微博中话题的传播分析及热点预测

侯凯，苏菲，庄伯金
（北京邮电大学信息与通信工程学院，北京 100876）

摘要：微博在舆论传播中的作用日益凸显，有很多微博是围绕同一相关事件而展开讨论的，这便构成了一个话题。研究话题的传播规律并对热点话题进行有效预测，有助于我们了解事件的发展动态及趋势，可用于新闻热点的挖掘和不良信息的监管等领域。论文将话题传播过程划分为三个阶段（即潜伏期、爆发期与消退期），采用 VIP 用户比例、用户粉丝数分布等四个特征刻画话题传播，取得了较好的实验结果；并提出了时序信号表示话题趋势的方法，最后采用了基于“潜在基底”假设的分类模型对话题进行热点预测。实验结果表明 83.5%的话题可以提前预测，平均提前时间约为 1.6 小时，验证了算法的有效性。

关键词：人工智能；微博话题；分类模型；热点预测
中图分类号：TP181

Hotspot Prodiction and Analysis of Propagating of Topics Based on Microblog

HouKai, SuFei, Zhuang Bojin
(School of Information and Communication Engineering ,Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: Microblog is playing a more and more important role in spread of public opinion. The contents of many microblogs are around the same event, which constitute a topic. The hotspot prediction and propagating analysis of topics could help people keep up with the latest trend about events, which can be used in the mining of news and the regulation of the bad information. We divide the propagating process into three stages: incubation, outbreak and recession period, use the four features to describe the propagation, such as the proportion of VIP, distribution of the fans number, etc. We propose a classification method based on a latent source model, and use the timing signal to represent the trends of topics. The result shows that, we can detect trends before Sina does 83.5% of the time, with a mean value of 1.6 hour in advance. The effectiveness of the method is demonstrated.

Key words: artificial intelligence; topic on microblog; classification method;hot prediction

0 引言

在社交领域，微博服务占据了重要的地位，它打破了传统熟人社交的形式，鼓励用户以简短的内容发布消息，并提供了简捷的转发服务。在微博平台上，用户可以关注任何自己感兴趣的用户，并转发他们的消息呈现给自己的朋友，大大降低了“参与门槛”，有效地促进了消息的传播。

微博的转发功能使得信息在用户群体中迅速扩张，传播速度呈几何增长。研究话题的发展趋势，可以反映出相关事件的发展状况，对商业营销、政治活动等具有重要的指导价值，可以给商家、活动主办方的未来策略提供良好的参考信息。同时，微博作为新兴的传播媒体，由于平台控制不规范、监管漏洞，微博中难免存在欺诈、虚假、非法信息。对某些话题进行实时检测，对热点话题进行有效地预警，可以为网络监管提供很大的帮助。

在大量的微博中，有很多微博是围绕某一相关事件展开讨论的，这些微博簇构成了不同

作者简介：侯凯（1990-），男，硕士研究生，主要研究方向：模式识别
通信联系人：苏菲（1973-），女，教授，主要研究方向：模式识别、图像处理等. E-mail: sufei@bupt.edu.cn

的话题。如 2014 年 7 月 14 日巴西世界杯决赛，当天世界杯相关微博数近 3500 万条，新浪微博中参与“世界杯德国夺冠”话题讨论的用户超过 11 万人。每时每刻，微博中都有大量的话题存在，有新的话题产生，旧话题的湮灭。微博话题往往围绕特定事件产生，分析话题的发展，可以更有效地了解网络舆论及事件动态。

虽然基于微博的研究越来越多，但由于微博的社交特性、媒体特性，导致话题的传播呈现混沌状态，人们对话题传播机制的理解依然不够。此外，研究话题的传播规律，可以使我们更加了解微博传播的影响因素，为话题预测提供更好的理论支持。

1 相关研究

随着微博在舆论传播中的作用日益凸显，人们对微博的研究工作也日益增多，但是目前为止，针对话题传播及话题热点预测的工作依然较少。

1.1 话题的传播分析

对话题的传播规律进行合理分析有助于对话题趋势的研究。田野[1]使用回归模型对特定事件的关注度趋势及情感极性趋势进行了回归预测；马社祥[2]从小波分解、SVR 等角度对非平稳时间序列进行了建模。尽管回归、拟合模型对某些事件的拟合程度较好，但是其拟合参数不具有普适性，而且难以将这些方法推广到话题的预测领域。

有很多工作从传播动力学的角度对话题（微博）的传播规律进行了分析。韩忠明[3]等基于 SIR（易感者-感染者-恢复者）模型，不同用户对同一话题具有不同的敏感程度，通过为不同用户分配不同的感染系数进行仿真，拟合传播模型。Wang Hao[4]等考虑到用户多次转发、外部场强（转发者从其他媒体了解到事件，非粉丝用户转发原创者微博）等特殊情况，对 SIR 模型变型得到拟合效果更好的传染病模型。传染病模型的基础假设是病毒（信息）的自由扩散，但是，微博中“关注-被关注”的层次网络结构在很大程度上限制了消息初期的自由传播。而且这种方法也很难应用于话题预测。

兰月新[5]在研究突发事件模型中，把事件的发展阶段分为各个阶段，并对各个阶段进行分别分析。张婧[6]首先分析了话题传播的特点：病毒式爆发、名人效应、关键点传播等，然后通过预测微博转发链上关键点的转发行为及转发量，预估微博下一时间窗的转发数，从而进行话题预警。这些对话题传播特性的基础研究就有较好的普适性，总结了所有话题的共性，对后期的预测工作具有一定的指导意义。

1.2 热点预测研究

在对话题的预测工作研究中，国外对 twitter 研究较多，目前针对国内微博的话题趋势的研究才刚刚起步。Manish[7]等提出了子主题的共现聚类方法，进而使用最近历史（24h）的话题微博特征构建当前时刻的话题向量，对下一时间段的关注度变化比进行预测。Stanislav[8]针对 twitter 数据集，关注于时间序列的二分类问题，考虑是否可以根据充足的历史样例来判断当前话题能否成为热点。基于他提出的数学模型，可以比 twitter 网站平均提前 1.4 小时检测到热点话题，并且具有较高的准确率。

