基于隐半马尔可夫模型的微博流行信息检测方法

谢柏林，黎琦，邝建

引用本文

谢柏林，黎琦，邝建. 基于隐半马尔可夫模型的微博流行信息检测方法[J]. 计算机科学, 2022, 49(6A): 291-296.

XIE Bai-lin, LI Qi, KUANG Jiang. Microblog Popular Information Detection Based on Hidden Semi-Markov Model[J]. Computer Science, 2022, 49(6A): 291-296.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800011

基于隐半马尔可夫模型的微博流行信息检测方法

谢柏林　黎　琦　邝　建
广东外语外贸大学信息科学与技术学院　广州 510006
广东外语外贸大学网络空间安全学院　广州 510006

摘　要　目前微博已成为人们发布信息和获取信息的一个重要平台。为了及早发现微博上的流行信息,以便及时发现微博上的热点事件,同时及时发现、抑制谣言信息的传播,使微博在网民的信息获取和信息发布中发挥更积极的作用,文中提出了一种基于隐半马尔可夫模型的微博流行信息检测方法。该方法以信息转发者的影响力等级和相邻两个转发者的时间间隔构建观测值,使用随机森林分类算法来自动得到转发者的影响力等级,利用隐半马尔可夫模型来刻画流行信息的传播过程,基于此来及早发现潜在的流行信息。该方法分为模型训练和流行信息检测两个阶段,在流行信息检测阶段,计算每条信息在传播过程中产生的观测序列相对于模型的平均对数似然概率,实时更新每条信息的流行度。使用采集的新浪微博数据集和 Twitter 数据集对所提方法进行了测试,实验结果表明了该方法的有效性。

关键词:微博;流行信息;隐半马尔可夫模型;流行度;传播过程
中图法分类号　TP391

Microblog Popular Information Detection Based on Hidden Semi-Markov Model

XIE Bai-lin,LI Qi and KUANG Jiang
School of Information Science and Technology,Guangdong University of Foreign Studies,Guangzhou 510006,China
School of Cyber Security,Guangdong University of Foreign Studies,Guangzhou 510006,China

Abstract　In recent years,microblog has become great places for people to communicate with each other and share knowledge. However,microblog has also become the main grounds for rumors' transmission. If we can identify popular information in early stage,then we can identify and quell rumors early,we can also identify hot topics early in microblog. Therefore,the research on popular information detection is important. In this paper a new method is presented for identifying popular information based on hidden semi-Markov model(HSMM),from the perspective of the transmission processes of popular information in microblog. In this method,the observation value is constructed based on the influence level of the information forwarder and the time interval between two adjacent forwarders,and the influence level of the forwarder is automatically obtained by using the random forest classification algorithm. The proposed method includes a training phase and an identification phase. In the identification phase,the average log likelihood of every observation sequence is calculated,and the popularity of information is updated in real time. So this method can identify the popular information in early stage. An experiment based on real datasets of Sina Weibo and Twitter is conducted to evaluate this method. The experiment results validate the effectiveness of this method.

Keywords　Microblog,Popular information,Hidden semi-Markov model,Popularity,Transmission process

1　引言

微博是一种基于用户关系的信息分享、传播以及获取平台,已成为人们生活中不可缺少的一部分。在微博上,用户主要通过发布博文(即简短文本)来实现信息的即时分享。用户在发布博文时无需长篇大论,因此微博具有较低的准入门槛。借助电脑、手机等设备,任何一个微博用户都可以在任何时间、任何地点随意地发布信息。微博信息主要依靠用户的转发进行传播,微博信息具有极快的传播速度[1-2]。信息发布的便捷性、信息的及时性和信息传播的快速性,使得微博已成为网民获取信息、发布信息的重要渠道。

然而,由于微博上的博文字数简短,比较难全面客观地传递信息,另外由于微博信息在传播过程中缺乏强有力的把关人(Gatekeeper)[3],以及有些用户为了吸引眼球喜欢故意发布、转发一些耸人听闻的谣言,导致微博成为谣言滋生的温床。例如 2020 年新冠肺炎疫情期间,微博上充斥着大量谣言和虚假信息。在新冠肺炎疫情所带来的社会恐慌面前,人们更愿意相信这些耸人听闻的虚假消息,甚至成为此类有害信息的主动转发者。这类有害信息不仅会造成社会恐慌,也会影响一国民众的认知模式和社会心态。

截至 2020 年 3 月,新浪微博日活跃用户数已达 2.41 亿,平均每天新增的博文超 2 亿条。在每天新增的海量信息中,

基金项目:广东省基础与应用基础研究基金(2018A0303130045);广州市科技计划项目(201904010334).
This work was supported by the Guangdong Basic and Applied Basic Research Foundation(2018A0303130045) and Science and Technology Program of Guangzhou(201904010334).
通信作者:谢柏林(bailinxie@gdufs.edu.cn)

绝大部分信息都不会在微博上广泛传播，例如普通用户发布的绝大部分信息。如果能在信息传播的前期识别出微博上的潜在流行信息，那么可以及时发现微博上的热点事件，进而也可以及时发现、抑制微博上谣言的传播，减小谣言对公众的影响，使微博在网民的信息获取和信息发布中发挥更积极的作用。因此，微博流行信息检测方法的研究就显得有必要。

