doi:10.11959/j.issn.1000-436x.2016052

基于时间序列分析的微博突发话题检测方法

贺敏[1,2]，徐杰[2]，杜攀[1]，程学旗[1]，王丽宏[2]
（1. 中国科学院计算技术研究所，北京 100080；2. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：针对微博信息噪音大、新颖度难以判断的问题，在动量模型的基础上进行优化，提出了基于时序分析的微博突发话题检测方法。通过动量模型提取候选突发特征后，对特征的动量时间序列分别借鉴信号频域分析理论和股票趋势分析理论进行建模，分析特征的频域特性来识别频繁伪突发特征，分析特征的新颖程度来识别间歇性伪突发特征，合并过滤后的有效突发特征形成突发话题。微博数据实验表明，该方法有效提高了突发话题检测的准确率和 *F* 值。

关键词：突发话题；微博；突发特征；时序分析
中图分类号：TP391 **文献标识码：**A

Bursty topic detection method for microblog based on time series analysis

HE Min[1,2], XU Jie[2], DU Pan[1], CHENG Xue-qi[1], WANG Li-hong[2]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2.National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Detecting bursty topics from microblogs was an important task to understand the current events attracting a large number of internet users. However, the existing hods suitable for news articles cannot be adopted directly for microblogs. Because microblogs have unique characteristics compared wi formal texts, including diversity, dynamic and noise. A detection method for microblog bursty topic was proposed based on time series analysis, which was an optimization method of momentum model. The candidate bursty features were extracted by momentum model. The time series of feature's momentum were modled by frequency domain analysis theory and stock trend analysis theory. The frequently pseudo-bursty features were filtered according to analysis results of frequency-domain characteristics. The intermittently pseudo-bursty features were filtered according to the novelty analysis result through stock trend theory. The bursty topics were finally emerged with combination of effective bursty features. The experiments are conducted on a real Sina microblog data set. It show that the proposed method improves the precis and *F*-measure remarkably compared with the momentum modle.

Key words: bursty topic, microblog, bursty feature, time series analysis

1 引言

近年来，随着 Web 2.0 社交网络的兴起，微博以其方便快捷的优点迅速流行起来，现在已经发展成为网络信息传播的主要途径。微博用户数量众多，每天产生的信息量非常庞大。在微博中，人人都是信息的生产者和传播者，信息发布、转发非常便捷，这使微博成为信息传播速度最快的网络媒体。社会上许多突发性话题，往往在微博平台上首发，借助其好友转发机制迅速传播，引起广泛的社会共鸣，进而波及传统媒体如新闻、论坛、博客等，产生巨大的社会影响。因此，微博平台上的社会突

收稿日期：2015-04-03；**修回日期：**2015-08-29
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（No. 2014AA015203）；国家科技支撑计划基金资助项目（No. 2012BAH46B01）
Foundation Items: The National High Technology Research and Development Program of China (863 Program)(No. 2014AA015203), The National Key Technology Support Program (No.2012BAH46B01)

发话题检测技术，对于社会热点及时发现、网络民意尽快感知、突发事件及早响应等方面都具有积极的现实意义。

这里的微博突发话题是指微博上新出现的可能在短时间内产生强大影响力的关于社会热点事件的网络话题。传统的突发话题检测方法主要面向新闻等长文档数据，而且以突发特征的有效识别为基础，扩展出突发话题。与传统新闻话题相比，微博话题作为大众媒体的产物，具有显著的特点。

1) 话题的多样性。同一时间微博上各种话题，如社会事件类话题、娱乐八卦类话题、个人生活琐事等多种话题掺杂在一起，特别是一些生活琐事类话题，可能表现出一些周期性的突发特点，如周一“不想上班”，周末“出游计划”，月末“月光族”等话题。

2) 话题的间歇性。同一个话题，通常会随着微博用户的关注程度和时间的推移经历一个产生、发展、成熟、衰退和消亡的完整生命周期。而且微博基于好友的转发机制，导致海量的信息冗余，产生大量的滞后过期信息，这使话题表现出一定的间歇性特征。