虽然国内微博和 twitter 在产品形式上有诸多不同，话题的传播特性也可能存在较大的区别，但 Stanislav 的算法思想值得借鉴，将微博话题转化为时间信号，使用历史热点及非热点话题的特征，建立起对当前话题趋势预测的非参数化模型。

2 话题的传播分析

通过研究话题传播过程的规律，可以发现话题传播过程中的影响因素，有助于提高话题热点预测的精度。

2.1 话题传播阶段的划分

为了分析影响话题传播的因素，需要分析话题传播过程中各种特征的变化特点及规律。首先要明确话题的各个传播阶段的含义，通过区分话题生命周期的各个阶段，可以更好地分析话题的特征变化。然后研究话题传播中，不同阶段用户特征及微博特征的变化。本文以“湖南校车落水”和“阿航失联客机坠毁”两个事件为例，说明话题的传播规律。

2014 年 7 月 10 日下午 5 时左右，湖南湘潭市一幼儿园校车在送孩子回家的途中，不慎翻入水塘。11 日凌晨 3 时许，涉事校车被打捞上岸，确认造成 11 人遇难。事件在微博中引起了人们对校车安全的广泛关注。2014 年 7 月 24 日，阿尔及利亚航空公司一架从瓦加杜古飞往阿尔及尔的 AH5017 航班于凌晨 1:07 离开瓦加杜古国际机场，但起飞 50 分钟后，飞机在位于马里空域失去联络。并与当天下午 6 时左右确认坠毁，机上乘客全部遇难。这两个话题都属于当时社会讨论的热点，而且随着事件后续的顺利解决以及新事件的产生，事件吸引人们关注的时间较短，可以分析它们整个生命周期过程中各个因素的变化规律。

网络中信息的传播大致服从 S 型的传播规律[5]。起初阶段，事件刚刚发生，消息只在发布者和他的受众用户间传播。随着得到消息的群体的扩大以及一些媒体或名人的介入，信息得到了爆炸式的传播。但是随着其他事件的爆发，消息推陈出新，人们会去关注更为新鲜的资讯报道，信息传播进入平稳期，并随着人们关注度的下降最终消亡。

上述传播过程的第一个阶段，我们称之为潜伏期，此时消息正在网络中酝酿；第二个阶段是爆发期，消息吸引了大量用户的关注并得到了快速广泛的传播；最后随着事件的结束，消息传播逐渐减弱并慢慢接近消逝，进入第三个阶段消退期。微博中消息的传播也服从网路消息传播的一般性规律，话题的传播阶段如图 1 所示，其中横轴表示时间，纵轴表示消息在人群中获得的累计关注度，对应到微博中，纵轴可表示为同一话题下各个时刻用户累计发布的微博数目，原点表示话题出现的时间。

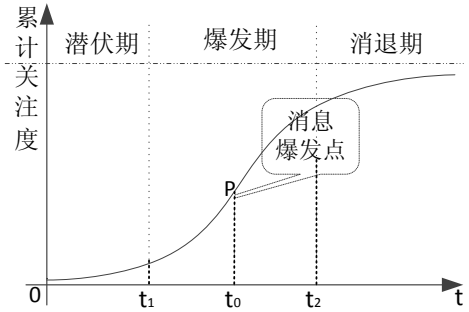

图 1　微博话题的传播阶段图

Fig. 1　The propagation stages of topic

在消息传播的数学模型中，假设某事件发生后，关于该事件的累计关注度是关于时间的连续可微函数，即 $f = f(t)$。网络中消息累计关注度的初值（零时刻）设为 s_0，消息传播的累计关注度的最高上限为 K。累计关注度的变化量与累计数量本身以及消息传播的剩余空间成正比[5]，即：

$$\frac{df}{dt} = r \times f \times (1 - \frac{f}{K}) \tag{1}$$

其中，$r > 0$ 表示正向增长率。上式中综合考虑了这两个因子的影响，求解微分方程得到：

$$f(t) = \frac{K}{1 + \left(\frac{K}{s_0} - 1\right) e^{-r \times t}} \tag{2}$$

可以根据消息传播的历史数据确定参数 K 和 r，并通过计算二阶导 $f''(t) = 0$ 以及 $f'''(t) = 0$ 来确认传播模型的三个关键时间点 t_0、t_1、t_2。这几个时间点将生命周期划分为三个阶段，这里定义话题出现到 t_0 这段时间为潜伏期。此后消息逐渐开始广泛传播，t_0 到 t_2 这段时间为爆发期，其中 t_1 时刻消息的传播速率达到最大，称之为爆发点。t_2 时刻以后，消息的传播速度逐渐下降，进入消退期。

将上述数学模型应用于上述两个话题，用它们的历史数据拟合曲线求解模型参数。“湖南校车落水”话题的拟合确定系数 $R^2 = 0.9812$，“阿航失联客机坠毁”话题的确定系数 $R^2 = 0.9828$，表明拟合程度较好。两个事件的阶段划分结果如图 2 所示，