由于微博信息在传播过程中具有很强的随机性和突发性，现有方法很难在信息传播的前期有效识别出潜在流行信息，为此本文将采用隐半马尔可夫模型来刻画流行信息的传播过程，基于此来实时评估微博信息的流行度，从而及早发现潜在的流行信息，该方法能在信息传播的前期就能识别出潜在的流行信息。

本文第 2 节介绍了相关研究；第 3 节介绍了微博流行信息检测的原理；第 4 节给出了实验结果；最后总结全文并展望未来。

2 相关研究

近年来，学者们在微博流行信息检测方面取得了一些成果[19,22-25,27-30]。Hong 等[4]提出利用博文的内容特征、网络的拓扑特征、信息传播的时间特征和元数据特征，并采用 Logistic 回归算法来预测某条信息是否会被转发，以及该信息在不久的将来会被转发多少次，在该方法中，博文的内容特征选为博文的话题分布，网络的拓扑特征选为节点的度分布、局部聚类系数等，信息传播的时间特征选为信息相邻两次被转发的平均时间间隔等，元数据特征选为当前时刻信息是否已被转发过，以及发布者发布的信息被转发了多少次。Bandari 等[5]利用博文所属的类型、所用的语言、信息出处等特征，使用回归算法和分类算法来识别 Twitter 上潜在的流行信息。Naveed 等[6]提出利用博文的内容特征，采用 Logistic 回归分类算法来识别微博上潜在的流行信息，在该方法中，博文的内容特征选为博文是否提及其他用户、博文中是否含有标签、博文中是否含有 URL、博文中是否含有“!”、博文中是否含有“?”、博文中是否含有正面的情感词、博文中是否含有负面的情感词、博文是否表达了发布者的积极情绪、博文是否表达了发布者的消极情绪、博文所表达的观点以及博文的话题特征等。Peng 等[7]提出利用博文内容特征、网络特征和时间特征，基于条件随机场模型来检测流行信息，在该方法中，博文内容特征选为博文的话题相似性、博文中是否含有 URL、博文中是否含有标签以及博文是否提及其他用户；网络特征选为信息发布者的背景（即发布者粉丝的数量、发布者发布的博文数量等），以及信息发布者与潜在转发者之间的社会关系；时间特征选为用户响应博文的时间。Gao 等[8]利用信息发布后 1 小时内的传播特征，基于分类算法来识别潜在的流行信息。Zhu 等[9]提出了一种多元线性回归模型，然后基于该模型来预测微博信息的流行度。Bao 等[10]利用信息传播前期的网络特征来识别潜在的流行信息。Gao 等[11]提出了一种强化的泊松过程模型，然后基于该模型来检测微博流行信息。Cao 等[12]根据信息流行度变化趋势对信息进行分组，然后训练得到每个组所对应的模型，最后基于这些模型来识别微博上的流行信息。Gao 等[13]对现有的微博信息流行度预测方法进行了对比分析和总结。另外，Wang 等[14]基于多元线性回归方程和社会学中的连接强度，提出了了一种针对 Facebook 知名主页的消息流行度预测模型。Xie 等[15]提出了一种基于多模变分编解码器的短视频流行度预测框架。Liu 等[20]利用模糊理论和神经网络来预测用户的转发行为，Yin 等[21]利用深度学习算法来预测信息的转发时间。

上述方法在识别潜在流行信息时的效果不太理想，尤其很难在信息传播的前期有效识别出潜在流行信息，这主要是因为微博信息在传播过程中具有很强的随机性和突发性。本文将采用隐半马尔可夫模型（Hidden semi-Markov Model，HSMM）[16-17]来刻画流行信息的传播过程，基于此来实时评估微博信息的流行度，从而及早发现潜在的流行信息，该方法能在信息传播的前期就能识别出潜在的流行信息。

3 微博流行信息检测方法

微博信息主要通过用户转发进行传播，当某个用户在微博上发布一条信息后，该信息的传播过程大体如图 1 所示（不考虑用户多次转发的情况）。其中，F 表示信息发送者，R^1，R^2，R^3，…，R^{15} 表示信息转发者，转发者可能是发送者的粉丝或粉丝的粉丝，也可能是其他用户，这些用户通过主页或搜索引擎发现了信息，然后对信息进行转发。图 1 中，圆圈表示用户的影响力，圆圈越大表示用户的影响力越大。

