设置为 0.5。

2) 用户 u_i 与 u_j 的观点接近度

本文基于用户 u_i 与 u_j 的观点距离获取观点接近度 SP_{ij} ，用户间观点距离越小表明其观点接近度越高。 SP_{ij} 的计算方法如式 (15) 所示：

$$SP_{ij} = \frac{1}{1 + |S_{u_i}(t) - S_{u_j}(t)|} \quad (15)$$

其中， $S_{u_i}(t)$ 、 $S_{u_j}(t)$ 分别表示用户 u_i 与 u_j 在 t 时刻的观点值，两者差值越小表明两用户间观点越接近。

3) 用户 u_i 对用户 u_j 的综合信任度

综合上述的用户全局影响力、用户间交互度、用户间观点接近度，可计算出用户 u_i 对 u_j 的综合信任度 $CT(u_i, u_j)$ ，其计算方法如式 (16) 所示：

$$CT(u_i, u_j) = \theta_1 GI(j) + \theta_2 INT_{ij} + \theta_3 SP_{ij} \quad (16)$$

其中， θ_1 、 θ_2 、 θ_3 为 3 个方面信任度的贡献权重，且 $\theta_1 + \theta_2 + \theta_3 = 1$ ，综合信任度 $CT(u_i, u_j)$ 即为用户 u_j 对 u_i 观点更新时的影响权重 α_{ij} 。

3.4 基于改进 HK 模型的观点演化预测

综合式 (5)、式 (8)~(11)、式 (16)，可以得到改进的 HK 模型，如式 (17)~(19) 所示：

$$I(u_i, S(t)) = I_1(u_i, S(t)) \cup I_2(u_i, S(t)) = \{u_j \mid |S_{u_i}(t) - S_{u_k}(t)| \leq \varepsilon\} \cup \{u_j \mid |S_{u_i}(t) - S_{u_k}(t)| \leq \varepsilon_{F_{ik}}\} \quad (17)$$

$$S_{u_i}(t+1) = \sum_{u_j \in I(u_i, S(t))} \alpha_{ij} S_{u_j}(t) \quad (18)$$

$$\alpha_{ij} = \theta_1 GI(j) + \theta_2 INT_{ij} + \theta_3 SP_{ij} \quad (19)$$

基于上述模型，通过有效筛选社交用户 u_i 在 t 时刻的信任用户集合 $I(u_i, S(t))$ ，可以获得预测 u_i 在 $t+1$ 时刻观点的信息来源 (t 时刻 $I(u_i, S(t))$ 中用户的观点值)，为了实现更为合理的预测，式 (19) 给出了不同信任用户对观点预测的影响权重。最终，通过对 t 时刻 u_i 的所有信任用户的观点值加权求平均就可以实现对 u_i 在 $t+1$ 时刻观点的预测，如式 (18) 所示。以此重复，就可以实现对社交用户 u_i 在 $t+2$ 、 $t+3$ 、 $\dots$ 、 $t+n$ 时刻观点的预测，进而实现对观点演化的预测。

4 实证研究

4.1 实验数据集

目前社交用户观点演化预测研究中尚无满足本文实验要求的权威数据集，故本文采用 Python 爬虫工具从新浪微博获取数据，以验证本文研究内容的有效性。2022 年 6 月 10 日，河北唐山某个烧烤店发生了一起暴力殴打事件，随后在微博引起广泛讨论，针对该事件采集了 2022 年 6 月 10 日—8 月 29 日的相关微博数据，共计 112 802 条微博，涉及 85 180 名用户，其中活跃的用户将作为观点预测模型的用户集，活跃用户的判定标准为在该话题的各个发展阶段均参与讨论的用户，依据该标准，最终选取了 112 个活跃用户，并采集了这些用户过去 1 个月的历史数据，将其作为确定用户初始观点值的数据来源。

为了确定“唐山打人事件”的具体发展阶段，

依据事件的发展时间对的相关博文的数量进行统计, 绘制舆情热度趋势图, 如图 2 所示。从图 2 可以看出, 该舆情事件具有明显的波动特征^[15], 且事件的发展经历了 4 个阶段, 分别为:

1) 2022 年 6 月 10—15 日为第一阶段: 唐山打人事件在社交网络中迅速传播, 从而引发了用户的广泛讨论, 进入第一个生命周期阶段。

2) 2022 年 6 月 16—18 日为第二阶段: 唐山打人事件初步调查结果通报, 从而引发第二次广泛讨论, 使舆情发展进入第二个生命周期阶段。

3) 2022 年 6 月 19—24 日为第三阶段: 唐山打人事件中相关嫌疑人全部抓获, 从而引发网友对该事件的讨论, 使舆情发展进入第三个生命周期阶段。

4) 2022 年 6 月 25 日后为第四阶段: 唐山打人事件舆情发展进入消退阶段, 用户对该事件的讨论趋于衰退。

在上述 4 个阶段中, 由于第四阶段属于消退期, 用户观点趋于平稳, 故本文后续实验将围绕前 3 个阶段的观点演化进行预测。

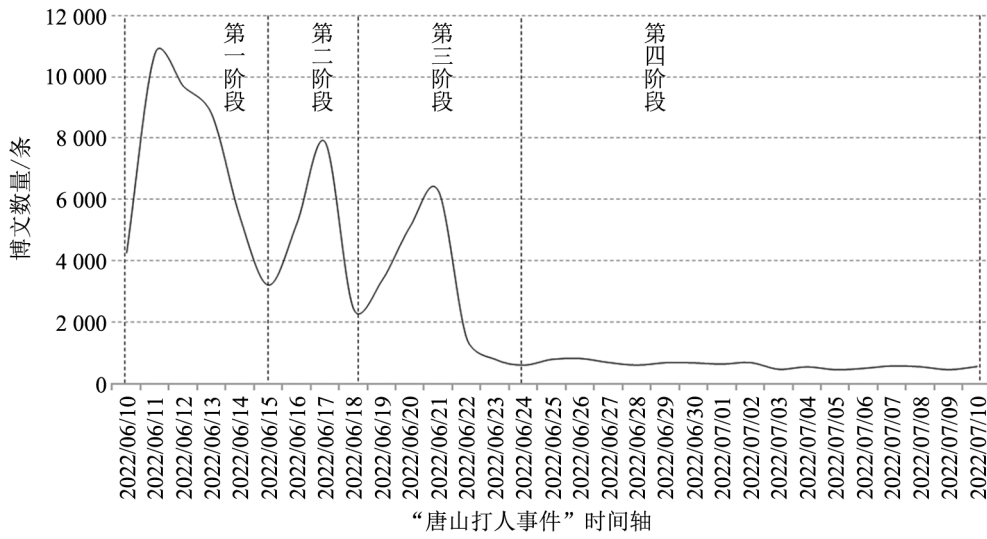


图 2 “唐山打人事件” 舆情热度趋势图

Fig. 2 Trend Chart of Public Opinion Heat for the “Tangshan Assault Incident”

4.2 评价标准

为了评价改进的 HK 模型用于社交用户观点演化预测的性能, 本文采用观点值的预测误差来对模型的性能进行评估。已有研究中常用的误差评价指标主要有均方误差 (MSE) 和平均绝对误差 (MAE) 等。均方误差 MSE 值可以评价数据的变化程度, MSE 的值越小, 说明预测模型具有更高的精确度, 而平均绝对误差 MAE 值能更好地反映预测值误差的实际情况。

本文采用以上两个常用指标进行预测模型的误差评价, 对模型的有效性进行验证。均方误差 (MSE) 和平均绝对误差 (MAE) 的计算方法如式 (20)、式 (21) 所示:

$$MSE = \frac{1}{n} \sum_{i=1}^n (S_{u_i}^{real} - S_{u_i}^{pre})^2 \quad (20)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |S_{u_i}^{real} - S_{u_i}^{pre}| \quad (21)$$