微博信息表现出的这些特点，对于传统的基于突发特征的突发话题发现方法提出了新的挑战。

1) 对于众多具有突发性特点的特征，如何过滤日常生活类的周期性突发特征，是提高突发话题准确性的一个关键问题。

2) 如何识别间断性突发特征，是确保突发话题新颖性的另外一个关键问题。

本文针对上述挑战，在有意义串动量模型识别突发话题[1]的基础上，进一步优化识别方法，采用信号频域分析的方法，分析频繁特征的频谱特性，区分社会性话题的突发特征与生活类话题的特征，过滤频繁伪突发特征；采用股票趋势分析的方法，利用趋势性信息间接判断突发特征的新颖度，过滤间歇性伪突发特征，提升突发特征识别的准确率，进而提高突发话题检测的准确率。

2　相关工作

话题检测的研究主要包括 3 类方法，第 1 类是基于聚类的方法，有层次聚类、增量聚类等多种方法；第 2 类是基于矩阵分解的方法，有 LSI、NMF 等模型；第 3 类是基于概率生成的方法，有 PLSI、LDA 等模型。但是，突发话题的检测方法主要是以突发特征的发现来驱动，再由突发特征映射到突发话题。Fung[1]首次提出了以特征为中心的话题聚类方法。该方法通过分析时间信息来获取突发特征，然后根据突发特征的分布进行突发话题聚类。He[2]借鉴了 Fung 的方法，通过使用谱分析方法对词语权重(如 TF-IDF)随时间变化的曲线进行分类，然后使用高斯模型和高斯混合模型分别对非周期性特征和周期性特征进行建模，寻找突发时间段，最后使用无监督的贪婪算法检测发现周期性和非周期性突发话题。Kleinberg[3]提出的二状态自动机方法具有开创性，该方法基于一个隐马尔可夫模型（HMM)，模型中的观测数据是主题词在不同时间点上的词频序列，隐变量是词语所处的状态（突发状态或非突发状态），利用参数解析度和状态翻转代价 2 个参数来触发状态转移，发现突发态和突发特征。

近年来，在传统方法的基础上结合了社交网络的新特性，提出了一些针对社交网络突发话题检测的新方法。Cui 等[4]提出了将“#”作为 Twitter 突发事件的指示，根据“#”出现的位置、频次分布、作者等信息定义了稳定性、名言的可能性、作者熵等属性来检测 Twitter 突发事件。Du[5]使用微博中用户影响力、信息的点击数、回复数、收藏数来综合表示关键词的能量，通过计算时间窗口内的平均能量发现突发关键词，对突发关键词进行相似度比较，合并发现突发话题。Shiva[6]提出了通过词典学习的方法来识别新话题，如果当前时刻的文档不能用从前一时刻文档中提取的词典线性表示，则将文档判定为新颖文档，再将所有新文档聚类产生新话题。Zhu[7]把网络论坛话题发现中 2 个有代表性的模型（TF-IDF 和 UF-ITUF）结合起来，从内容特征和用户参与度两方面计算主题和话题的相似度，由此来更新原话题和产生新话题。

上述方法中，Cui[5]和 Du[7]的方法仅考虑了话题的突发程度，Shiva [6]和 Zhu[7]仅考虑了话题的新颖程度。而微博信息纷繁复杂，充斥着大量的历史过期信息和个人生活信息，需要将突发性与新颖性结合起来分析，才能更加准确地识别突发话题。本文在使用动量模型判断特征突发程度的基础上，进一步通过分析特征的时间序列判断特征的频繁程度和新颖程度，准确识别新颖的突发话题特征，有效检测突发话题。