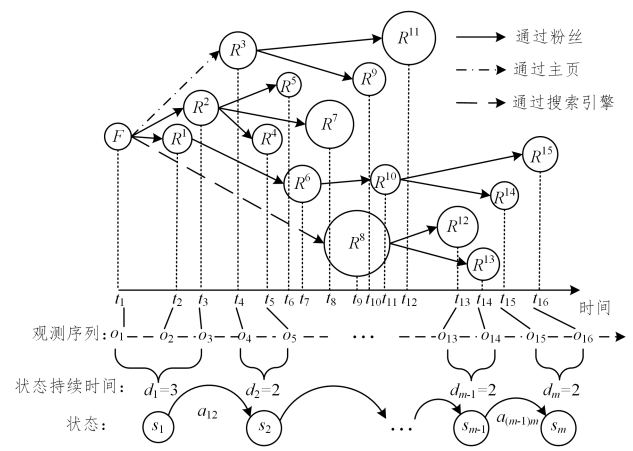

图 1 微博信息传播模型

Fig. 1 Transmission model of microblog information

微博流行信息在传播过程中的传播速度和转发者的影响力通常会发生变化（例如流行信息在传播的前期、中期和后期，其传播速度是不一样的），导致相邻两次转发的时间间隔、转发者影响力的统计分布会发生变化。可以把相邻两次转发的时间间隔和转发者影响力的不同统计分布作为状态，假设不同状态的个数为 M，即 s_1，s_2，…，s_M。状态的转移过程则可以看作是一个马尔可夫过程，即当前状态只与前一个状态有关。

令 $O=o_1o_2\cdots o_t=O_{t\to t}$ 表示某条流行信息在传播过程中产生的一个长度为 t 的二维观测序列，其中 $o_t=(x_t,h_t)$ 表示第 t 个观测值，x_t 表示第 t 个转发者的影响力等级（x_1 表示信息发布者的影响力等级），h_t 表示第 t 个转发者与前一个转发者之间的时间间隔，h_t 取的是离散化的整数值，其中 $h_1=0$，$h_t\in\{0,1,2,3,\cdots\}$，本文中 h_t 的时间单位为 s。假设转发者的影响力可分为 Y 个等级，则 $1\leqslant x_t\leqslant Y$。在现有研究和实验结果的基础上，本文使用用户粉丝的数量、用户关注的用户数量、用户是否为认证用户、用户的性别、用户发表/转发的博文数量、用户注册时的时间，并基于分类算法来自动判别转发者

的影响力等级。

当某条信息在微博上传播时，本文使用分类算法来自动获得第 t 个转发者 Z_t 的影响力等级 x_t，其步骤如下：

步骤 1　在微博上采集大量流行信息在传播过程中出现的转发者的相关数据，然后提取出转发者的特征值，即粉丝的数量、关注的用户数量、是否为认证用户、性别、发表/转发的博文数量、注册的时间；

步骤 2　基于特征值，使用 k-means 聚类算法把步骤 1 中的转发者聚成 Y 个不同的簇，然后对每个簇中的转发者进行人工检验和筛选，使得每一个簇对应一种影响力等级，且同一簇中转发者的影响力等级都相同，最后把同一簇中的转发者都标记相同的类标号；

步骤 3　使用步骤 2 中已标记的数据来训练和测试基于不同分类算法的分类器，然后根据测试结果选择出最佳分类器；

步骤 4　从 Z_t 的个人资料中提取出特征值，然后使用训练好的最佳分类器对 Z_t 进行分类，从而得到 x_t 的取值。

由于观测值与状态不具有一一对应的关系，即给定一个观测值 o_t，我们不能直接得到此时的状态，因此这是一个隐马尔可夫过程。令 $\boldsymbol{A}=\{a_{mn};1\leqslant m,n\leqslant M\}$ 是状态转移概率矩阵，它的元素 a_{mn} 表示状态从 s_m 跳转到 s_n 的概率。由于 h_t 主要与转发者发现信息的时间和网络的传输时延有关，而 x_t 主要与用户粉丝的数量、用户关注的用户数量、用户是否为认证用户、用户的性别、用户发布/转发的博文数量等有关，因此可以近似假定，对于给定的状态 s_m，x_t 和 h_t 是相互独立的。令 $\boldsymbol{B}=\{b_m(y,\varphi);1\leqslant m\leqslant M\}$ 为观测值概率矩阵，它的元素 $b_m(y,\varphi)$ 表示对于在给定状态 s_m 下，当观测值 $x_t=y$，$h_t=\varphi$ 时的概率，其中 $1\leqslant y\leqslant Y$，$0\leqslant\varphi$。$b_m(y,\varphi)$ 可表示为：

$$\begin{aligned}b_m(y,\varphi)&=\Pr[x_t=y,h_t=\varphi|q_t=s_m]\\&=\Pr[x_t=y|q_t=s_m]\times\Pr[h_t=\varphi|q_t=s_m]\\&=b_m^1(y)\times b_m^2(\varphi)\end{aligned}\tag{1}$$

其中，q_t 表示第 t 个转发者出现时模型所处的状态。

令 $\pi=\{\pi_m;1\leqslant m\leqslant M\}$ 为初始状态概率矩阵，它的元素 π_m 表示流行信息在被发布时模型状态为 s_m 的概率。令 $\boldsymbol{P}=\{p_m(d);1\leqslant m\leqslant M,1\leqslant d\leqslant D\}$ 为状态持续时间概率矩阵，它的元素 $p_m(d)$ 表示在状态 s_m 下，连续出现 d 个转发者的概率，其中 D 为状态持续的最大时间。流行信息的传播过程是由多种因数决定的，因此状态持续时间可能是一个比较复杂的分布，不一定是几何分布，而在隐马尔可夫模型中，状态持续时间必须服从几何分布，因此流行信息在传播过程中，模型状态转移过程实际上是一个隐半马尔可夫过程。图 1 给出了某条流行信息在传播过程中的状态转移过程，其中 o_t 为第 t 个观测值，s_1，s_2，s_{m-1}，s_m 为隐半马尔可夫模型的状态，a_{12}，$a_{(m-1)m}$ 为状态之间的转移概率，d_1，d_2，d_{m-1}，d_m 为状态持续的时间。