其中, n 表示用户总数量, $S_{u_i}^{real}$ 表示用户 u_i 的

真实观点值, $S_{u_i}^{pre}$ 表示用户根据本文改进模型预测得出的观点值。后续实验将通过计算改进 HK 模型所得预测观点值和真实观点值之差表示两者的差距, 进而获得所有用户观点值的误差之和, 并将其平均值作为整个模型的误差。

4.3 实验结果及分析

本文的实验主要包括 3 部分: 第一部分为相关参数的确定, 第二部分为模型有效性检验, 第三部分为不同模型预测性能的比较。在第一部分实验中, 涉及的相关参数包括 5 个, 分别是: 文本相似度阈值 ξ 、信任阈值 ε 、交互行为贡献权重 μ_i ($i=1, 2, 3$)、综合信任度计算时的影响因素权重 θ_j ($j=1, 2, 3$) 及用户间亲密度计算时的调和参数 ϕ 。

4.3.1 相关参数的确定

1) 文本相似度阈值 ξ 的确定

文献[16]指出, 用户对舆情事件的情感是用户对事件观点的直接体现, 故本文基于情感分析获取

用户对舆情事件的观点值。为了确定不同相似度阈值下获取的初始观点值的准确性，将不同阈值下的初始观点值与用户真实初始观点值进行比较，其中，通过对爬取到的用户初始时刻发布博文进行情感分析获得其真实初始观点值 $S_{u_i}^{true}$ ，通过筛选不同文本相似度阈值下的用户历史博文获得不同阈值 ξ 下的初始观点值 $S_{u_i}^\xi$ ；然后计算两者的均方误差 MSE 值和平均绝对误差 MAE 值，从而确定最优取值。

借鉴代晓丽等^[17]对文本相似度阈值的设置规则，将文本相似度阈值设置为 0.5、0.6、0.7、0.8、0.9。通过反复实验法尝试确定文本相似度阈值 ξ 。实验结果如表 1 所示。

表 1 不同文本相似度阈值 ξ 下的 MSE 值和 MAE 值

Tab. 1 MSE and MAE Values under Different Text Similarity Threshold ξ

ξ	MSE 值	MAE 值
0.5	0.1098	0.2619
0.6	0.1076	0.2720
0.7	0.1034	0.2585
0.8	0.1099	0.2638
0.9	0.1122	0.2683

从表 1 可以看出，当文本相似度阈值 ξ 为 0.7 时，MSE 值和 MAE 值都较低，故本文设定文本相似度阈值为 0.7。

2) 用户信任阈值 ε 的确定

通过上述实验，已经确定了文本相似度阈值 ξ 为 0.7，进而能够计算获得当前事件 E 中活跃用户的初始观点值。为获取待预测用户的信任集合，需首先确定用户的信任阈值 ε 。将基于活跃用户真实聚类结果与演化预测结果进行比较，从而确定合适的信任阈值 ε 。首先，通过将活跃用户的真实观点值进行聚类，获得真实观点的演化聚类数量；其次，利用本文改进的 HK 模型进行观点演化预测，获得活跃用户不同信任阈值下的观点演化预测结果；最后，比较真实观点聚类数量和不同信任阈值下的观点演化预测数量，确定用户的信任阈值 ε 。

本文通过对活跃用户真实博文数据的观点值进行 K-means 聚类，得到活跃用户真实观点聚合情况。考虑到不同观点值数量的设定对聚类效果的影响，本文使用轮廓系数^[18]度量设定不同最终观点值聚合数量下的聚类效果，对于用户 u_i 的观点值

S_{u_i} ，其轮廓系数 S_i 的计算方法如式 (22) 所示。

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (22)$$

其中， a_i 表示用户 u_i 的观点值 S_{u_i} 与其他活跃用户观点值的平均距离， b_i 表示用户 u_i 的观点值 S_{u_i} 与所有活跃用户观点值平均距离的最小值， $\max(a_i, b_i)$ 表示 a_i 和 b_i 中的较大值。通过计算所有活跃用户观点值的平均轮廓系数，度量聚类的效果。平均轮廓系数越大，聚类效果越好。