3 基于特征时序分析的微博突发话题检测方法

3.1 基于有意义串动量模型的微博突发话题检测方法

基于有意义串动量模型的突发话题识别方法[8]中实时检测有意义串，发现微博中不断涌现的新词，将新词作为突发话题检测的基本特征；利用动力学原理建模这些基本特征的动态变化特性，通过对特征变化的动量和加速度分析，衡量其变化趋势和突发程度，识别微博的突发性特征，进而发现突发性话题。

有意义串提取[9]是一种回顾性检测，具体的提取过程为：首先通过重复串发现得到候选字符串；然后计算重复串的上下文邻接类别，来衡量候选串是否满足语用多样性；最后通过语言模型来判断字符串的语义完整性，经过两层过滤得到有意义串。有意义串在真实语境中大量使用，比词语粒度更大，可以涵盖正在使用的新词和术语，能够更加准确有效地反映实时微博话题的关键信息。

动态提取观察时间窗口内微博信息的有意义串，作为局部微博信息的特征，借鉴动力学原理对特征进行建模，定义特征的“质量”、“位置”、“速度”、“加速度”、“动量”等基本属性，来反映特征在事件发展过程中的变化趋势及能量大小，进而识别突发特征。特征的若干物理学基本属性的定义如下。

定义 1 特征的“质量”m 指特征的重要性，它不随时间变化，是特征的基本属性，在一段较长时间内基本恒定。该值采用传统的 $tf \cdot idf$ 来衡量，通过统计特征在大量信息中的 tf 和 idf 计算得到。特征 i 的质量 $m(i) = tf(i) \cdot idf(i)$。

定义 2 特征的“位置”x 与时间相关，指特征在某一时刻的流通度或关注度，随时间动态变化。该值与特征在时刻 t 出现的频次、文档频次、参与博主数等相关，计算式为

$$x(t,i) = a \cdot tf(t,i) + b \cdot df(t,i) + c \cdot af(t,i) \qquad (1)$$

其中，$x(t,i)$ 表示特征 i 在时刻 t 的“位置”，$tf(t,i)$ 表示特征 i 在时刻 t 出现的频次，$df(t,i)$ 表示特征 i 在时刻 t 出现的文档频次，$af(t,i)$ 表示在时刻 t 的微博内容包含特征 i 的博主数。a、b、c 是调节参数，通过大量数据的统计分析产生。

上述定义中，特征的“质量”m 是在大量信息中统计得到的，反映了特征在普通文本流中的重要性。特征的“位置”x 是与时间相关的值，反映了特征在时刻 t 的热度。由这 2 个基本的定义，可以计算特征 i 在时刻 t 的速度 v，动量 p 和加速度 a。

根据动力学定义，动量 p 反映了特征在时刻 t 的能量大小及变化趋势，加速度 a 反映了特征在时刻 t 与时刻 $t-1$ 的二阶变化趋势，即时刻 t 的增长率与时刻 $t-1$ 的增长率相比是加快还是放缓。只有当特征的 p 和 a 都满足一定条件时，表明特征在当前时刻的瞬时能量比较大，而且有迅速增长的趋势，该特征才是突发特征。

最后，根据突发特征的共现情况对突发特征聚类，得到突发话题。

3.2 基于频域分析的频繁伪突发特征识别

在 3.1 节方法中，突发特征识别的准确率直接决定了突发话题检测的准确率。动量模型虽然反映了特征的瞬时能量变化趋势，但是不能体现特征在较长时间段的历史能量情况。在真实微博信息中，存在这样一类频繁特征，如“工作人员”、“上半年”，“短信”等，它们周期性或者非周期性的频繁出现，但每天出现的频次不会特别高。由于语言的多样复杂性，这类频繁特征可以在多重语境中重复出现，偶尔呈现瞬时爆发增长趋势，但实际上并非真正的突发话题关键特征，称为频繁伪突发特征。在这种情况下，动量模型将这些特征误识别为突发特征，最终产生错误的突发话题。