隐半马尔可夫模型是在隐马尔可夫模型（Hidden Markov Model，HMM）[18] 的基础上发展起来的，其模型状态持续时间可以为任意分布，更适合于描述非稳态和非 Markovian 分布的随机过程。因此，隐半马尔可夫模型的应用范围比隐马尔可夫模型更广。

本文提出的微博流行信息检测方法可分为模型训练和流行信息检测两个阶段。

（1）模型训练阶段

基于文献[17]中的前向-后向算法（Forward-Backward Algorithm），采用多序列来训练隐半马尔可夫模型。在微博上，采集大量流行信息在传播过程中产生的观测序列，将其作为模型训练的数据集。假设训练数据集为 Ω，包含 L 个不同的观测序列，即 $\Omega=\{O^{(l)},l=1,2,\cdots,L\}$，其中 $O^{(l)}=o_{1\to T_l}^{(l)}$ 为第 l 个观测序列，T_l 为对应序列的长度。

令 λ 表示模型的参数集，即 $\lambda=\{a_{mn},b_m^1(y),b_m^2(\varphi),\pi_m,p_m(d)\}$。参数更新的计算式如式(2)－式(6)所示。在式(3)中，当 $x_t^{(l)}=y$ 时，$\delta(x_t^{(l)}-y)=1$，否则 $\delta(x_t^{(l)}-y)=0$；在式(4)中，当 $h_t^{(l)}=\varphi$ 时，$\delta(h_t^{(l)}-\varphi)=1$，否则 $\delta(h_t^{(l)}-\varphi)=0$。其中，$x_t^{(l)}$，$h_t^{(l)}$ 分别表示观测序列 $O^{(l)}$ 中的第 t 个转发者的影响力等级和时间间隔。

$$\bar{a}_{mn}=\frac{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{t=2}^{T_l}\xi_t^{(l)}(m,n)]}{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{n=1}^{M}\sum_{t=2}^{T_l}\xi_t^{(l)}(m,n)]}\tag{2}$$

$$\bar{b}_m^1(y)=\frac{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{t=1}^{T_l}\gamma_t^{(l)}(m)\times\delta(x_t^{(l)}-y)]}{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{y=1}^{Y}\sum_{t=1}^{T_l}\gamma_t^{(l)}(m)\times\delta(x_t^{(l)}-y)]}\tag{3}$$

$$\bar{b}_m^2(\varphi)=\frac{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{t=1}^{T_l}\gamma_t^{(l)}(m)\times\delta(h_t^{(l)}-\varphi)]}{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{\varphi}\sum_{t=1}^{T_l}\gamma_t^{(l)}(m)\times\delta(h_t^{(l)}-\varphi)]}\tag{4}$$

$$\bar{\pi}_m=\frac{\sum_{l=1}^{L}\frac{1}{\chi_l}\times\gamma_1^{(l)}(m)}{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{m=1}^{M}\gamma_1^{(l)}(m)]}\tag{5}$$

$$\bar{p}_m(d)=\frac{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{t=1}^{T_l}\eta_t^{(l)}(m,d)]}{\sum_{l=1}^{L}\frac{1}{\chi_l}\times[\sum_{d=1}^{D}\sum_{t=1}^{T_l}\eta_t^{(l)}(m,d)]}\tag{6}$$

其中，$\xi_t^{(l)}(m,n)$ 为状态跳转联合概率，它表示在观测序列为 $O^{(l)}$ 的基础上，当第 t 个转发者出现时，模型从状态 s_m 跳转到状态 s_n 的概率，$\xi_t^{(l)}(m,n)$ 的定义如式(7)所示，具体计算式如式(8)所示，$q_t^{(l)}$ 表示在观测序列 $O^{(l)}$ 中第 t 个转发者出现时模型所处的状态。

$$\xi_t^{(l)}(m,n)\equiv\Pr[O^{(l)},q_{t-1}^{(l)}=s_m,q_t^{(l)}=s_n]\tag{7}$$

$$\xi_t^{(l)}(m,n)=\frac{\alpha_{t-1}^{(l)}(m,1)a_{mn}b_n^1(x_t^{(l)})b_n^2(h_t^{(l)})}{(\sum_d p_n(d)\beta_t^{(l)}(n,d))^{-1}}\tag{8}$$

其中，$\gamma_t^{(l)}(m)$ 为状态和观测值联合概率，它表示在观测序列为 $O^{(l)}$ 的基础上，当第 t 个转发者出现时，模型此刻正处于状态 s_m 的概率。$\gamma_t^{(l)}(m)$ 的定义如式(9)所示：

$$\gamma_t^{(l)}(m)\equiv\Pr[O^{(l)},q_t^{(l)}=s_m]\tag{9}$$

当 $1\leqslant t<T_l$ 时，$\gamma_t^{(l)}(m)$ 的计算式如式(10)所示，当 $t=T_l$ 时，$\gamma_{T_l}^{(l)}(m)=\sum_d\alpha_{T_l}^{(l)}(m,d)$。

$$\gamma_t^{(l)}(m)=\gamma_{t+1}^{(l)}(m)+\sum_{n\neq m}[\xi_{t+1}^{(l)}(m,n)-\xi_{t+1}^{(l)}(n,m)]\tag{10}$$