实验中将活跃用户最终观点值聚合数量依次取值 2~8，步长为 1，计算不同观点值聚合数量下的平均轮廓系数，结果如图 3 所示。由图 3 可知，当活跃用户观点值的聚合数量为 4 时，平均轮廓系数最大，可知最终活跃用户的观点聚合数量为 4 个。

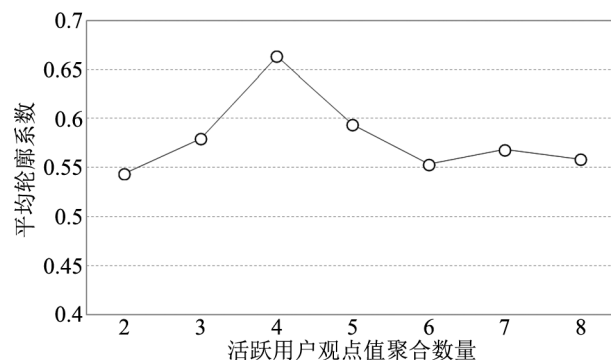


图 3 不同观点值聚合数量下的平均轮廓系数

Fig. 3 Average Silhouette Coefficient for Different Aggregation Quantities of Viewpoint Values

通过分析聚类结果，发现“唐山打人事件”最终形成 4 个观点，分别为：①媒体指责此次暴力打人事件，要求严惩；②媒体希望政府部门披露受伤女孩状况；③唐山黑恶势力猖獗，质疑唐山“官黑一家”；④担忧女性安全，探讨女性如何保护自己。

完成上述观点聚类后，本文将基于论文提出的方法可视化展现观点的演化及聚合，对比二者的拟合情况。由于当信任阈值较大时，会增加原本不交流用户间的交流性，从而影响模型的预测效果。因此本文借鉴曹春萍等^[10]的研究成果，将信任阈值设置为 0.1、0.2、0.3，并利用本文实验得出的初始观点值进行活跃用户的观点演化过程，得到不同信任阈值 ε 下的活跃用户观点演化趋势图，如图 4 所示。

观察图 4 的 3 幅实验结果图可以发现，当信任阈值 $\varepsilon=0.1$ 、 $\varepsilon=0.2$ 及 $\varepsilon=0.3$ 时，活跃用户的观点聚合数量分别为 4 个观点、2 个观点及 1 个观点。

