基于 HMM 的话题风险状态预测方法研究

蔡婷婷[1],朱恒民[1,2],魏　静[1]
(1. 南京邮电大学 管理学院,江苏 南京 210003;
2. 江苏高校哲学社会科学重点研究基地—信息产业融合创新与应急管理研究中心,
江苏 南京 210003)

摘　要:如何在互联网海量信息中预测话题风险性和演化趋势,是舆情监管部门的工作重点。针对话题演化趋势预测研究中存在的不足:话题状态划分方法单一、话题状态演化预测研究缺乏等,从话题预警的视角,提出话题风险状态预测方法,为舆情监管部门提供预警依据。首先,基于向心度和密度指标划分不同等级的话题风险状态,直观地刻画话题引发舆论危机的风险程度;其次,基于 HMM(Hidden Markov Model,隐马尔可夫模型)对各话题风险状态构建模型,并将各风险状态下所对应的观测序列数据作为训练集训练模型;最后,根据极大相似准则选用最佳模型预测话题观测值,进而借助平面坐标映射法得到话题在未来时刻的风险状态。以新冠肺炎疫情事件为研究样本话题,验证基于 HMM 的话题风险状态预测方法的有效性,交叉检验的平均预测准确率达到 90% 以上,相比于 BP 神经网络、LSTM 以及 RNN 时间序列预测模型,该方法的预测误差更小。

关键词:隐马尔可夫模型;话题状态;话题演化;趋势预测;风险预警
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2023)05-0029-06
doi:10.3969/j.issn.1673-629X.2023.05.005

Research on Prediction Method of Topic Risk States Based on HMM

CAI Ting-ting[1],ZHU Heng-min[1,2],WEI Jing[1]
(1. School of Management,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. Jiangsu University Philosophy and Social Science Key Research Base—Information Industry Integration
Innovation and Emergency Management Research Center,Nanjing 210003,China)

Abstract:Predicting the riskiness and evolutionary trend of topics in the massive information on the Internet is work priority of public opinion supervisory departments. In view of the problems in the current analysis of topics,such as the single method of topic state classification and the lack of research on topic state evolution prediction,a prediction method of topic risk states was proposed from the perspective of topic early warning,aiming to provide the basis for public opinion supervision departments. First,the method classified different levels of topic risk states based on two indexes of centrality and density,in order to visually portray the risk level of topics triggering public opinion crisis. Then,the model was built for each risk state based on the hidden Markov model,and the observation sequence data corresponding to each risk state was used as the training set to train the model. Finally,the best model was selected to predict the observed value of the topic according to the maximum likelihood principle,so as to obtain the risk state of the topic at the future moment with the help of the plane coordinate mapping method. The COVID-19 epidemic event was taken as the research sample topic to verify the effectiveness of the topic risk state prediction method. The average prediction accuracy of cross validation is over 90%. In addition,the proposed method has lower prediction error than the time series prediction models of BP neural network, LSTM and RNN.

Key words:hidden Markov model;topic state;topic evolution;trend prediction;risk warning

0　引　言

互联网社交媒体是用户发布、传播和获取海量话题信息的重要平台。网络话题是在不断演化的,话题的迅速发酵与扩散会引发网络舆论,甚至是舆情危机。

收稿日期:2022-07-10　　**修回日期**:2022-11-15
基金项目:国家自然科学基金项目资助项目(71874088,71704085);江苏省研究生科研与实践创新计划项目资助项目(KYCX21_0835)
作者简介:蔡婷婷(1997-),女,硕士研究生,通信作者,研究方向为舆情传播研究;朱恒民(1974-),男,教授,博士,研究方向为数据挖掘、舆情管理研究;魏　静(1982-),女,教授,博士,研究方向为复杂网络、舆情传播研究。

话题的状态可用于描述话题本身的发展趋势和舆论爆发的风险性,对话题的状态演化趋势进行预测有助于舆论监管部门及时采取措施,避免引发舆情危机,进而实现社交网络信息传播的有效监管。