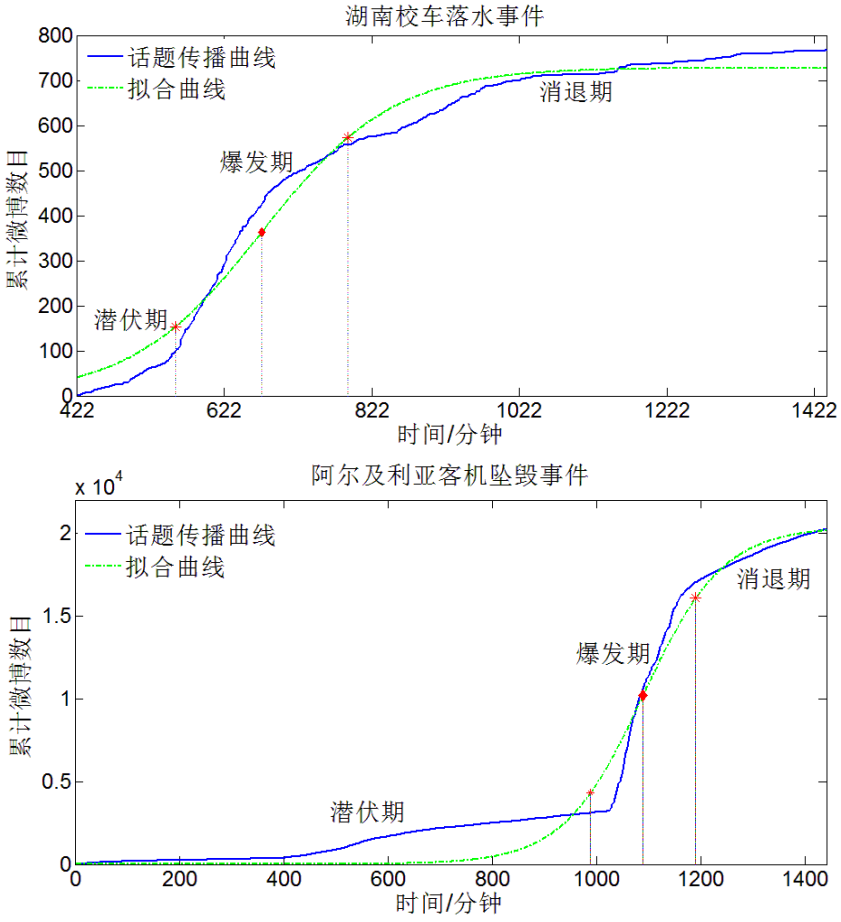

图 2　话题的传播阶段，其中上方为“湖南校车落水”事件；下方为“阿航失联客机坠毁”事件

Fig. 2　The propagation stages of two topics. Top: "school bus overboard in hunan" event. Bottom: "airliner crash of airalgerie" event

2.2　话题传播中的特征分析

话题传播的不同阶段受众群体有着不同的分布，同时微博的内容是影响消息扩散的重要

因素，从用户和内容两方面挖掘在传播过程中变化显著的因素，可用于话题传播的描述。研究发现：在话题传播的不同阶段，用户特征中变化比较明显的特征有：VIP 用户的比例和用户粉丝数的分布；内容特征中变化比较明显的特征有：#话题标签的比例和 url 外链的比例。

微博平台为用户提供了 VIP 会员服务，VIP 用户具有众多特权，这部分用户对微博的黏着度较高，一般也较为活跃。“湖南校车落水”和“阿航失联客机坠毁”话题在各个阶段 VIP 用户的比例如图 3 所示。截止到 2013 年，新浪微博的会员用户约为 1110 万，月活跃用户 1.29 亿，注册用户超过 5.36 亿，会员用户约占总用户量的 2.1%。

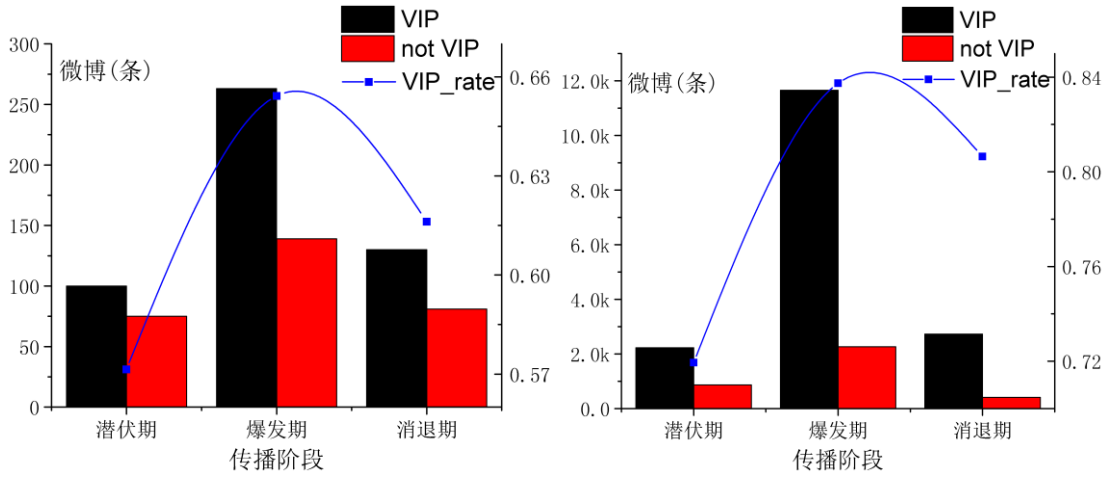

图 3　各个阶段 VIP 用户发布微博的比例，左侧为“校车落水”话题，右侧为“客机坠毁”话题

Fig. 3　The proportion of VIP in each stage. Left: ”school car” topic. Right: “airliner crash” topic

在话题的微博中，有超过一半的微博是 VIP 用户发布的，一方面 VIP 用户比较活跃，平均发布的微博条数要高于普通用户；另一方面，VIP 用户在微博平台中较为积极活跃，他们更关注时事，更热爱分享传播信息。观察两个话题各个阶段，发现在爆发期 VIP 发布微博数目比重要高于潜伏期，而后在消退期所占比例有所下降。在爆发期，大量 VIP 用户的参与促进了消息的传播。

用户的粉丝数是用户的重要属性，粉丝数是用户影响力的重要体现，也是影响微博传播的重要特征。通过观察数据整体分布，可将用户按照粉丝数 500、5000 和 100k 阈值划分为四个区间。分析“湖南校车落水”和“阿航失联客机坠毁”两个话题的实验结果如图 4 所示。