综合以上两部分实验可知，当信任阈值为 0.1

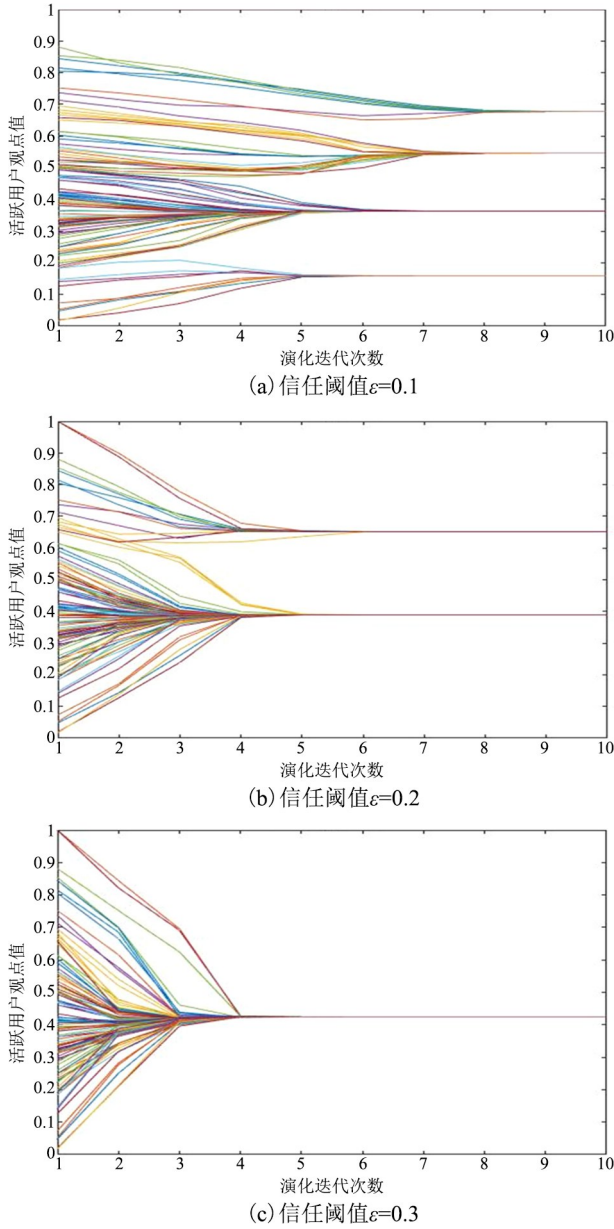


图 4 不同信任阈值下的活跃用户观点演化趋势图
Fig. 4 Trend Chart of Opinion Evolution among Active Users under Different Trust Thresholds

时, 本文的观点演化预测趋势聚合为 4 个观点, 该结果与真实活跃用户的观点聚合数量一致, 故后续将 0.1 作为活跃用户观点演化预测的信任阈值。

3) 交互行为贡献权重 μ_i 的确定

用户间的交互行为包括转发、评论和点赞, 不同的交互行为对用户间交互度以及博文影响力的计算影响不同, 本文采用层次分析法^[19]确定各种交互行为的贡献权重。首先, 依据“1~9 标度方法”, 对交互行为的贡献度进行两两比较, 构建出如表 2 所示的权重判定矩阵。

根据表 2, 可求出上述判定矩阵的 CR 值为 0.0051, 通过了一致性检验。该判定矩阵的最大特

表 2 交互行为贡献权重判定矩阵

Tab. 2 Matrix for Determining Weighted Contribution of Interactive Behaviors

贡献权重	转发	评论	点赞
转发	1	2	3
评论	1/2	1	2
点赞	1/3	1/2	1

征值 $\lambda_{\max} = 3.009$, 此时对应的特征向量 $W = (0.847, 0.466, 0.256)^T$, 将向量 W 标准化后可得向量 $W' = (0.540, 0.297, 0.163)^T$, 每一维的值分别对应各交互行为为权重, 即转发行为的贡献权重 $\mu_1 = 0.540$, 评论行为的贡献权重 $\mu_2 = 0.297$, 点赞行为的贡献权重 $\mu_3 = 0.163$ 。

4) 亲密度计算时调和参数 ϕ 的确定

本文采用 Tunning 实验法^[20]确定调和参数 ϕ 的取值。 ϕ 的取值会对后续信任集合的选取产生影响, 在特定话题事件下, 用户之间的交互关系比用用户间的关注关系更能体现用户之间的影响情况^[21], 故本文实验将调和参数 ϕ 的取值范围设定为 $[0.1, 0.5]$ 。采用平均文本相似度 Ave_sim 指标来确定调和参数 ϕ 值, 平均文本相似度 Ave_sim 的计算公式如式 (23) 所示:

$$Ave_sim = \frac{\sum_{i=1}^n \sum_{j \in I_2(u_i, S(t))} \cos(d_i, d_j)}{n} \quad (23)$$

其中, $\cos(d_i, d_j)$ 表示用户 u_i 历史博文与亲密信任集合中用户历史博文的余弦相似度, $\sum_{i=1}^n$

$\sum_{j \in I_2(u_i, S(t))} \cos(d_i, d_j)$ 表示所有用户的文本相似度之和, n 表示社交用户总数。本文实验在 ϕ 取值 0.1、0.2、0.3、0.4、0.5 时, 分别获取对应的 Ave_sim 值, 得到不同调和因子 ϕ 与平均文本相似度之间的关系, 如图 5 所示。由图 5 可知, 当调和因子取值为 0.3 时, 平均文本相似度较高, 故本文设定调和因子 ϕ 为 0.3。