为了识别上述频繁伪突发特征，需要对特征在较长历史时间段的频繁程度及变化规律进行分析。而信号的频域分析能够直观看到信号在不同频率成份上的大小分布，直接反映了信号的频繁程度，揭示了信号随出现频率的能量变化规律。因此，借鉴信号频域分析的理论，对特征的动量时间序列建模，将特征在一段时间的动量时间序列看做离散时间信号，变换到频域空间，来观察特征的能量分布规律特性。从离散时间信号变换到离散频域信号的方法从采用信号处理中应用广泛的离散傅里叶变换，变换式如下

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}kn} \qquad (2)$$

其中，$x[n]$ 是特征的动量时间序列，N 是离散时间序列的样点数，$X[k]$ 是变换后的频域信号序列。信号在各出现频率上的能量大小为 $|X[k]|^2$。

例如，通过 3.1 节方法发现在 2013 年 10 月的数据中发现了“事业单位”、“今天下午”、“坠入湄公河”、“暴力恐怖袭击”这几个突发特征，将特征连续 28 天的动量时间序列，通过离散傅里叶变换变换到频域中，在不同频率上的能量分布曲线如图 1 所示。

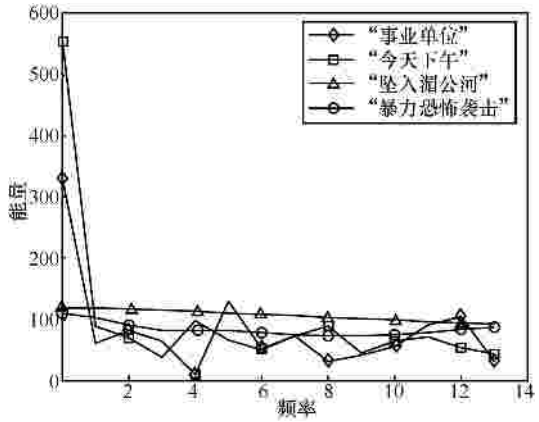

图 1　特征频域能量分布

从图 1 中看出，“事业单位”、“今天下午”2 个特征的能量分布曲线中，0 频率上的能量比较大，而其他频率上的能量相对较小，曲线有较明显的冲击，而实际上它们是频繁伪突发特征；“坠入湄公河”、“暴力恐怖袭击”2 个特征的能量分布曲线中，0 频率上的能量与其他频率上的能量差距相对较小，曲线比较平缓，而它们是真正的突发特征。曲线中 0 频率代表信号的恒定分量，它的大小反映出特征每天出现的稳定情况，曲线中的非 0 频率代表信号的变化分量，它的大小反映出特征的变化情况。如果特征每天出现的频次恒定，那么频谱曲线中将只有 0 频率的能量，其他频率能量为 0。为了区分频繁伪突发特征和突发特征，利用上述特性给出能量比 S 的定义。

定义 3　信号的频谱分布中，0 频率的能量与其他频率能量平均值的比值称为能量比，用符号 S 表示

$$S=\frac{|x[0]|^2}{\frac{1}{K}\left(\sum_{k=1}^{K}|x[k]|^2\right)} \tag{3}$$

S 可作为特征是否为频繁特征的度量，S 的值越大，特征是频繁特征的可能性越大。实际应用中通过大量的标注数据训练得到阈值 S_T，采用与阈值比较的方法过滤掉频繁伪突发特征。

3.3　基于趋势分析的间歇性伪突发特征识别

话题呈现出一定的生存周期，有些话题的产生、发展、消亡是连续的，从话题产生开始到结束期间，每天都产生相关消息，但是有些话题的发展是间歇性的，话题产生后沉寂几天才会有新的进展和消息，在话题存在的整个周期中，不一定每天都出现相关讨论。因此，在 3.1 节方法中，存在一部分突发特征误判，这类特征间歇性的出现，实际上是一个话题，但由于动量模型只判断了瞬时的动量和加速度，尚未判断特征更长时间段是否活跃，产生误判。例如，图 2 中所示的赣南脐橙被染色的话题中的特征“赣南脐橙”，最早在 10 月 25 日出现，在接下来的 3 天内讨论比较少，而在 10 月 29 日又再次爆发，根据动量模型方法，它被再次识别为新的突发话题，而该话题实际上仍是 10 月 25 日话题的延续，并非新颖的话题。