话题演化是对已有话题随着时间演化情况进行的分析[1-2]。话题的状态演化属于话题演化分析的研究范畴,已有工作多是基于生命周期的视角来回溯话题状态的演化过程。Chen 等提出一种基于生命周期的老化理论,将话题发展分为萌芽、生长、衰退和消亡四个周期,并将其与传统的 single-pass 聚类算法相结合,自适应地检测和跟踪在线序列话题事件[3];贾亚敏和曹树金等结合话题生命周期理论将话题状态分为起始、爆发、波动和平息四个阶段,探索每个阶段的话题演化规律[4-5]。部分学者通过定义指标来回溯话题所处的生命周期阶段:Y. Tu 等基于老化理论提出新颖指数,并结合已发表量指数来探测处于生命周期新生阶段的热点话题[6];Collon 等基于共词分析法提出了向心度和密度两个指标,用于评价科技文献主题的重要性和成熟度[7];刘自强等基于这两个指标,通过平面坐标映射法将科技文献主题划分为新生、成长、收缩、消亡四个生命周期阶段,以期描述主题在整个生命周期的演化过程[8]。相对于科技文献中的专业词汇,网络自由文本中包含了大量同义、近义等具有复杂语义关系的词汇,且词之间的共现频率较低,因此共词分析法并不适用于复杂语义关系的自由文本。

关于话题演化趋势的预测,现有工作多是通过时间序列预测话题热度等指标来分析话题的演化趋势:马晓宁基于粒子群算法优化的 BP 神经网络方法对话题热度进行趋势预测[9];刘晨等融合 LSTM 与卷积神经网络方法预测话题的热度趋势[10]。然而关于话题在未来时刻状态趋势预测的已有研究相对较少。范云满等在 Y. Tu 等[6]的研究基础上新增被引量指标,并利用多项式拟合曲线的方法预测话题状态趋势[11];Kong 等结合与话题相关的各动态因素的贡献和模式匹配的方法,从微观和宏观两个层面探索话题流行度状态在未来的发展趋势[12]。隐马尔可夫模型(Hidden Markov Model,HMM)作为一种成熟的概率统计模型,能考虑时间序列的影响,在描述对象统计特性的动态随机过程上面具有突出优势[13],已经成功应用于手势识别、寿命预测等领域[14-15]。话题状态演化可看作是由话题内部状态和外部观测特征构成的一种双重随机过程,它适用于 HMM 模型,已有少量研究工作将 HMM 运用于话题状态趋势预测中。Zeng 等基于话题内容相似度对舆情话题进行分类,并基于 HMM 构建话题预测模型来预测话题生命周期阶段[16];Liu 等以博文数量和增长率作为观测指标,运用 HMM 对多个话题分别构建状态预测模型并建立模型库,通过人工判别待预测话题与模型库中已有话题是否相似,从而选择相应模型预测话题未来的生命周期阶段[17]。上述工作提出的话题状态预测模型人工干预量和预测误差较大。而且,话题生命周期受多方面因素和偶发情况影响,准确预测话题未来的生命周期状态具有较大的挑战性。

综上所述,目前相关工作多是从生命周期的视角来回溯话题状态的演化过程,对演化中的话题在未来时刻的状态趋势预测研究较少。该文从话题预警的视角,基于向心度和密度指标将演化中的话题划分为不同等级的风险状态;基于 word2vec 模型[18]计量状态指标,解决了共词分析法不能有效处理网络自由文本中的复杂语义这一问题;基于 HMM 提出话题未来时刻的风险状态趋势预测方法,为话题的有效预警提供科学依据。

1 话题风险状态定义

话题状态是对话题当前及潜在影响力的度量,它描述了话题本身的发展趋势和引发舆论危机的风险性。从话题预警的视角将话题状态划分风险等级,可以直观地刻画话题引发舆论危机的风险程度,也是下一阶段话题趋势预测的目标。

Collon 等[7]针对科技文献主题提出向心度和密度两个指标,向心度表示主题与其他主题关联的强弱,向心度越大,该主题越接近议题的“中心”,因此向心度反映了主题的重要性。密度表示构成主题的特征词之间的紧密程度,在主题演化的过程中,主题在内容上从分散逐渐收敛,密度也随着增大,因此密度反映了主题的成熟度。向心度和密度可被借鉴来度量网络话题当前及潜在的影响力,该文采用这两个指标来刻画话题的风险状态,进而对话题可能引发舆论危机的风险等级进行划分。考虑到网络自由文本包含同义词、近义词等复杂语义关系,区别于文献[7-8]中采用共词分析法计算话题向心度和密度,该文基于 word2vec 模型计量两个指标值。

1.1 向心度指标计量

在描述网络话题时,向心度是指一个话题与其他话题关联的强弱程度。向心度越大说明话题与其他话题关联越强,该话题在所有话题中越接近于“中心”位置,越容易受到网民的关注,从而容易引发舆论危机,因此向心度可以反映话题的风险状态。