5) 综合信任度贡献权重 θ_j 的确定

贡献权重 $\theta_j (j=1, 2, 3)$ 是指全局影响力、用户间交互度和用户间观点接近度 3 个影响因素的权重。本文采用熵值法^[22]确定 θ_1 、 θ_2 、 θ_3 的取值: 计算全局影响力、用户间交互度、用户间观点接近度的信息熵 H'_1 、 H'_2 、 H'_3 , 然后基于 3 个信息熵计算贡献权重 θ_1 、 θ_2 、 θ_3 的值。计算方法如式 (24) 和

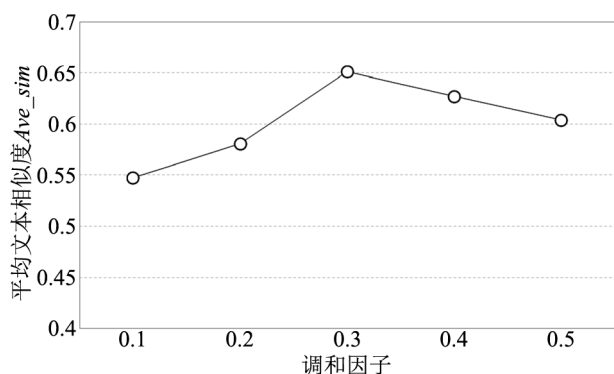


图 5 不同调和因子 ϕ 下的平均文本相似度

Fig. 5 Average Text Similarity under Different Harmonization Factors ϕ

式 (25) 所示:

$$H'_i = -\frac{1}{\ln m} \sum_{j=1}^m p_{ij} \cdot \ln p_{ij} \quad (24)$$

$$\theta_i = \frac{1-H'_i}{3-\sum H'} \quad (25)$$

其中, p_{ij} 表示第 i 种属性第 j 种取值的频率。 H'_i 表示熵值法中第 i 种属性的信息熵, 由 m 个不同属性值的分布决定。属性取值越丰富, H'_i 越小, 则相应属性在综合信任度中影响程度越高。最终得到 $H'_1 = 0.937$ 、 $H'_2 = 0.911$ 、 $H'_3 = 0.903$ 、 $\theta_1 = 0.253$ 、 $\theta_2 = 0.357$ 、 $\theta_3 = 0.390$ 。

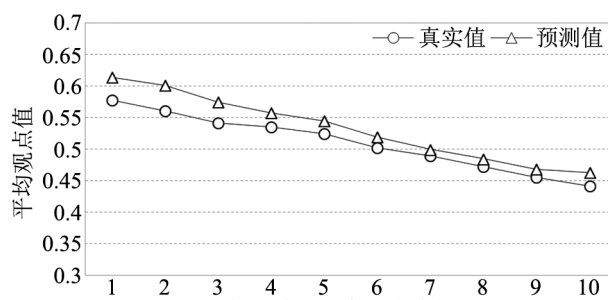
4.3.2 模型有效性检验

依据“唐山打人事件”舆情热度趋势图, 选取事件前 3 个发展阶段验证对比预测观点值和真实观点值, 通过拟合程度来检验模型的优劣。该实验首先计算了 112 个用户在不同阶段的观点演化值, 考虑到用户对事件的观点由群体用户共同决定, 将利用每一发展阶段的群体平均观点值判断预测值和真实值的拟合情况, 拟合结果如图 6 所示。

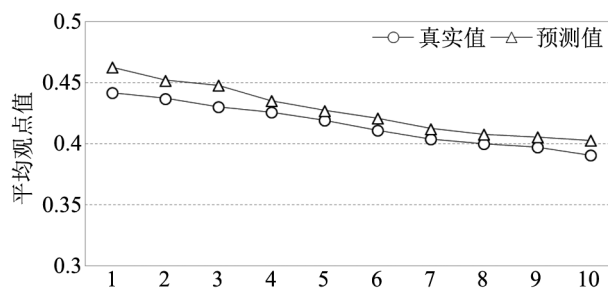
从图 6 中可以看出, 将“唐山打人事件”的舆情发展的 3 个阶段分别进行观点演化预测值和真实值的拟合, 3 个阶段都获得了较好的拟合效果, 从而验证了将改进 HK 模型用于社交用户观点演化预测的可行性和有效性。