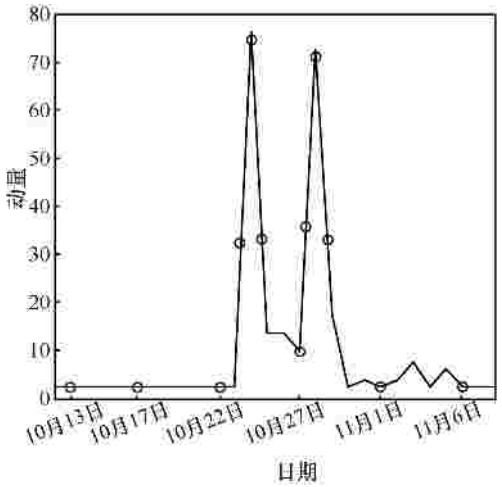

图 2　“赣南脐橙”动量分布

为了识别间歇性伪突发特征，需要分析特征在一定周期内的新颖性。特征在一个话题周期内第一次大量出现称为突发，而当特征在一个话题周期内再次出现时，即使表现出瞬时爆发，但是由于此次爆发与第一次真正的突发存在间歇期，从阶段性趋势来看，经过间歇期后的爆发点应该已经处于特征的下降趋势范围。股票趋势分析方法就是针对一定时间段的价格变化进行平滑，分析价格在一定时间范围的上涨或者下跌趋势。因此，借鉴股票趋势分析的方法，对话题周期内的特征动量进行平滑，分析其阶段性的趋势特点，进而判断突发特征的新颖性，识别出间歇性伪突发特征。下面给出几个定义。

定义 4　指数移动平均(EMA, exponential

moving average)将特征的动量时间序列进行 n 天指数平均，平均后的动量值与前 n 天的动量值相关，对于较近的动量值权重较大。

$$EMA(n)[x]_t = a\,x_t + (1-a)\,EMA(n-1)[x]_{-1} = \sum_{k=0}^{n} a\,(1-a)^k\,x_{t-k} \tag{4}$$

其中，x_t 是第 t 天的动量值，$EMA(n)[x]_t$ 是 n 天指数平均动量值，它对于近期动量的变化要快于简单的 n 天平均值。a 取值与 n 相关，一般为 $\frac{2}{n+1}$。

定义 5 移动平均收敛发散（MACD, moving average convergencc-divergence）指标是由 2 条曲线构成：一条实线（称为 MACD 线）与一条虚线（称为 signal 线）。MACD 线是较快的 EMA 线和较慢的 EMA 线的差值，它对于动量值变动的反应比较敏感。较快的 EMA 线与较慢的 EMA 线相比，n 取值更小，受影响的历史区间更小，对当前值反应更快。signal 线是 MACD 线是经过指数平均之后的另一条 EMA 线，它对于动量值变动的反应比较缓慢。计算式如下

$$MACD(n_1,n_2) = EMA(n_1) - EMA(n_2), n_1 < n_2 \tag{5}$$

$$signal(n_1,n_2,n_3) = EMA(n_3)\left[MACD(n_1,n_2)\right], \quad n_1 < n_3 < n_2 \tag{6}$$