基于 word2vec 模型,通过计算两话题之间特征词的相似度来衡量话题之间的关联程度,话题与其他话题特征词的相似度越高,话题的向心度值越大。假设 T_i 是基于 LDA 模型提取出的话题,则其可表示成 T_i =

$[(w_{i1},t_{i1}),(w_{i2},t_{i2}),\cdots,(w_{im},t_{im})]$,其中 t_{ik} 和 w_{ik} 分别表示构成话题 T_i 的第 k 个特征词及其权重,m 是特征词的数量;令 v_{ik} 是由 word2vec 模型训练出的特征词 t_{ik} 对应的向量,则两话题 T_i 与 T_j 之间的相似度 $\mathrm{Sim}(T_i,T_j)$ 可由式(1)计算可得。

$$\mathrm{Sim}(T_i,T_j)=\frac{1}{|T_i|\times|T_j|}\sum_{k=1}^{m_i}\sum_{x=1}^{m_j}w_{ik}w_{jx}\times \mathrm{Sim}(v_{ik},v_{jx}) \tag{1}$$

其中,$\mathrm{Sim}(v_{ik},v_{jx})$ 为向量 v_{ik} 和 v_{jx} 之间的余弦相似度,$|T_i|$ 和 $|T_j|$ 分别为话题 T_i 和 T_j 的模,计算公式如下:

$$|T_i|=\sqrt{\sum_{k=1}^{m_i}\sum_{x=1}^{m_i}w_{ik}w_{ix}\times\mathrm{Sim}(v_{ik},v_{ix})} \tag{2}$$

设 T_{set} 为所有话题的集合,即 $T_{\mathrm{set}}=\{T_1,T_2,\cdots,T_n\}$,则话题 T_i 的向心度 C_i 可由话题 T_i 与其他话题之间相似度的均值求得,即:

$$C_i=\frac{1}{n-1}\sum_{T_j\in T_{\mathrm{set}}-|T_i|}\mathrm{Sim}(T_i,T_j) \tag{3}$$

由式(3)可知,话题向心度 C_i 的值域为[0,1]。当 $C_i=1$ 时,表明话题 T_i 与其他所有话题均有强关联,位于最“中心”;当 $C_i=0$ 时,该话题与其他所有话题毫无关联,为整个话题空间的孤立点。

1.2　密度指标计量

在描述网络话题时,密度是指话题内特征词之间的紧密程度。围绕话题展开的讨论越集中,话题会越聚焦,特征词之间的紧密度越高,密度值越大,话题也会趋于成熟。因此,话题的密度反映了话题讨论的集中程度,也是话题风险的表征指标之一。

该文采用话题内部特征词之间的相似度来衡量其紧密程度,话题内特征词之间的相似度越高,话题的密度值越大。话题 T_i 中第 k 个特征词 w_{ik} 与其他特征词之间的平均相似度 A_{ik} 由式(4)计算可得。

$$A_{ik}=\frac{1}{m_i-1}\sum_{j\neq k}^{m_i}\mathrm{Sim}(v_{ik},v_{ij}) \tag{4}$$

话题的密度可用各个特征词的加权平均相似度表示,如式(5)所示。

$$D_i=\frac{1}{|T_i|}\sum_{k=1}^{m_i}w_{ik}A_{ik} \tag{5}$$

由式(5)可知,话题密度 D_i 的值域为[0,1]。D_i 值越大,表明话题 T_i 中的特征词语义越趋于集中;反之,则话题中的特征词语义越趋于分散。

1.3　话题风险等级划分

正如上文所述,向心度和密度分别从不同的角度反映了话题引发舆论危机的风险。借鉴科技文献中划分主题状态的方法——平面坐标映射法,将话题的向心度和密度分别作为平面坐标系的横轴和纵轴,并将两个指标的均值作为坐标原点,则可以把话题的状态空间划分为四个象限,分别对应了话题的四种风险状态,如图 1 所示。

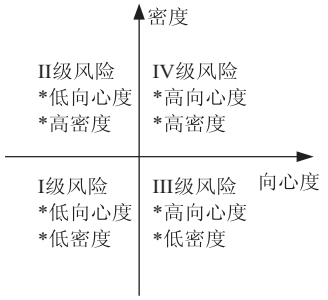

图 1　基于向心度、密度两个特征划分的话题状态类别

(1)I 级风险状态:话题的向心度和密度均较低,意味着该话题与其他话题关联弱,处于议题的边缘位置,且话题讨论分散,不聚焦,难以引发舆论危机,因此该类话题定义为 I 级风险状态。