其中，$\eta_t^{(l)}(m,d)$ 为状态持续联合概率，它表示在观测序列为 $O^{(l)}$ 的基础上，当第 t 个转发者出现时，模型从其他状态跳转到状态 s_m 且将在状态 s_m 下连续出现 d 个转发者的概率。$\eta_t^{(l)}(m,d)$ 的定义和计算式分别如式(11)、式(12)所示，$\tau_t^{(l)}$ 表示在观测序列 $O^{(l)}$ 中模型在状态 $q_t^{(l)}$ 下将连续出现的转发者的数量。

$$\eta_t^{(l)}(m,d)\equiv \Pr[O^{(l)}, q_{t-1}^{(l)}\neq s_m, q_t^{(l)}=s_m, \tau_t^{(l)}=d] \tag{11}$$

$$\eta_t^{(l)}(m,d)=\frac{\sum_{n\neq m}\alpha_{t-1}^{(l)}(n,1)a_{nm}}{(b_m^1(x_t^{(l)})b_m^2(h_t^{(l)})p_m(d)\beta_t^{(l)}(m,d))^{-1}} \tag{12}$$

其中，$\alpha_t^{(l)}(m,d)$为观测序列 $O^{(l)}$ 中的前向变量，表示对第 l 个观测序列而言，在观测到前 t 个转发者的基础上，当第 t 个转发者出现时，模型处于状态 s_m 且还将在状态 s_m 下连续出现 d 个转发者的概率。其定义如式(13)所示。

$$\alpha_t^{(l)}(m,d)\equiv Pr[o_{1\to t}^{(l)}, q_t^{(l)}=s_m, \tau_t^{(l)}=d] \tag{13}$$

当 $t=1$ 时，$\alpha_1^{(l)}(m,d)=\pi_m b_m^1(x_1^{(l)})b_m^2(h_1^{(l)})p_m(d)$，当 $t\neq 1$ 时，$\alpha_t^{(l)}(m,d)$的计算式如式(14)所示。

$$\alpha_t^{(l)}(m,d)=\alpha_{t-1}^{(l)}(m,d+1)b_m^1(x_t^{(l)})b_m^2(h_t^{(l)})+[\sum_{n\neq m}\alpha_{t-1}^{(l)}(n,1)a_{nm}]\times b_m^1(x_t^{(l)})b_m^2(h_t^{(l)})p_m(d) \tag{14}$$

其中，$\beta_t^{(l)}(m,d)$为观测序列 $O^{(l)}$ 中的后向变量，表示对第 l 个观测序列而言，在第 t 个转发者出现时模型处于状态 s_m 且还将在状态 s_m 下连续出现 d 个转发者的条件下，产生观测序列 $o_{t+1\to T_l}^{(l)}$ 的概率。其定义如式(15)所示。

$$\beta_t^{(l)}(m,d)\equiv \Pr[o_{t+1\to T_l}^{(l)} \mid q_t^{(l)}=s_m, \tau_t^{(l)}=d] \tag{15}$$

当 $t=T_l$ 时，$\beta_{T_l}^{(l)}(m,d)=1$；当 $d=1$，且 $t\neq T_l$ 时，$\beta_t^{(l)}(m,1)$的计算式如式(16)所示；当 $d>1$，且 $t\neq T_l$ 时，$\beta_t^{(l)}(m,d)$的计算式如式(17)所示。

$$\beta_t^{(l)}(m,1)=\sum_{n\neq m}\frac{a_{mn}b_n^1(x_{t+1}^{(l)})b_n^2(h_{t+1}^{(l)})}{(\sum_d p_n(d)\beta_{t+1}^{(l)}(n,d))^{-1}} \tag{16}$$

$$\beta_t^{(l)}(m,d)=b_m^1(x_{t+1}^{(l)})b_m^2(h_{t+1}^{(l)})\beta_{t+1}^{(l)}(m,d-1) \tag{17}$$

其中，χ_l 为训练集中观测序列 $O^{(l)}$ 相对于模型的平均对数似然概率(Average Log Likelihood)，为观测序列的加权系数，其计算式如式(18)所示。

$$\chi_l=\log\{\Pr[O^{(l)}\mid\lambda]\}/T_l=\log(\sum_m\sum_d\alpha_{T_l}^{(l)}(m,d))/T_l \tag{18}$$

模型训练结束后，得到$\{a_{mn}, b_m^1(y), b_m^2(\varphi), \pi_m, p_m(d)\}$的值，即得到流行信息在微博上的传播模型。