当快速的 MACD 线穿越慢速的 Signal 线，动量的趋势发生变化。用 histogram 来表示，它是 MACD 和 signal 的差值，计算式如下

$$histogram(n_1,n_2,n_3) = MACD(n_1,n_2) - signal(n_1,n_2,n_3) \tag{7}$$

histogram 扩大了特征的平均动量值和局部波动之间的差异，当 *histogram*＞0 时，表示动量处于上升趋势，当 *histogram*＜0 时，表示动量处于下降趋势。它可用来反映特征动量的变化趋势，作为衡量特征新颖性的指标。在经过动量模型判断特征是否突发后，再次根据特征的 *histogram* 值是否大于 0，来判断特征是否是间歇性伪突发特征。间歇性突发特征在首次出现时，*histogram*＞0，动量呈现上升态势，属于突发特征；当间歇性特征在话题周期内再次出现时，*histogram*＜0，动量呈现下跌态势，则表明特征在前期出现过大规模爆发，判断该特征不是新颖的突发特征，该时间点不是突发点。例如，“赣南脐橙”的趋势分析如图 3 所示，m 表示特征的动量，在 10 月 29 日时，虽然动量和加速度都比较大，但是 *histogram* 值小于 0，表明该特征不是新颖的特征，属于间歇性的伪突发特征。

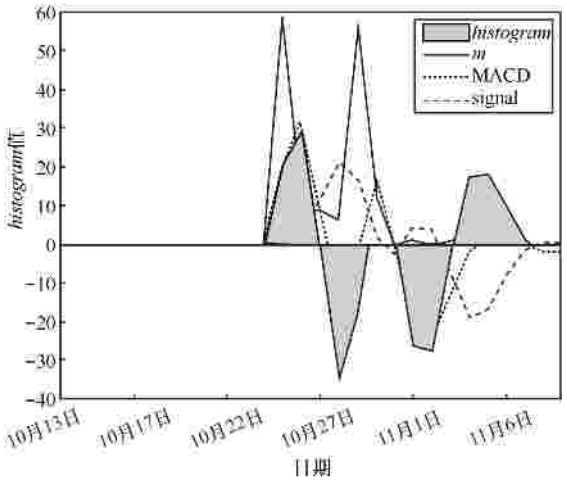

图 3 “赣南脐橙”趋势分析

在突发特征的趋势分析中，参数 n_1、n_2、n_3 的取值与话题周期相关。变化较慢的 EMA 线中 n_2 的取值为一个话题周期，因为间歇性伪突发特征的识别限定在一个话题周期内，对于时间超过一个话题周期的突发特征可能是真正的突发特征。参考股票趋势分析方法的参数选取，n_1 的取值为 n_2 的一半，n_3 取值介于 n_1 和 n_2 之间，能够反映出趋势的变化情况。通过大量统计分析和训练发现，大部分话题的生存周期在一周以内，n_1 取值为 3，n_2 取值为 7，n_3 取值为 5，对于间歇性突发特征的识别比较有效。对于具体领域的突发话题检测，n_1、n_2、n_3 的取值需要根据应用场景重新调整。

4 实验及结果分析

4.1 实验数据及评价标准

本文通过互联网采集新浪微博由 1 000 个加 V 活跃博主从 2013 年 8 月~11 月发表的 78 万余条微博信息作为实验数据。实验将 8 月~10 月中旬两个半月的数据作为训练语料，将 10 月 13 日~11 月 9 日的数据作为微博信息流检测每天的突发话题。由 2 名舆情分析领域的专业人员对每天的数据进行标注，分别产生 185 和 193 个突发话题，取 2 人标注的交集共 180 个突发话题做为作为评价实验结果的标准。根据标注结果计算突发话题的准确率 P、召回率 R 和综合指标 F 值，以此评价算法。

4.2　实验结果

由于微博文本高度稀疏，采用聚类的突发话题检测方法在微博上效果较差，本实验采用将内容与用户结合起来的经典模型 TF-IDF&UF-IUF[7]方法与动量模型方法、两类改进型的动量模型方法来作对比。其中，2 类改进型的动量模型方法是在动量模型基础上通过对特征频域分析和趋势分析进行的优化。实验结果如表 1 所示。

表 1　　突发话题检测结果

方法	*P*	*R*	*F*
TF-IDF&UF-ITUF	70.27%	72.22%	71.23%
动量模型	87.71%	87.22%	87.47%
动量模型+频域分析	91.18%	86.11%	88.57%
动量模型+趋势分析	92.35%	87.22%	89.71%
动量模型+频域分析+趋势分析	96.27%	86.11%	90.91%