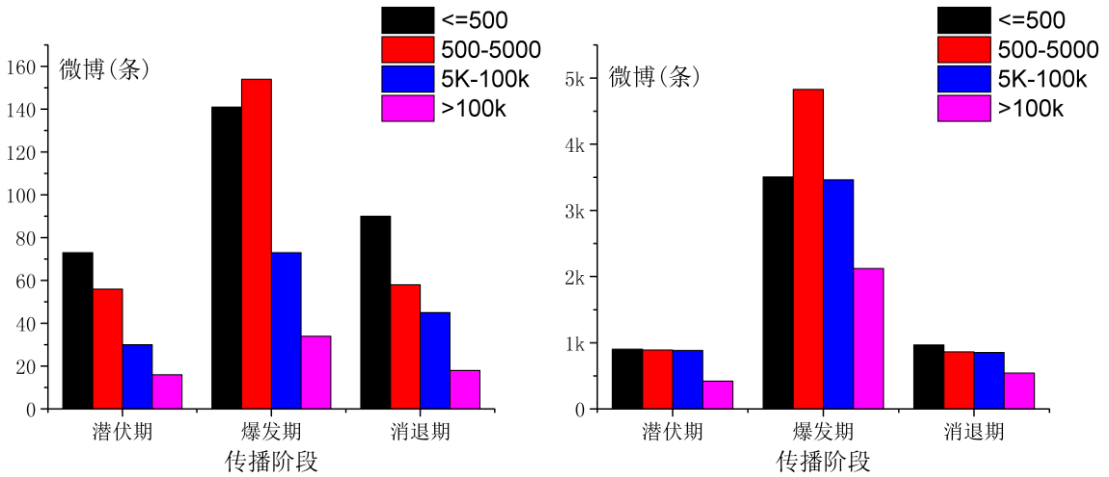

图 4　各阶段用户粉丝数分布图，左侧为“湖南校车落水”话题，右侧为“阿航失联客机坠毁”话题

Fig. 4　The distribution of fans number in each stage. Left: ”school car” topic. Right: “airliner crash” topic

新浪微博中可以通过#号来指定讨论的话题，通过指定微博内容所属话题，有利于吸引别

人来参与讨论，而新加入讨论的用户在发布微博是为了指明自己的讨论主旨，也很可能会添加#的话题标志。实验分析了“湖南校车落水”和“阿航失联客机坠毁”两个话题的不同阶段，微博外链数的分布规律，结果如图 5 所示。

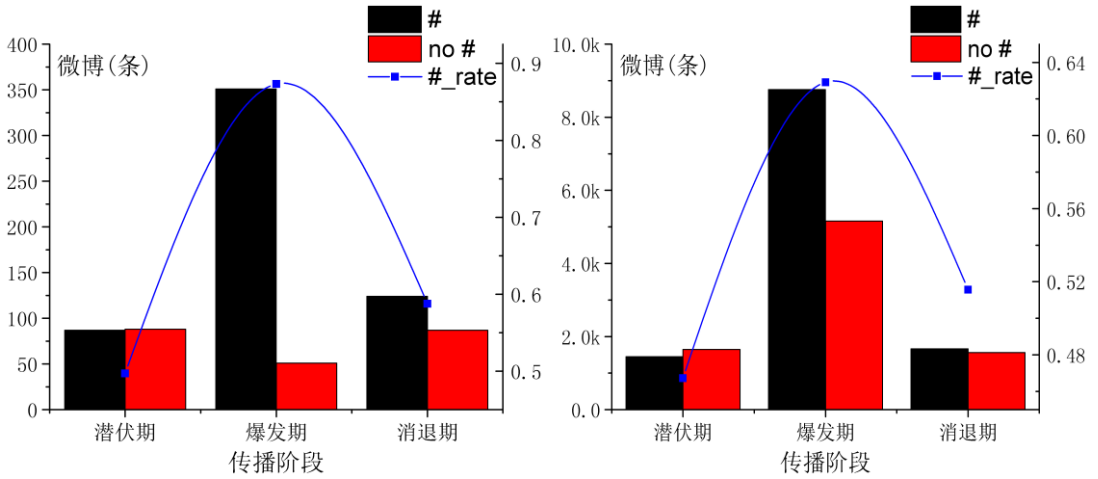

图 5　各阶段包含#话题的微博分布，左侧为“湖南校车落水”话题，右侧为“阿航失联客机坠毁”话题

Fig. 5　The proportion of topic flag in each stage. Left: ”school car” topic. Right: “airliner crash” topic

在话题潜伏期用户数目较少，随着用户的广泛参与，事件被提炼出简单的话题标签，此后人们在发表关于事件的讨论时往往会添加该标签，导致在爆发期包含#话题标志的微博所占比例有明显上升。在消退期，随着人们对改话题关注度的下降，讨论更为泛泛，包含“#话题#”的微博比例有所下降。

微博中可以包含话题相关事件详细情况的 url 外部链接，而且用户在发布原始微博时可能会附加消息来源的网页 url，来表明消息的可靠性。这些 url 通常是传统的新闻媒体网站，比如凤凰网、新浪新闻以及一些博客。微博中 url 外部链接的数量在一定程度上反映了微博外的网络世界中对话题的讨论热度[7]。实验分析了“湖南校车落水”和“阿航失联客机坠毁”两个话题的不同阶段，微博外链数的分布规律，结果如图 6 所示。

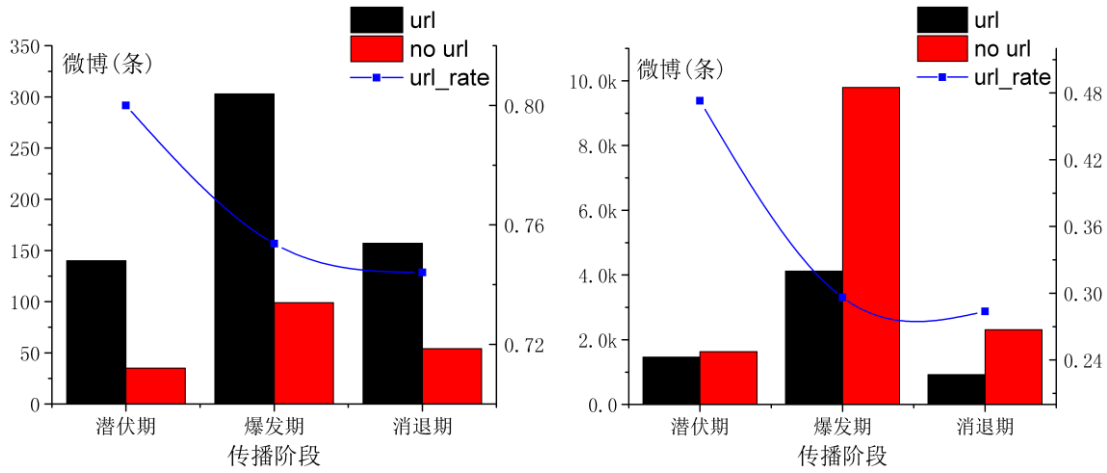

图 6　各阶段包含外链的微博分布，左侧为“湖南校车落水”话题，右侧为“阿航失联客机坠毁”话题

Fig. 6　The proportion of external links in each stage. Left: ”school car” topic. Right: “airliner crash” topic