(2)流行信息检测阶段

当某个用户在微博上发布某条信息时，按照以下步骤来判别该信息是否为流行信息：

步骤 1　令 $t=1$，计算出 x_1 的值，记录该信息被发布的时间 c_1，此时 $h_1=0$；

步骤 2　当该信息被转发时，则令 $t=t+1$，并计算出 x_t 的值，记录该信息被转发的时间 c_t，统计出 h_t 的值；

步骤 3　求出观测序列 $o_{1\to t}$，然后计算观测序列 $o_{1\to t}$ 相对于模型的平均对数似然概率 Θ，Θ 的计算式如式(19)所示。为了提高 Θ 的计算速度，我们假定模型状态持续时间概率 $p_m(d)$为 Gamma 分布，并采用文献[26]中的方法来训练得到 $p_m(d)$的值。

$$\Theta=\log\{\Pr[o_{1\to t}\mid\lambda]\}/t=\log[\sum_{m=1}^{M}\sum_{d=1}^{D}\alpha_t(m,d)]/t \tag{19}$$

步骤 4　如果 Θ 大于某个阈值，则认为该信息是潜在的流行信息，退出循环，从而实现对流行信息的及早检测；否则跳转到步骤 2。

上述循环过程中，Θ 的取值反映了信息的流行程度，我们把 Θ 作为微博信息的流行度。

4　实验测试及结果分析

我们使用采集到的新浪微博数据集和 Twitter 数据集来测试本文提出的流行信息检测方法。在新浪微博的热门微博排行榜中找到 6835 条流行信息，另外通过新浪微博的搜索功能找到 5426 条非流行信息，然后使用新浪微博的 API 对这些信息进行采集。在数据采集过程中，把信息传播过程中出现的发布者和转发者的相关资料都采集下来，保存于数据库中，将其作为训练测试集。采集到的新浪微博数据的时间跨度为 2013 年 12 月－2017 年 12 月。

另外，使用 Twitter 的 API 从 Twitter 上采集到了一些流行信息和非流行信息，在 Twitter 数据采集过程中，通过 Twitter 的搜索功能，搜索与某些热门事件相关的流行信息，例如搜索关键词设置为“Britain left”等，“Britain left”对应的是英国脱欧事件。我们共采集到 6532 条流行信息和 5517 条非流行信息。使用 Twitter 的 API，得到这些信息传播过程中出现的所有发布者和转发者的相关资料。采集到的 Twitter 数据的时间跨度为 2014 年 1 月－2017 年 12 月。

4.1　转发者影响力等级自动判别方法测试

参照目前流行的网络购物商品评价星级制度，我们把转发者影响力等级分为 5 级。原因是：如果转发者影响力的等级类别过多，则人工在标注转发者影响力等级时其准确率较难保证，如果转发者影响力等级类别过少，则很难反映出不同转发者影响力的差异。

我们从采集到的数据集中随机挑选出大量转发者，并提取出每个转发者的特征值，即转发者粉丝的数量、关注的用户数量、是否为认证用户、性别、发表/转发的博文数量、注册的时间。基于转发者的上述特征值，使用 k-means 聚类算法把这些转发者聚成 5 个不同的簇，然后对每个簇中的转发者进行人工检验和筛选，使得每一个簇对应一种影响力等级，且同一簇中转发者的影响力等级都相同，接着把同一簇中的转发者都标记相同的类标号，即 Class_1，Class_2，Class_3，Class_4 和 Class_5。最后，从已标记的数据集中通过随机抽取样本的方式来生成测试数据集，使得测试数据集包含 5 个类别，并且每个类别包含 5000 个转发者。

基于上述测试数据集，我们采用交叉验证方法，分别测试了随机森林(Random Forest，RF)分类算法、C4.5 分类算法、朴素贝叶斯(Naïve Bayesian，NB)分类算法、支持向量机(SVM)分类算法在自动判别转发者影响力等级时的性能，测试实验所用主机的配置为：CPU 为 Intel 酷睿 i7-9700(八核心，主频 3GB)，内存为 16GB。RF，C4.5，NB 和 SVM 分类算法的精度(Precision)和召回率(Recall)如表 1 所列。

表 1　4 种分类算法的性能

Table 1　Performance of four classification algorithms

分类算法	精度	召回率
RF	0.98	0.97
C4.5	0.94	0.95
NB	0.89	0.88
SVM	0.99	0.98

RF，C4.5，NB 和 SVM 分类算法的建模时间和测试时间如表 2 所列。由表 1 可知，SVM 分类算法的准确率略高于 RF 分类算法，然而由表 2 可知，SVM 分类算法的测试时间多于 RF 分类算法。当流行信息在微博上传播时，短时间内会出现大量转发者，此时判别转发者影响力等级的速度就显得非常重要。由表 2 可知，RF 分类算法的测试时间最短，非常适合在线分类，因此本文采用 RF 分类算法来自动判别转发者的影响力等级。

表 2 分类算法的建模时间和测试时间

Table 2 Modeling and testing time of classification algorithms

（单位：s）

分类算法	建模时间	测试时间
RF	1.25	0.036
C4.5	0.17	0.029
NB	0.08	0.974
SVM	1.03	1.121

4.2 流行信息检测方法测试

我们使用本文提出的转发者影响力等级自动判别方法对采集到的新浪微博数据集进行预处理，得到信息传播过程中产生的观测序列，最终得到 6835 条流行信息产生的观测序列和 5426 条非流行信息产生的观测序列。