从表 1 中看出，不管是动量模型方法，还是在其基础上优化的频域分析和趋势分析方法，实验结果的准确率和召回率都比 TF-IDF&UF-IUF 方法高很多。这是因为动量模型较好地反映了特征的瞬时能量变化，能够快速发现突发特征。频域分析方法和趋势分析方法进一步对特征进行较长时间段能量变化分析，与动量模型方法相比，实验结果对召回率的影响较小，准确率都有较大的提升。趋势分析方法与频域分析方法相比，准确率和召回率都略高，因为趋势分析方法过滤掉的突发话题均为错误结果，对召回率没有影响，准确率提高较多；而频域方法在过滤掉大部分错误结果的同时，也将个别正确的突发话题过滤掉，在准确率提高的同时，对召回率有影响。实验验证了频域分析方法能够准确地识别频繁伪突发特征，趋势分析方法能够准确地识别间歇性伪突发特征，有效地过滤了动量模型中的伪突发特征，进而减少了错误的突发话题发现结果。经过频域分析和趋势分析方法两层过滤后，剔除了大部分的错误结果，突发话题发现的准确率已经高达 96.27%。

为了进一步分析话题准确率提高的原因，表 2 显示 3 种方法对于突发特征的识别结果对比。从表中看出，频域分析在过滤掉一些频繁伪突发特征的同时，也将一部分正确的突发特征当作频繁特征错误过滤，但是过滤的频繁伪突发特征比例仍高于误判的正确突发特征比例，所以在突发特征准确率上仍有较大提升。虽然过滤掉一部分正确的突发特征，但是从表 1 看出，该步骤对正确突发话题的检测数量只产生微小影响，这是因为检测发现的多个突发特征最终对应一个正确的突发话题，只要尚未将一个突发话题对应的所有突发特征都过滤掉，仍然可以通过特征聚类产生该突发话题。例如，10 月 14 日的数据中，“高考改革、分值、英语科目、北京高考”是一个突发话题，“分值”通过频域分析方法作为一个频繁特征被删除，但是该话题的其他几个突发特征仍然存在，合并后突发话题仍然可以准确识别。而“短信、客服”这样的突发话题本来就是错误结果，对应的突发特征数量一般比较少，通过频域分析能够将其全部过滤。

从表 2 的结果可以看出：趋势分析的方法能够减少错误的突发特征数量，而准确的突发特征数量几乎没有下降，突发特征发现的准确率得到提升。因为趋势分析方法仅将已经出现过的间歇性突发特征过滤，对于正确的突发特征发现影响很小。例如，对于图 2 所示的话题，通过趋势分析能够在 10 月 28 日判断出“赣南脐橙”和“催熟染色”这 2 个突发特征并非新颖特征，将其从突发特征中删除。

表 2　　突发特征识别结果

方法	识别数量	准确数量	准确率
动量模型	1 262	1 089	86.29%
动量模型+频域分析	1 024	933	91.11%
动量模型+趋势分析	1 185	1 087	91.73%
动量模型+频域分析+趋势分析	945	933	98.73%

5　结束语

本文针对动量模型方法对突发特征误判的现象，提出了采用时间序列分析方法来过滤伪突发特征来检测突发话题的优化方法。在动量模型的基础上，对特征的动量时间序列分别借鉴信号频域分析理论和股票趋势分析理论进行建模，通过特频域分析过滤掉频繁伪突发特征，通过趋势分析过滤掉间歇性伪突发特征，最终对有效的突发特征聚类产生突发话题。实验中，在对突发话题检测召回率影响较小的情况下，该方法相对动量模型方法将突发话题检测准确率从 87.71%提升到 96.27%，能够有效提升突发话题检测的准确率和 *F* 值。

基于特征时序分析的微博突发话题检测方法有效提升了突发话题的准确率，但未来仍需在如下 2 个方向上继续探索：1）优化突发特征的识别策略，

采用学习方法或产生式策略加以整合识别；2）提升突发特征识别的召回率，通过利用好友关系、链接关系、转发关系等丰富的关联关系，弥补动量模型方法对突发特征识别的漏检。

参考文献：

[1] FUNG G, YU J, YU P, et al. Parameter free bursty events detection in text streams[C]//Conference on 31th VLDB. Trondheim, Norway, c2005: 181-192.