(2)II 级风险状态:话题的向心度较低,但密度较高,意味着该话题讨论相对聚焦,成熟度高,但与其他话题关联弱,处于边缘位置。一旦引爆只是小面积发酵,难以在全网范围内激起舆论危机,因此该类话题定义为 II 级风险状态。

(3)III 级风险状态:话题的密度较低,但向心度较高,意味着虽然该话题讨论不够聚焦,尚未成熟,但与其他话题关联强。随着围绕该话题展开的讨论增多,话题内容趋向聚焦,很容易在全网范围内引发舆论危机,因此该类话题定义为 III 级风险状态。

(4)IV 级风险状态:话题的向心度和密度均较高,意味着该话题与其他话题关联强,处于议题的“中心”位置,且话题聚焦,讨论集中,极易引发全网范围内的舆论危机,因此该类话题定义为 IV 级风险状态。全民关注的热点与焦点话题往往属于该类风险状态。

相对于 I 级和 II 级风险状态,III 级和 IV 级风险状态话题引发舆论危机的可能性较大,政府和舆论监管部门需要格外关注话题走向,必要时采取预警措施,干预话题进一步扩散,营造良好的网络舆论氛围。

2　话题风险状态预测方法

话题风险状态预测方法是根据当前时刻的话题观测数据预测出下一时刻话题所处的风险状态。话题状态随着时间推移不断演化,虽然无法直接观察到话题状态,但可以通过向心度、密度等观测指标来反映。因此,话题状态演化过程是由外部观测指标反映内部话题状态的双重随机过程,可用隐马尔可夫模型描述。

2.1　模型构建

隐马尔可夫模型是一个双重随机过程,一个过程是描述隐藏状态转移的马尔可夫链,另一个过程是描述隐藏状态与观测状态之间的映射关系[19]。图 2 描

述了一段时间内隐藏状态之间的转移关系及隐藏状态与观测状态之间的对应关系。

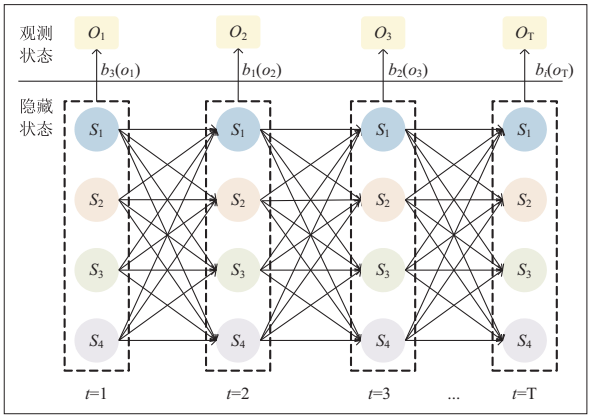

图 2 话题风险状态转移序列与观测序列关系

该文基于 HMM 构建话题风险状态模型,模型参数描述如下:隐马尔可夫模型可用参数 $\lambda = \{\pi, A, B\}$ 来表示,话题风险状态预测模型参数选取及模型训练的初始值设置描述如下:

(1)隐藏状态集合 S:$S = \{s_1, s_2, s_3, s_4\}$,$s_1$、$s_2$、$s_3$、$s_4$ 分别对应话题的 I 级、II 级、III 级、IV 级风险状态,状态数量 $N = 4$。令话题在 t 时刻的状态为 q_t,$q_t \in S$。

(2)观测序列 O:$O = \{o_1, o_2, \cdots, o_t\}$,表示在 $1 \sim t$ 时间段内由话题各时刻二维观测值组成的观测序列,o_t 表示 t 时刻下话题 T 的向心度和密度值组成的二维观测值。

(3)初始状态概率分布 π:$\pi = \{\pi_i\}$,$\pi_i = P(q_t = s_i)$,$1 \leqslant i \leqslant N$。其中,$\pi_i$ 为出现状态 s_i 的概率,满足 $\sum_{i=1}^{N} \pi_i = 1$。HMM 模型初始化参数 π 的初始化选择对模型的最终收敛结果影响不大[20],将初始状态均匀化分布。