在话题潜伏期包含外链的微博所占比例相对后面两个时期较高，主要是因为一方面，有很多门户网站会获得一手的新闻资讯，然后消息才得以在微博平台传播；另一方面，博客和新闻报道并没有字数的限制，它们对事件的描述较为详细，事件爆发初期，用户更倾向于在微博中添加外部链接来体现消息的准确性和可靠性。

2.3 话题趋势特征的构建

话题特征是随时间变化的数据，当前时刻话题的影响力可以用截至目前话题下包含的累计微博数来表示。当前时刻最近一段时间的话题走势对话题状态的变化具有重要影响，Li Kuang[9]和 M. Gupta[7]在研究微博中消息传播中均考虑了历史趋势特征。为了研究话题的变化趋势，可以将话题各时刻的特征用时间序列来表示，即构建话题的趋势特征相关的时间信号。

首先根据话题下微博的发布时间，按照一定的时间长度间隔将微博划分为各个子集。试验中采用的时间间隔 τ 为两分钟，$\rho[n]$ 表示第 n 个时间区间内包括的微博数，即 $2(n-1)\sim 2n$ 时间段内，话题下发布的微博数目。截止到 $t\cdot\tau$ 时刻，话题下累计微博数为：

$$v[t]=\sum_{r\le t}\rho[r] \tag{3}$$

因此，$\rho[r]$ 是累计微博数 $v[t]$ 关于时间的离散导数，即 $\rho[t]=v'(t)$。$\rho[t]$ 可以体现出微博数目在第 t 个时间区间的变化率。

有些热门话题的微博数较多，和非热门话题的微博数完全不在一个数量级上，需要数据的规范化，一方面通过数据规范化，使不同话题序列信号特征落在可比的区间内，利于距离的计算；另一方面，由于爬虫抓取能力的变化，规范化后更能真实反映话题在某时刻具有的讨论程度。为了更加真实地反映话题在各个时刻的传播趋势，需要选择一个基准来进行数据的规范化，定义规划化的基准为同一时间区间内，爬虫获取的总体微博数。

$$b[n]=o[n]+\sum_{i=1}^{N}\rho_i[n] \tag{4}$$

其中，$\rho_i[n]$ 表示第 i 个话题在第 n 个区间的微博数目，$o[n]$ 表示 n 时刻不属于任何话题的微博数。规范化后的话题信号为：

$$\rho_b[n]=\left(\frac{\rho[n]}{b[n]}\right)^{\beta}\cdot c \tag{5}$$

指数参数 β 控制了规范化的奖罚程度,试验中设置 $\beta=1$。为了使规范化后信号的值更加合理，更具有可比性，添加了平衡因子 c，当 $c=1$ 时，话题信号的值域是 $[0,1]$，c 越大规范后话题曲线的分离程度越直观，默认设置 c 等于 1000。

热点话题和非热点话题除了微博总量的差异，更为显著的区别是话题信号曲线中峰值的数量和幅度，热点话题传播过程中各个传播阶段的区别较大，一般存在微博数的跳跃变化，并且随着事件的发展会产生多次关于话题的讨论，相比非热点话题其信号曲线波动更为明显，峰值出现次数更多而且幅度更大。

试验中我们强调信号曲线的峰值突变，便于区分波动信号曲线与那些传播过程平滑，峰值变化不明显的曲线，定义新的话题信号量[8]：

$$\rho_{b,s}=\left|\rho_b[n]-\rho_b[n-1]\right|^{\kappa} \tag{6}$$

其中，ρ_b 表示规范化后的数据，参数 κ 控制了对峰值突变的激励程度，根据经验选取 κ 等于 1.2。

如果简单考虑相邻时刻时间序列的变化，难免产生误差。在时间序列曲线中，为了消除外界噪声，需要对信号进行平滑处理，选取平滑窗口的大小为 N_{smooth}，平滑处理后信号的

计算公式为：

$$\rho_{b,s,c}[n]=\sum_{m=n-N_{smooth}+1}^{n}\rho_{b,s}[m] \tag{7}$$

经平滑处理后，信号曲线更能反映话题的总体发展趋势，设置 N_{smooth} 等于 10。

由于微博平台中“关注-被关注”的用户关系网络结构，信息在微博中的流通可以视作为分支过程。虽然我们无法准确知道消息传播的分支细节，但是可以确定微博传播初期的受众群体数目是呈指数规律的。因此在信号处理的最后一步，可以对信号取对数，既可以是话题数据变化更加平稳，而又保持数据的性质和关系，反应信息的真实传播过程。信号的最终计算方法为：

$$\rho_{b,s,c,l}[n]=\log\left(\rho_{b,s,c}[n]+\xi\right) \tag{8}$$

其中 $\rho_{b,s,c}[n]$ 表示平滑处理后的信号，ξ 是较小的正数，设置其大小为 0.0001。

为了保证分类预测的准确性和可靠性，取热点话题曲线中以其爆发时刻为中心的 $2\cdot T_w$ 长度的子串，作为热点参考集。由于非热点话题在整个传播周期内变化并不显著，很难和热点话题的爆发段形成混合，可以随机选取 $2\cdot T_w$ 以上长度的非热点话题信号作为到非热点参考集。最终两类信号曲线效果如图 7 所示，热门话题和非热门话题信号曲线可以较好地区分开来。