4.3.3 对比实验

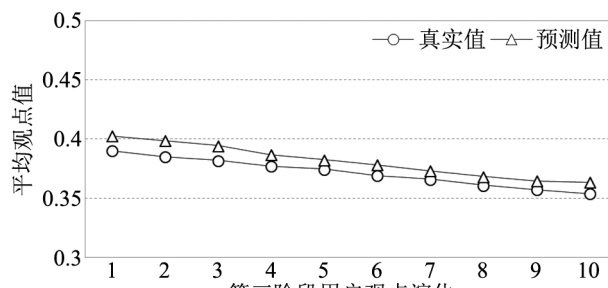
本文基于用户的历史数据以及用户间的社交关系等对 HK 模型中初始观点值、信任阈值以及观点影响权重 3 部分进行改进, 用于社交用户观点预测的研究。为验证研究内容的有效性, 将对比本文提出的改进 HK 模型与经典 HK 模型、文献[10]中 URHK 模型在社交用户观点演化预测中的性能, 性



(a) 第一阶段观点演化拟合结果图



(b) 第二阶段观点演化拟合结果图



(c) 第三阶段观点演化拟合结果图

图 6 用户观点演化预测拟合结果

Fig. 6 Fitted Results of User Opinion Evolution Prediction

能评价指标采用 MSE 值和 MAE 值, 实验结果如表 3 和表 4 所示。

表 3 模型不同阶段演化预测 MSE 值

Tab. 3 MSE Values for Evolutionary Prediction at Different Stages of the Model

模 型	MSE-1	MSE-2	MSE-3	平均 MSE
经典 HK 模型	0.1154	0.1122	0.1081	0.1119
URHK 模型 ^[10]	0.1098	0.1077	0.1049	0.1075
改进 HK 模型	0.1055	0.1037	0.1011	0.1034

从表 3 和表 4 中可以得出, 本文提出的改进 HK 模型在用于观点演化预测时, 其均方误差 MSE 值和平均绝对误差 MAE 值均优于其他两个模型。这是因为本文改进的 HK 模型在进行社交用户的观点演化预测时引入了用户的历史数据, 从而使得用户的初始观点值的设定更加符合现实, 并且在用户

表 4 模型不同阶段演化预测 MAE 值
Tab. 4 MAE Values for Evolutionary Prediction at
Different Stages of the Model

模 型	MAE-1	MAE-2	MAE-3	平均 MAE
经典 HK 模型	0.3003	0.2842	0.2552	0.2799
URHK 模型 ^[10]	0.2933	0.2782	0.2527	0.2747
改进 HK 模型	0.2712	0.2561	0.2482	0.2585

间的观点交互时考虑了用户间的真实交互行为以及用户间的交互影响权重,使得该模型用于用户观点演化预测时,更加符合用户间的真实交互行为,降低了用户观点演化预测的误差。

5 结 语

本文通过分析 HK 有界信任模型的不足,提出了改进的 HK 模型,并将其用于社交用户观点演化预测研究,该模型虽然取得了较好的性能,但本文研究尚存在以下不足:①在社交用户间观点影响权重的计算中,本文仅计算了事件发生初始阶段的用户间观点影响权重,未考虑其动态变化;②在实验分析中,本文改进的模型在分阶段观点预测时,可能会出现预测误差的波动,即在舆情事件演化阶段发生转变时,存在误差升高的情况;③本文的研究仅是预测社交网络中用户对事件的观点演化,并没有深入分析其对舆情发展态势的影响。在未来研究中,将围绕上述不足进行深入研究:①针对社交用户在舆情事件中的观点影响权重动态变化展开研究,探索更加即时的用户间观点影响权重,以使模型的预测效果更准确;②将注重舆情发展阶段发生转变时的用户观点变化,以探究事件发展阶段发生转变时用户观点变化的影响因素,提升 HK 观点演化模型的准确性;③尝试将用户观点的演化融入到网络舆情态势预测中,助力实现对各类网络舆情的准确预测。