本文模型的 ROC（Receiver Operating Characteristic）曲线如图 3 所示，当模型的假正率（False Positive Rate，FPR）为 5%时，模型的真正率（True Positive Rate，TPR）为 99.5%。TPR 和 FPR 的计算式如式（20）、式（21）所示。

$$TPR=\frac{\text{正确识别出的流行信息数量}}{\text{流行信息的总数量}} \tag{20}$$

$$\text{FPR}=\frac{\text{非流行信息误判为流行信息的数量}}{\text{非流行信息的总数量}} \tag{21}$$

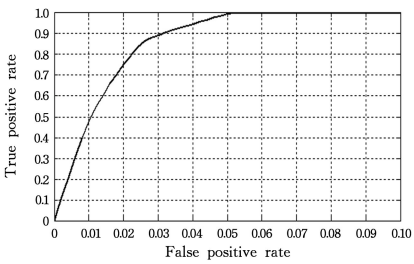

图 3 新浪微博数据集下模型的 ROC 曲线

Fig. 3 Model's ROC curve under Sina Weibo dataset

为了进一步测试本文提出的流行信息检测方法，我们使用本文提出的转发者影响力等级自动判别方法对采集到的 Twitter 数据集进行预处理，得到信息传播过程中产生的观测序列。本文模型的 ROC 曲线如图 4 所示，当模型的假正率为 5%时，模型的真正率为 98.3%。

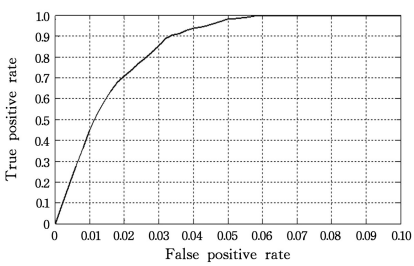

图 4 Twitter 数据集下模型的 ROC 曲线

Fig. 4 Model's ROC curve under Twitter dataset

Twitter 用户的地理分布比新浪微博用户的地理分布更广阔，因此 Twitter 上的流行信息具有更强的突发性。另外，用于模型训练的新浪微博数据集的规模大于 Twitter 数据集，因此本文提出的流行信息检测方法在新浪微博数据集上的性能优于 Twitter 数据集上的性能。

从上述测试实验的结果可知本文方法具有较好的性能。

4.3 对比测试

我们使用采集到的新浪微博数据集和 Twitter 数据集对本文方法和文献[4-6]中的方法进行了对比测试，新浪微博数据集下的精度（Precision）和召回率（Recall）如表 3 所列，Twitter 数据集下的测试结果如表 4 所列，其中 HSMM 代表本文中的方法，Hong 代表文献[4]中的方法，Bandari 代表文献[5]中的方法，Naveed 代表文献[6]中的方法。文献[4]是基于博文的内容特征、网络的拓扑特征、信息传播的时间特征和元数据特征，采用 Logistic 回归算法来检测微博流行信息。文献[5]是基于博文所属的类型、所用的语言、信息出处等特征，使用回归算法和分类算法来检测微博流行信息。文献[6]是基于博文的内容特征，采用 Logistic 回归分类算法来检测微博流行信息。由表 3、表 4 可知，本文方法具有更好的性能。

表 3 新浪微博数据集下的对比测试

Table 3 Comparison test under Sina Weibo dataset

（单位：%）

测试方法	精度	召回率
HSMM	99.5	99.2
Hong	86.3	85.1
Bandari	75.5	73.2
Naveed	81.2	80.4

表 4 Twitter 数据集下的对比测试

Table 4 Comparison test under Twitter dataset

（单位：%）

测试方法	精度	召回率
HSMM	98.3	97.2
Hong	84.3	82.8
Bandari	72.8	71.5
Naveed	80.4	79.6

文献[4-6]中的方法通过利用微博的一些特征，例如博文的内容特征、网络的拓扑特征等，基于回归算法或分类算法来检测微博上的流行信息，这些方法主要从静态的角度来识别流行信息，未充分考虑流行信息的动态传播过程，而信息的动态传播过程更能体现信息的流行趋势，因此文献[4-6]中的方法在识别潜在的流行信息时其准确率不太理想，其中文献[5]中的方法使用的特征最少，其性能表现最不理想。本文方法从信息传播的动态过程来识别潜在的流行信息，该方法具有很高的精度和召回率。

结束语 本文提出了一种基于隐半马尔可夫模型的微博流行信息检测方法，该方法以信息转发者的影响力等级和相邻两个转发者的时间间隔构建观测值，使用随机森林分类算法来自动得到转发者的影响力等级，采用隐半马尔可夫模型来刻画流行信息的传播过程，基于此来识别微博上潜在的流行信息。我们使用采集的新浪微博数据集和 Twitter 数据集对本文方法进行了测试，实验结果表明该方法具有很高的精度和召回率。

未来将使用更多的微博数据集来测试该方法的性能，以及在实际微博平台上在线测试该方法的性能。

参 考 文 献

[1] YE S, WU S F. Measuring message propagation and social influence on Twitter. com[C]// Proceedings of the Second International Conference on Social Informatics. 2010: 216-231.