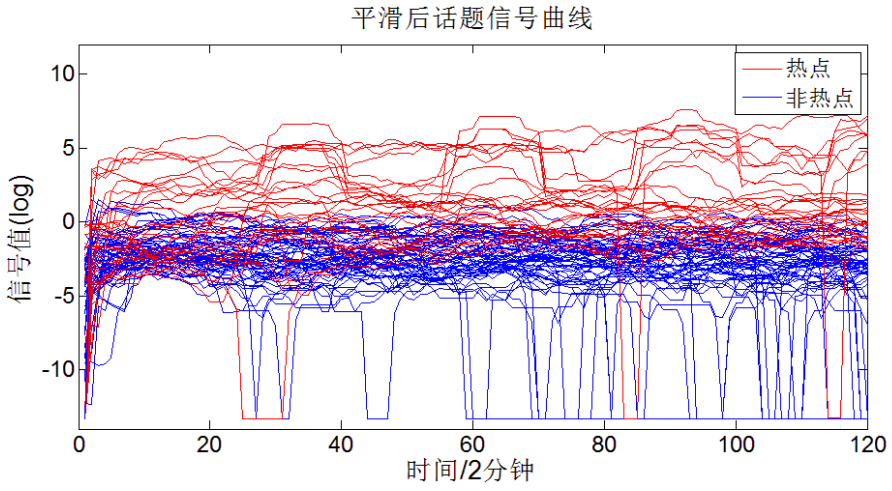

图 7　热点与非热点数据对比图

Fig. 7　The comparison of hot and no-hot data

通过话题的信号变换，可以实现话题趋势特征的构建。在实际中，可以取检测时间点前 T_w 长度的信号作为该时刻话题的趋势表示，默认设置 T_w 为 2 小时。进而考虑 VIP 用户比例、用户粉丝数分布、#话题标签比例和 url 外链比例这四个特征，统计它们在 T_e 时间区间内大小，和信号特征组合在一起，组成话题在检测时刻的特征表示，最后进行热点判别，试验中 T_e 也设置为 2 小时。

3　预测模型

本文采用了基于“潜在基模型”的有监督分类算法[8]。该方法构建了新的模型空间，模型空间由一些未知的“潜在基（基底）”来决定。基本假设是数据是由这些“潜在基底”按照某种方式组合产生的，而且基底的数目要相对小于观测数据的数量。模型中并不刻意设置

这些基底的数量和集合，而是希望通过人们已知的数据来表征这些“潜在基底”。

定义正基底为 $t_1,t_2,...,t_n$，对应类别标签+，负基底为 $q_1,q_2,...,q_\ell$，对应类别标签为-。对于任意+类的观测数据，可以看作是潜在正基底 $t_1,t_2,...,t_n$ 中某个的噪声版本，基底加随机噪声构成了观察数据，反之亦然。

如果观测数据 s 是基底 q 添加噪声得到的，称 s 由 q 生成。定义基底 q 和观测值 s 之间的生成模型为：

$$P(s \leftarrow q) \propto \exp(-\gamma \cdot d(s,q)) \tag{9}$$

其中 $d(s,q)$ 表示 s 和 q 之间的距离，γ 为尺度参数。观测数据与基底的距离越小，它来源于改基底的可能性越高，实际中可以采用欧氏距离：

$$d(s,q) = \sum_{i=1}^{N} (s_i - q_i)^2 \tag{10}$$

当然任何对称性、正定性的凸函数 d 都可以用于这里的距离计算。

3.1 类别检测

为了计算观测数据 s 属于各个类别的概率，首先需要确定各个类别的参考数据集：R_+ 表示+类的参考集合，$R\_$ 表示-类的参考集合。参考集是人工标注的历史数据，对于话题数据，根据话题传播期间的传播范围和影响程度可以归为热点或非热点类别。基于上述模型，如果观测数据 s 和 R_+ 集合中的某个数据具有相同的基底，那么它肯定属于+类。反之，如果参考数据 s 和 $R\_$ 集合中的某个数据具有相同的基底，它一定属于-类。因此观测数据 s 属于+类的概率满足：

$$\begin{aligned} P(+|s) &\propto \sum_{r\in R_+} \sum_{j=1}^{n} P(s \leftarrow t_j, r \leftarrow t_j) \\ &= \sum_{r\in R_+} \sum_{j=1}^{n} \exp(-\gamma \cdot d(s,t_j)) \exp(-\gamma \cdot d(r,t_j)) \\ &= \sum_{r\in R_+} \sum_{j=1}^{n} \exp(-\gamma \cdot (d(s,t_j) + d(r,t_j))) \end{aligned} \tag{11}$$

其中 t_j 表示+类的第 j 个“潜在基底”，n 表示+类基底的总数，r 表示参考集 R_+ 中的数据。当参数 γ 足够大时，关于 j 的求和项的大小主要由最大的指数项确定，此时最大指数项要远大于其他项的值，近似替代可得：

$$P(+|s) \propto \sum_{r\in R_+} \exp\left(-\gamma \min_j (d(s,t_j) + d(r,t_j))\right) \tag{12}$$

然而，表达式中仍包含关于未知的潜在基底 t_j 的最小化项，我们希望消除 t_j 变量将式子修改为关于参考数据 r 和的形式。这里假设“潜在基底”可以合理地覆盖观测数据和参考数据的向量空间，这样关于“潜在基底”集合的最优解 t_{j^*} 应该和整个向量空间的全局最优解 t_* 近似相等。即局部最优解可以近似达到全局最优解的效果，得到观测数据 s 属于+类的概率满足：

$$
\begin{aligned}
P(+|s) &\propto \sum_{r\in R_+}\exp\left(-\gamma \min_j(d(s,t_j)+d(r,t_j))\right)\\
&\approx \sum_{r\in R_+}\exp(-\gamma\cdot C\cdot d(s,r))\\
&= \sum_{r\in R_+}\exp(-\lambda\cdot d(s,r))
\end{aligned}
\tag{13}
$$

同理，可以得到观测数据 s 属于−类的概率满足：

$$
P(-|s)\propto \sum_{r\in R_-}\exp(-\lambda\cdot d(s,r)) \tag{14}
$$

定义观察数据属于+类的概率可以近似表示为：

$$
\begin{aligned}
P(+|s) &= \frac{P(+|s)}{P(+|s)+P(-|s)}\\
&= \frac{\sum_{r\in R_+}\exp(-\lambda\cdot d(s,r))}{\sum_{r\in R_+}\exp(-\lambda\cdot d(s,r))+\varepsilon\cdot\sum_{r\in R_-}\exp(-\lambda\cdot d(s,r))}
\end{aligned}
\tag{15}
$$