[2] HE Q. CHANG K, LIM E. Analyzing feature trajectories for event detection[C]//Conference on 30th SIGIR. Amsterdam, c2007: 208-214.

[3] KLEINBERG J. Bursty and hierarchical structure in steam[C]// Conference on KDD'02. Edmonton, Alberta, Canada, c2002: 91-101.

[4] CUI A, ZHANG M, LIU Y, et al. Discover breaking events with popular hashtags in twitter[C]// Conference on CIKM'12. Maui, HI, USA, c2012: 1796-1798.

[5] DU Y Y, HE Y X, TIAN Y. Microblog bursty topic detection based on userrelationship[C]// 6th IEEE Information Technology and Artificial Intelligence Conference. Chongqing, China, c2011: 260-263.

[6] SHIVA P K, PREM M, ARINDAM B. Emerging topic detection using dictionary learning[C]//Conference on CIKM'11. Glasgow, Scotland, UK, c2011: 745-754.

[7] ZHU M L, HU W M, WU O. Topic detection and tracking for threaded discussion communities[C]// IEEE/WIC/ACM International Conferences on Web Intelligences and Intelligent Agent Technology. c2008: 77-83.

[8] **贺敏, 杜攀, 张瑾, 等. 基于有意义串动量模型的微博突发话题检测方法**[J]. **计算机研究与发展**, 2015, 52(5): 1022-1028.
HE M, DU P,ZHANG J, et al. Microblog bursty topic detection method based on momentum model [J]. Journal of Computer h and Development, 2015, 52(5):1022-1028

[9] **贺敏．面向互联网的有意义串挖掘**[D]. **北京：中国科学院计算技术研究所，**2007.
HE M. Web-oriented Chinese meaningful string mining[D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2007.

[10] ALAN R, MAUSAM, O E. Open domain event extraction from twitter[C]// Conference on KDD'12. Beijing, China, c2012: 1104-1112.

[11] ANDREW J, YASHAR M, JOEMON M. Building a large-scale corpus for evaluating event detection on twitter[C]// Conference on CIKM'13. San Francisco, CA, USA, c2013: 409-418.

[12] DIAO Q M, JIANG J, ZHU F D, et al. Finding bursty topics from microblogs[C]// The 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Korea, c2012: 536-544.

[13] POPESCU A M, PENNACCHIOTTI M, PARANJPE D A. Extracting events and event descriptions from twitter[C]// Conference on WWW'11. Hyderabad, India, c2011: 105-106.

[14] LI C L, SUN A X, DATTA A. Twevent: segment-based event detection from tweets[C]// Conference on CIKM'12. Maui, HI, USA, c2012: 155-164.

作者简介：

贺敏（1982-），女，山西忻州人，中国科学院计算技术研究所博士生，主要研究方向为网络信息安全、舆情分析、自然语言处理等。

徐杰（1982-），男，山西五寨人，博士，国家计算机网络应急技术处理协调中心工程师，主要研究方向为网络信息安全和多媒体技术。

杜攀（1981-），男，河南南阳人，中国科学院计算技术研究所助理研究员，主要研究方向为文本挖掘、信息检索、机器学习等。

程学旗（1971-），男，安徽安庆人，中国科学院计算技术研究所研究员、博士生导师，主要研究方向为信息检索、文本挖掘、社会计算等。

王丽宏（1967-），女，辽宁沈阳人，国家计算机网络应急技术处理协调中心副总工程师、研究员，主要研究方向为网络信息安全、舆情分析等。