(4)状态转移概率矩阵 $\boldsymbol{A}$:$\boldsymbol{A} = \{a_{ij}\}$,$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$,$1 \leqslant i, j \leqslant N$。其中,$a_{ij}$ 为话题风险状态从 s_i 转移到 s_j 的概率,满足 $\sum_{j=1}^{N} a_{ij} = 1$。同样,$\boldsymbol{A}$ 的初始化选择对模型的最终收敛结果影响也不大[20],因此可假设话题在初始时刻处于原来状态和转移到其他状态的概率是相同的。

(5)观测状态概率分布 B:$B = \{b_i(o_t)\}$,$b_i(o_t) = P(o_t \mid q_t = s_i)$。其中,$b_i(o_t)$ 为 t 时刻隐藏状态为 s_i 对应观测状态为 o_t 的概率。当 HMM 的观测值为连续值时,状态 s_i 生成观测状态的概率可以用高斯模型(Gaussian Model,GM)来拟合,即隐藏状态 s_i 对应的观测值服从均值为 u_i、协方差为 Σ_i 的二元高斯概率密度函数,如式(6)所示。该文将话题在四类风险状态下对应的二维观测数据的平均值和协方差作为模型初始均值和协方差。

$$b_i(o_t) = \frac{1}{\sqrt{2\pi\Sigma_i}} \exp\left(-\frac{1}{2}(o_t - u_i)^{\mathrm{T}} \Sigma_i^{-1} (o_t - u_i)\right) \tag{6}$$

文中话题风险状态预测方法是将各个风险状态下对应的观测序列数据作为该状态的表征,分别针对不同的话题风险状态构建 HMM 模型,从而预测话题演化过程中风险状态的变化趋势,它能够避免原有模型[17]对不同类型话题建模导致模型普适性较低的问题,弥补话题生命周期波动性较高带来的模型稳定性较低的不足。根据平面坐标映射方法,提取出各个风险状态下对应的多条观测序列,作为 HMM 模型的训练数据,对四类话题风险状态进行模型训练,以期提高模型稳定性和预测效果。

2.2 模型训练

将风险状态 s_i 下的全部观测样本序列表示为 $O(s_i)$,作为各话题风险状态模型的训练数据,并利用 Baum-Welch 算法(EM 算法)训练模型,得到模型集合为 HMMs = {HMM1, HMM2, HMM3, HMM4},对应参数集 $\Omega = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$,将 EM 算法的最大迭代次数设置为 100,收敛阈值为 0.001,经过多次迭代后得到每个模型的最优重估参数。模型训练过程如图 3 所示。

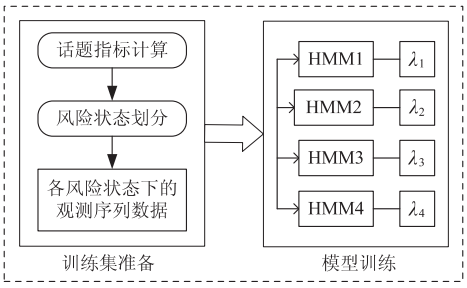

图 3 各个话题状态的模型训练过程

以话题 I 级风险状态的模型训练为例。首先,E 步是函数 $Q(\lambda, \hat{\lambda})$,$Q(\lambda, \hat{\lambda}) = \sum_I \log P(O, I \mid \lambda) P(O, I \mid \hat{\lambda})$。

其中,$\hat{\lambda}$ 是 HMM 模型参数的当前估计值,λ 是要极大化的 HMM 参数,$O = \{o_1, o_2, \cdots, o_t\}$ 是观测状态序列,$I = \{q_1, q_2, \cdots, q_t\}$ 是隐藏状态序列。然后,EM 算法的 M 步是极大化 Q 函数 $Q(\lambda, \hat{\lambda})$,利用拉格朗日乘子法求最优模型参数 λ。

2.3 话题状态预测

对四类话题状态模型训练后,分别得到最优参数 $\lambda_1, \lambda_2, \lambda_3, \lambda_4$。将话题当前时刻的观测值输入各个模型中,根据极大相似准则 $i^* = \arg\max_i P(O \mid \lambda_i)$,$i \in \{1, 2, 3, 4\}$,找出概率值最大的模型,即为最佳模型[21]。由此从当前时刻的状态出发,来进一步预测话