基于的假设是观测数据属于+类和−类的概率和右侧指数和呈线性正相关，ε 与 R_+、R_- 参考集有关，默认设置为 1。

应用于热点话题的检测，由于话题数据是一个实时时间流数据，话题传播阶段特征的变化是整体服从一定趋势但在特定时刻可能存在上下的波动，因此在判断 t 时刻话题的类别时，需要考虑历史时刻话题数据对话题类别的影响。定义 $t-1$ 时刻话题对应数据为 s_{-1}，此时刻属于热点的概率为 $P(+|s_{-1})$，那么 t 时刻话题属于热点的概率为：

$$
\begin{aligned}
&R(+|T_t)=\alpha\cdot P(+|s_0)+(1-\alpha)\cdot R(+|T_{t-1})\\
&R(+|T_{t-1})=\alpha\cdot P(+|s_{-1})+(1-\alpha)\cdot R(+|T_{t-2})\\
&\cdots\cdots
\end{aligned}
\tag{16}
$$

其中，α 确定了话题的历史数据对当前时刻的影响程度，且与当前时刻距离越久，话题特征对当前时刻的影响越小。

3.2 在线话题距离的计算

设话题在当前时刻的观测数据为 s，s 可以表示为 T_w 长度的序列（向量），参考集中的某一样本为 r，r 的长度要大于 T_w，假设其长度为 T_{ref}。需要来衡量观测数据 s 和参考数据 r 的距离，距离描述了两个曲线的相似程度。首先计算短信号和长信号的子段之间的距离（取长信号与短信号等长的一部分），然后定义与子序列的最短距离作为两个信号的距离[10]。如果短信号和长信号的某部分完全相同，便终止比较，认为它们之间的距离为 0。比较过程如图 8 所示。

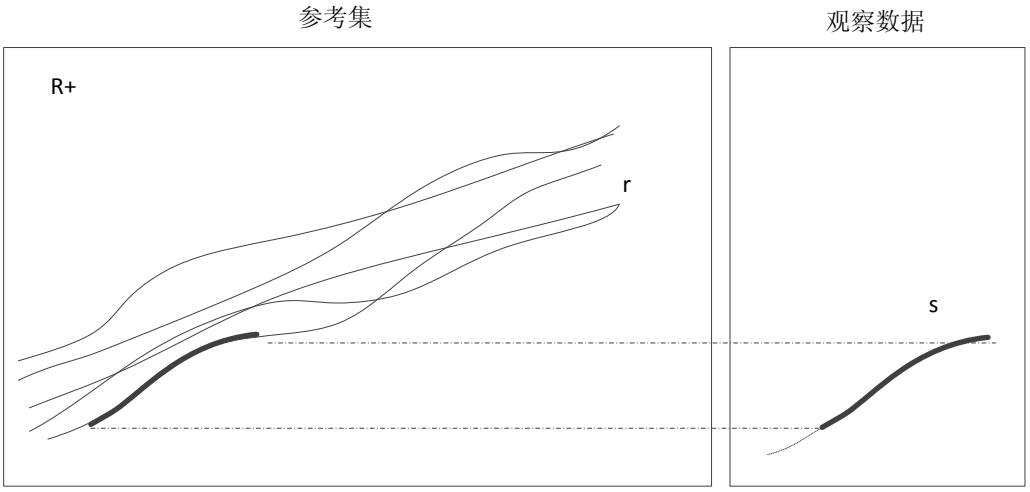

图 8　短信号和长信号的距离测量

Fig. 8　The distance of short and long signals

对于连续的信号，可以从长信号的开头逐渐向后滑动，改变与短信号距离比较的起点，得到两个数据之间的距离：

$$d(r,s)=\min_{k=1,\ldots,T_{ref}-T_w+1} d(r_{k:k+T_w-1},s) \tag{17}$$

其中，$r_{k:k+T_w-1}$ 表示与 s 长度相同的参考信号子序列。

4　结果与分析

实验使用了 2014 年 1 月 1 号到 2014 年 7 月 31 号 7 个月的数据，总共约 1.53 亿条微博。经 LDA 降维、Single-Pass 聚类得到 3864 个话题，对这些话题进行人工标注，最终得到 1269 个热点话题和 2595 个一般性话题（非热点）。

试验数据分为三个部分：模型参考集合（R_+ 与 R_-）、参数寻优集和测试集。随机选取 769 个热门话题作为正类参考集，随机选取 1595 个非热点话题作为负类参考集。使用人工标注的 100 个热点话题和 200 个非热点话题进行模型参数寻优。在最优参数下，使用已标注的 400 个热门话题和 800 个非热门话题对算法进行测试。

4.1　评价标准

假设 S_{hot} 和 S_{normal} 分别表示测试集中热点话题和非热点话题的数量，N_{hot} 为热点话题中检测也为热点话题的总数，N_{normal} 为非热点话题中检测为非热点话题的数目，算法的评估指标主要包括以下四个方面：

- 话题分类的准确率

$$\frac{N_{hot}+N_{normal}}{S_{hot}+S_{normal}}\cdot 100\% \tag{18}$$

- 热点话题预测的准确率

$$\frac{N_{hot}}{S_{hot}}\cdot 100\% \tag{19}$$

- 热点判别正确且提前预测的比率

$$\frac{N_{early}}{S_{hot}} \cdot 100\% \tag{20}$$

其中，N_{early} 为在真实热点话题成为热点之前判为热点话题的话题总数，比率反映了衡量的综合指标，比率越高，算法性能越好。

- 提前预测的平均时间

$$\frac{\sum T_{early}}{N_{early}} \tag{21}$$

其中，T_{early} 表示某热点话题预测为热点的时刻与成为热点的真实时刻之间的差值，平均提前时间反映了算法的有效性。

4.2 试验结果

在分类算法中涉及到一些参数，为了达到最好的实验效果，需要对参数进行寻优，可以对每组参数统计相应的分类性能，然后进行选择。模型主要的参数有两个，第一个是在概率计算中指数系数中距离的缩放因子 λ，试验中选取其值为 0.2、0.5、1 和 2。另一个参数是概率计算中的 ε，ε 可视作为两个相关因子的比值，试验中选取其值为 1、2、5 和 10。