参 考 文 献

[1] Grabowski A, Kosiński R A. Ising-based Model of Opinion Formation in a Complex Network of Interpersonal Interactions [J]. Physica A: Statistical Mechanics and Its Applications, 2006, 361 (2): 651-664.
[2] Sznajd-Weron K, Sznajd J. Opinion Evolution in Closed Community [J]. International Journal of Modern Physics C, 2000, 11 (6): 1157-1165.
[3] Holley R A, Liggett T M. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model [J]. The Annals of Probability, 1975: 643-663.

[4] Galam S. Application of Statistical Physics to Politics [J]. Physica A: Statistical Mechanics and Its Applications, 1999, 274 (1-2): 132-139.
[5] Deffuant G, Neau D, Amblard F, et al. Mixing Beliefs Among Interacting Agents [J]. Advances in Complex Systems, 2000, 3 (1): 87-98.
[6] Hegselmann R, Krause U. Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation [J]. Journal of Artificial Societies and Social Simulation, 2002, 5 (3): 1-33.
[7] 李根强, 刘莎, 张亚楠, 等. 信息熵理论视角下网络集群行为主体的观点演化研究 [J]. 情报科学, 2020, 38 (1): 42-47, 86.
[8] 祁凯, 黎冰雪, 杨志. 基于平台推荐机制和有界信任影响的网络群体观点演化研究 [J]. 情报杂志, 2021, 40 (5): 132-142, 185.
[9] 徐涵, 管安琪, 李紫琦. 基于意见气候的 Hegselmann-Krause 观点动力学建模与分析 [J]. 中国传媒大学学报 (自然科学版), 2021, 28 (4): 12-19.
[10] 曹春萍, 康远路. 融合用户相似性的 Hegselmann-Krause 模型研究 [J]. 小型微型计算机系统, 2021, 42 (11): 2344-2349.
[11] 张晓君, 黄微, 李瑞. 心理距离视角下网络舆情传播意愿影响因素研究 [J]. 图书情报工作, 2019, 63 (12): 86-93.
[12] Choi J, Croft W B. Temporal Models for Microblogs [C] // Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012: 2491-2494.
[13] 苏轲恒, 张敏, 王锦. 图书情报领域情感研究现状述评 [J]. 图书情报工作, 2022, 66 (6): 137-147.
[14] 曾子明, 王婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例 [J]. 情报学报, 2019, 38 (1): 89-96.
[15] 曹树金, 岳文玉. 突发公共卫生事件微博舆情主题挖掘与演化分析 [J]. 信息资源管理学报, 2020, 10 (6): 28-37.
[16] 苏轲恒, 张敏, 王锦. 图书情报领域情感研究现状述评 [J]. 图书情报工作, 2022, 66 (6): 137-147.
[17] 代晓丽, 刘世峰, 宫大庆. 基于 NLP 的文本相似度检测方法 [J]. 通信学报, 2021, 42 (10): 173-181.
[18] Rousseeuw P J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.
[19] Tam M C Y, Tummala V M R. An Application of the AHP in Vendor Selection of a Telecommunications System [J]. Omega, 2001, 29 (2): 171-182.
[20] Bertsimas D, Gupta V, Kallus N. Data-driven Robust Optimization [J]. Mathematical Programming, 2018, 167 (2): 235-292.
[21] 吴岷辉, 张晖, 赵旭剑, 等. 基于用户行为网络的微博意见领袖挖掘算法 [J]. 计算机应用研究, 2015, 32 (9): 2678-2683.
[22] 杨宇. 多指标综合评价中赋权方法评析 [J]. 统计与决策, 2006, (13): 17-19.

(责任编辑: 郭沫含)