[2] GUILLE A, HACID H, FAVRE C, et al. Information diffusion in online social networks: A survey[J]. ACM Sigmod Record, 2013, 42(2): 17-28.

[3] WESTERMAN D, SPENCE P R, VAN DER HEIDE B. A social network as information: The effect of system generated reports of connectedness on credibility on Twitter[J]. Computers in Human Behavior, 2012, 28(1): 199-206.

[4] HONG L, DAN O, DAVISON B D. Predicting popular messages in twitter[C]// Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011: 57-58.

[5] BANDARI R, ASUR S, HUBERMAN B A. The pulse of news in social media: Forecasting popularity[C]// Sixth International AAAI Conference on Weblogs and Social Media. 2012: 26-33.

[6] NAVEED N, GOTTRON T, KUNEGIS J, et al. Bad news travel fast: A content-based analysis of interestingness on twitter [C]// Proceedings of the 3rd International Web Science Conference. ACM, 2011: 1-7.

[7] PENG H K, ZHU J, PIAO D, et al. Retweet modeling using conditional random fields[C]// 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011: 336-343.

[8] GAO S, MA J, CHEN Z. Popularity prediction in microblogging network[C]// Asia-Pacific Web Conference. Cham: Springer, 2014: 379-390.

[9] ZHU H L, YUN X C, HAN Z S. Weibo Popularity Prediction Method Based on Propagation Acceleration[J]. Journal of Computer Research and Development, 2018, 55(6): 1282-1293.

[10] BAO P, SHEN H W, HUANG J, et al. Popularity prediction in microblogging network: a case study on sina weibo[C]// Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 177-178.

[11] GAO S, MA J, CHEN Z. Modeling and predicting retweeting dynamics on microblogging platforms[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 107-116.

[12] CAO Q, SHEN H, GAO H, et al. Predicting the popularity of online content with group-specific models[C]// Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee. 2017: 765-766.

[13] GAO X, CAO Z, LI S, et al. Taxonomy and Evaluation for Microblog Popularity Prediction[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2019, 13(2): 15-54.

[14] WANG X M, FANG B X, ZHANG H L, et al. TSL: predicting popularity of Facebook content based on tie strength[J]. Journal on Communications, 2019, 40(10): 1-9.

[15] XIE J Y, ZHU Y C, ZHANG Z B, et al. A Multimodal Variational Encoder-Decoder Framework for Micro-video Popularity Prediction[C]// Proceedings of the Web Conference 2020. 2020: 2542-2548.

[16] YU S Z. Hidden semi-Markov models[J]. Artificial intelligence, 2010, 174(2): 215-243.

[17] YU S Z, KOBAYASHI H. An efficient forward-backward algorithm for an explicit-duration hidden Markov model[J]. IEEE Signal Processing Letters, 2003, 10(1): 11-14.

[18] 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.

[19] ZHOU F, XU X, TRAJCEVSKI G, et al. A survey of information cascade analysis: Models, predictions, and recent advances [J]. ACM Computing Surveys(CSUR), 2021, 54(2): 1-36.

[20] LIU Y, ZHAO J, XIAO Y. C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network[J]. Neurocomputing, 2018, 275: 733-746.

[21] YIN H, YANG S, SONG X, et al. Deep fusion of multimodal features for social media retweet time prediction[J]. World Wide Web, 2020, 24(4): 1027-1044.

[22] ROY S, SUMAN B K, CHANDRA J, et al. Forecasting the Future: Leveraging RNN based Feature Concatenation for Tweet Outbreak Prediction[C]// Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. 2020: 219-223.

[23] LYMPEROPOULOS I N. RC-Tweet: Modeling and predicting the popularity of tweets through the dynamics of a capacitor[J]. Expert Systems with Applications, 2021, 163: 113785.

[24] XIAO C, LIU C, MA Y, et al. Time sensitivity-based popularity prediction for online promotion on Twitter[J]. Information Sciences, 2020, 525: 82-92.

[25] ZHANG Z, YIN Z, WEN J, et al. DeepBlue: Bi-layered LSTM for tweet popularity Estimation[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2021. https://ieeexplore.ieee.org/abstract/document/9314897.

[26] XIE Y. An efficient algorithm for parameterizing HsMM with Gaussian and Gamma distributions[J]. Information Processing Letters, 2012, 112(19): 732-737.

[27] CHEN L, DENG H. Predicting User Retweeting Behavior in Social Networks With a Novel Ensemble Learning Approach[C]// IEEE Access. 2020: 148250-148263.

[28] SHANG J, HUANG S, ZHANG D, et al. RNe2Vec: information diffusion popularity prediction based on repost network embedding[J]. Computing, 2021, 103(2): 271-289.

[29] CAO Q, SHEN H, GAO J, et al. Popularity prediction on social platforms with coupled graph neural networks[C]// Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 70-78.

[30] ZHOU F, YU L, XU X, et al. Decoupling Representation and Regressor for Long-Tailed Information Cascade Prediction[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1875-1879.

XIE Bai-lin, born in 1982, Ph. D, assistant professor, is a member of China Computer Federation. His main research interests include online social network, network security.