题在下一时刻的状态转移结果。

基于最佳模型的最优重估参数,根据 $o_{t+1}=\sum_{j=1}^{N}A(i,j)E(b_j(o_t))$ 直接预测出话题在 $t+1$ 时刻所对应的观测值,再借助平面坐标映射法判别出话题风险状态的预测结果。

3　实验与分析

3.1　数据来源及预处理

以“疫情、肺炎、新冠”为关键词爬取微博数据,时间跨度为 2019 年 12 月 31 日至 2020 年 5 月 19 日,获得微博数据共 307 932 条。对数据进行清洗、分词等预处理,运用 LDA 算法进行话题识别,采用主题一致性指标确定最佳话题数为 6。实验以周为时间单位,计算每个话题在时间跨度为 20 周的向心度和密度指标值,得到总共 120 条数据。将 120 条数据映射到坐标系中,获得属于 I 级、II 级、III 级、IV 级风险状态的观测序列数据分别为 24 条、35 条、24 条和 37 条。

3.2　实验结果分析

观测数据尽管不多,但基本上反映了国内疫情大范围爆发那段时期的微博话题讨论情况。实验采取 K 折交叉验证法(K-fold Cross Validation)对实验结果进行评估,K 取值为 4。将 120 条数据序列分成 4 等份,每次都取其中的 3 份(90 条观测数据)作为训练集,取剩下的 1 份(30 条观测数据)作为测试集。如此循环 4 次,在每一次交叉验证中,利用训练集数据中属于各风险状态的观测数据对各个状态训练 HMM 模型,再利用测试集数据进行状态预测。

表 1　采用话题风险状态方法的混淆矩阵

实际风险状态	预测风险状态				
	I 级	II 级	III 级	IV 级	预测准确率/%
Ⅰ级	23	0	0	1	95.83
Ⅱ级	1	34	0	0	97.14
Ⅲ级	1	2	19	0	86.36
IV 级	3	0	1	30	88.24

该文采用 $t+1$ 时刻的二维观测数据预测值与实际值的误差来评估模型预测效果,选取平均绝对百分比误差(Mean Absolute Percentage Error,MAPE)作为模型预测精度的评价指标。MAPE 的计算方式如式(7)所示。

$$\mathrm{MAPE}=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{R_{t+1}-P_{t+1}}{R_{t+1}}\right|\times 100\% \tag{7}$$

其中,n 为预测次数,R_{t+1} 为 $t+1$ 时刻的实际值,P_{t+1} 为 $t+1$ 时刻的预测值。经过计算得出,模型预测的向心度值 MAPE 为 14.13%,密度值 MAPE 为 11.99%。向心度与密度的实际值与预测值对比如图 4 所示,其中两个指标的预测值与实际值趋势一致,相比向心度,密度值的预测误差更小。

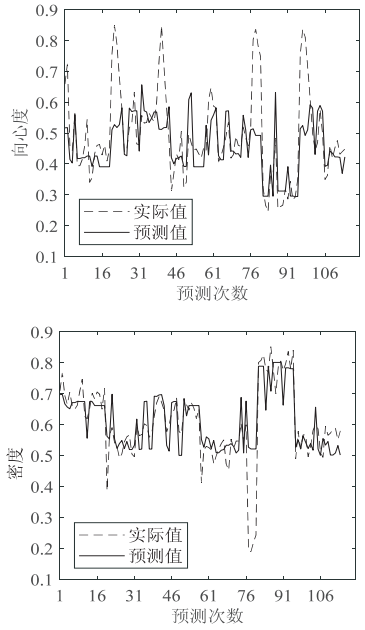

图 4　话题状态预测模型的向心度和密度预测值与实际值对比

根据预测出的 $t+1$ 时刻观测值判别话题风险状态后,得出话题风险状态预测的混淆矩阵,如表 1 所示。该方法预测风险状态的平均准确率为 92.11%,其中,III 级和 IV 级风险状态更具现实意义,两种状态预测准确率均达到 86% 以上,说明该预测方法能够有效捕捉话题引发舆论危机的风险性。

3.3　对比验证

为验证该研究方法的准确性和有效性,采用 BP 神经网络(BPNN)模型、LSTM 模型、RNN 模型进行对比实验。选取数据预处理得到的 6 个话题前 10 周观测值为训练集,将后 10 周观测值作为测试集评估预测效果。实验采用精确率、召回率与 F1 值进行模型评估,结果如表 2 所示。

表 2　实验模型效果对比

模型	精确率/%	召回率/%	F1 值/%
HMM	92.56	88.07	90.26
BPNN	83.65	84.92	84.28
LSTM	86.67	88.37	87.51
RNN	85.34	83.48	84.40