在参数寻优集合上进行实验并统计预测结果，不同参数下分类准确率和热点准确率的变化如图 9 所示。ε 越小，热点准确率越高，但非热点话题越容易产生误判，分类准确率可能下降。综合考虑两项指标，既要较高的热点准确率，又具有较高的分类准确率，可选取 λ=0.5，ε=5 作为模型的最优参数。

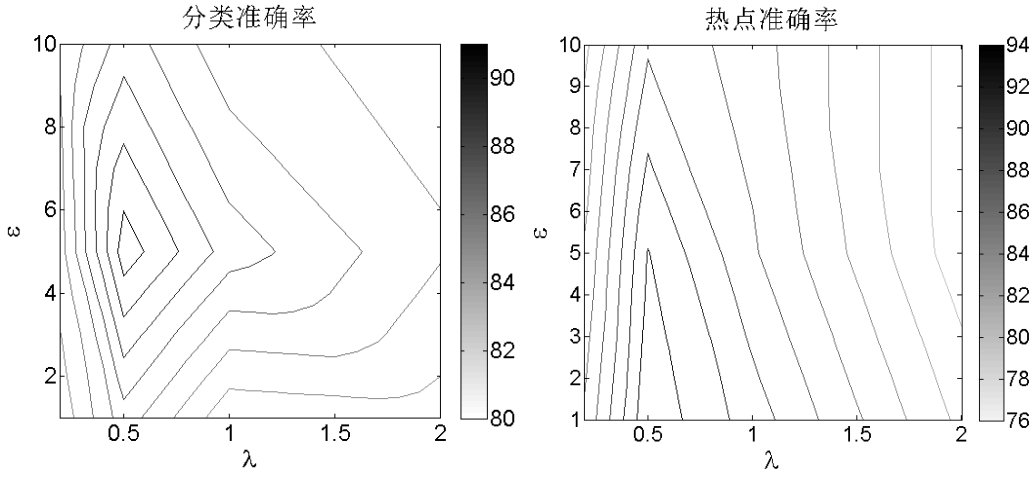

图 9　准确率变化的等高线图

Fig. 9　The contour map of accurate rate

在最优参数下，对测试集进行试验，统计预测结果 N_{hot} 和 N_{normal}，对于判别为热点且实际中人工标注为热点的话题，记录预测其为热点话题的时间，计算该时间点与实际中成为热点的时间间隔 T_{early}。结果表明：400 个热点话题中有 374 个话题可被判别为热点，800 个非热点话题中有 713 个话题被判别为非热点，分类准确率为 90.6%，热点话题准确率为 93.5%。

对热点话题试验结果进一步分析，374 个正确分类的热点话题中 334 个话题的预警时间提前于实际标注为热点的时间，有 83.5%的话题可以提前预测。话题的提前量 T_{early} 分布如图 10 所示，平均提前时间约为 1.6 小时。

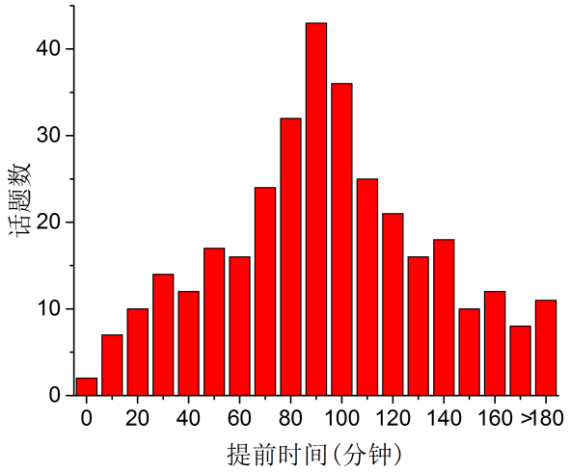

图 10　热点话题提前量分布图

Fig. 10　The distribution of time in advance

5　结论

本文讨论了话题传播的特征，主要详细介绍了话题的趋势特征构建。然后引入了基于“潜在基底”的分类检测方法，由于在实际数据中，观测数据和参考集数据的维度大小不同，文中介绍了一种不同长度序列之间距离的计算方法。最后实验验证了模型的有效性和可靠性。

[参考文献] (References)

[1] 田野. 基于微博平台的时间趋势分析及预测研究[D]. 武汉：武汉大学，2012.
[2] 马社祥. 基于小波分析的非平稳时间序列分析与预测[J]. 系统工程学报，2000，15（4）：305-311.
[3] 韩忠明. 基于内容的热点话题传播模型[J]. 智能系统学报，2013，8（3）：233-239.
[4] Wang Hao, Li Yiping, Feng Zhuonan, Feng Ling. Retweeting analysis and prediction in microblogs: an epidemic inspired approach[J]. China Communications, 2013, 17(3): 13-24.
[5] 兰月新. 突发事件网络衍生舆情监测模型研究[J]. 现代图书情报技术， 2013，19（3）：51-57.
[6] 谢婧.中文微博的话题检测及微博预警[D]. 上海：上海交通大学，2013.
[7] M. Gupta, J. Gao, Ch. Zhai, J. Han. Predicting future popularity trend of events in microblogging platforms[A]. ASIS&T 75th Annual Meeting[C]. Baltimore, Maryland, 2012. 26-30.
[8] Stanislav Nikolov. Trend or no trend: a novel nonparametric method for classifying time series[D]. Massachusetts Institute of Technology, 2012.
[9] Li Kuang , Xiang Tang , Kehua Guo. Predicting the times of re-tweeting in micro-blogs[J]. Mathematical Problems in Engineering, 2014, 11(2): 21-30.
[10] George H.Chen, Stanislav Nikolov, Devavrat Shah. A latent source model for online time series classification[A]. Neural Information Processing Systems[C]. NIPS, 2013. 665-674.