从实验结果可以看出,对于文中的话题数据集,HMM、BPNN、LSTM 和 RNN 模型得到的准确率、召回率和 F1 值均高于 80%。其中,HMM 模型得到的话题风险状态预测的 F1 值达到 90.26%,相较于适用较大数据量的神经网络模型,HMM 模型在预测话题风险状态时更具有优势。

4 结束语

为了预测处于演化过程中的话题状态趋势,从话题预警的视角,基于向心度和密度指标将演化中的话题划分为不同等级的风险状态,为话题状态划分提供了新思路。由于话题状态演化过程是由外部观测指标反映内部话题状态的双重随机过程,该文基于 HMM 提出话题风险状态预测方法,以新冠肺炎疫情事件为例进行了验证。实验结果表明,该方法预测风险状态的平均准确率为 92.11%,相对于 BP 神经网络、LSTM 以及 RNN 时间序列预测模型,该方法预测话题风险状态的误差更小。基于 HMM 的话题风险状态预测方法为舆情监管部门及时预警话题风险性提供了科学依据。

参考文献:

[1] 吴 迪,张梦甜,生 龙,等. 改进在线词对主题模型的微博热点话题演化[J]. 计算机工程与应用,2021,57(24):179-184.

[2] 王振飞,刘凯莉,郑志蕴,等. 面向时间序列的微博话题演化模型研究[J]. 计算机科学,2017,44(8):270-273.

[3] CHEN C C,CHEN Y T,MENG C C. An aging theory for event life-cycle modeling[J]. IEEE Transactions on Systems,Man,and Cybernetics - Part A:Systems and Humans,2007,37(2):237-248.

[4] 贾亚敏,安 璐,李 纲. 城市突发事件网络信息传播时序变化规律研究[J]. 情报杂志,2015,34(4):91-96.

[5] 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.

[6] TU Y N,SENG J L. Indices of novelty for emerging topic detection[J]. Information Processing & Management,2012,48(2):303-325.

[7] CALLON M,COURTIAL J P,LAVILLE F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemsitry[J]. Scientometrics,1991,22(1):155-205.

[8] 刘自强,王效岳,白如江. 多维主题演化分析模型构建与实证研究[J]. 情报理论与实践,2017,40(3):92-98.

[9] 马晓宁,王 惠. 基于 PSO 优化 BP 神经网络的话题趋势预测[J]. 计算机工程与设计,2018,39(9):2907-2911.

[10] 刘 晨,刘 超. 融合循环神经网络与卷积神经网络的话题趋势预测方法[J]. 工业控制计算机,2022,35(8):102-104.

[11] 范云满,马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报,2014,33(7):698-711.

[12] KONG Q,MAO W,CHEN G,et al. Exploring trends and patterns of popularity stage evolution in social media[J]. IEEE Transactions on Systems,Man,and Cybernetics:Systems,2018,50(10):3817-3827.

[13] PETRIE B T. Statistical inference for probabilistic functions of finite state Markov chains[J]. Annals of Mathematical Statistics,1966,37(6):1554-1563.

[14] 刘珠峰,周 良,丁秋林. 基于隐性马尔可夫模型的手势识别设计和优化[J]. 计算机应用研究,2011,28(6):2386-2388.

[15] 刘文溢,刘勤明,叶春明,等. 基于改进退化隐马尔可夫模型的设备健康诊断与寿命预测研究[J]. 计算机应用研究,2021,38(3):805-810.

[16] ZENG J,ZHANG S,WU C,et al. Predictive model for internet public opinion[C]//Fourth international conference on fuzzy systems and knowledge discovery. Haikou:IEEE,2007:7-11.

[17] LIU R F,GUO W B. HMM-based state prediction for internet hot topic[C]//Proceedings of 2011 IEEE international conference on computer science and automation engineering. Shanghai:IEEE,2011:240-244.

[18] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16)[2016-09-07]. https://arxiv.org/abs/1301.3781.

[19] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE,1989,77(2):257-286.

[20] WELCH L. Hidden Markov models and the Baum-welch algorithm[J]. IEEE Information Theory Society Newsletter,2003,53(4):10-13.

[21] COUTROT A,HSIAO J H,CHAN A B. Scanpath modeling and classification with hidden Markov models[J]. Behav Res Methods,2018,50(1):362-379.