DOI:10.7544/issn1000-1239.2015.20131549
52(5):1022-1028,2015

基于动量模型的微博突发话题检测方法

贺 敏[1,2] 杜 攀[1] 张 瑾[1] 刘 悦[1] 程学旗[1]
[1](中国科学院计算技术研究所 北京 100190)
[2](国家计算机网络应急技术处理协调中心 北京 100029)
(heminsmile@163.com)

Microblog Bursty Topic Detection Method Based on Momentum Model

He Min[1,2], Du Pan[1], Zhang Jin[1], Liu Yue[1], and Cheng Xueqi[1]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

Abstract Microblogs reflect the general public's real-time reaction to major events. Finding bursty topics from microblogs is an important task to understand the current events which attract a large number of Internet users. However, the existing methods suitable for news articles aren't adopted directly for microblogs, because microblogs have unique characteristics compared with formal texts, including diversity, dynamic and noise. In this paper, a new detection method for microblog bursty topic is proposed based on momentum model. The meaningful strings are extracted from micorblog posts in the special time window as the microblog dynamic features. The dynamic characteristics of these features are modeled by the principle of momentum. The velocity, accelerated velocity and momentum of the features are defined by the dynamic frequencies at different dimensions. The bursty features are detected with the combination of momentum, variation trend and second order change rate. By merging the detected bursty features with mutual information, the bursty topics are obtained. The experiments are conducted on a real Sina microblog data set containing around 526 thousand posts of 1000 users, and results show that the proposed method improves the precision and recall remarkably compared with the conventional methods. The proposed method could be well applied in online bursty topic detection for microblog information.

Key words bursty topic; microblog; bursty feature; meaningful string; momentum model

摘 要 针对微博特征空间动态变化、信息噪音大的特点,提出一种基于有意义串动量模型的微博突发话题检测方法.提取时间窗口内微博信息流的有意义串,作为微博信息的动态特征,根据动力学原理对特征进行动量建模,结合特征能量大小、变化趋势以及二阶变化率检测突发特性有意义串,即突发特征,合并突发特征形成突发话题.微博数据实验表明,该方法适用于在线微博突发话题检测,在准确率和召回率上都有明显提升.

关键词 突发话题;微博;突发特征;有意义串;动量模型

中图法分类号 TP391

微博是近年来兴起的 Web2.0 新媒体.用户可以通过手机、即时通信工具、Email、Web 等媒介在个人微博上发布 140 字以内的文本信息及图片、影音等多媒体内容,展现个人最新动态,分享身边实时

收稿日期:2013-11-29;**修回日期**:2014-08-12
基金项目:国家"八六三"高技术研究发展计划基金项目(SS2014AA012303);国家自然科学基金项目(61303156)

信息.微博平台上每天产生数量庞大的信息,据统计,新浪微博 2012 年 11 月日均发微博量约 1.366 亿条,平均每分钟约 94 907 条.而且,由于微博与多种媒体关联,信息发表、转发非常便捷,微博成为信息传播速度最快的媒体.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发话题检测技术,对于最新社会热点发现、网络民意及时感知、舆情检测、应急处置等方面都具有积极的现实意义.

但不同于传统的新闻文档,微博数据具有内容短小、数量巨大、信息零碎、用语不规范等显著特性,这些新特点为面向微博的突发话题检测技术带来了新的挑战:

1) 微博信息用词不规范,须及时识别微博新词.每个用户随时都可以发表微博,信息具有原创性和时效性的同时,也表现出草根性和随意性,用词口语化、不规范现象严重,简称、缩略语大量存在.随着网络事件的事态发展,微博空间不断涌现出表达话题核心语义的新词,只有及时动态地发现这些重要新词,才能准确地表达话题内容.因此,新词的不断涌现,对突发话题发现技术提出了新的挑战.

2) 微博信息数量庞大,突发话题容易被信息噪音淹没.微博用户根据个人兴趣每天发表大量身边发生的事件,信息琐碎零散.基于好友转发的传播机制,导致海量的信息冗余.因此,对于突发话题,虽然其在话题相关的微博数量上增长迅猛,但总量有限,很容易被各种噪声信息、热点话题等所淹没,难以识别.

本文针对上述挑战,采用实时检测有意义串的方法来发现微博中不断涌现的新词,作为突发话题检测的基本特征.利用动量模型建模这些基本特征的动态变化特性,通过对微博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.

1　相关工作

突发话题检测属于话题跟踪与检测(topic detection and tracking, TDT)的一个分支,与其他的 TDT 方法不同,突发话题检测是从话题随时间的动态特性出发,通过提取动态特征来进行话题检测的一系列方法.与传统的 TDT 以文档为中心的聚类方法不同,Fung 等人[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,聚类的结果是若干突发特征的集合.He 等人[2]借鉴了 Fung 等人的方法,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解和分类;并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来找到突发时间段;最后使用无监督的贪婪算法对周期性和非周期性突发话题进行检测.Kleinberg[3]提出的二状态自动机方法是具有开创性的突发特征检测方法,该模型是一个隐 Markov 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).在非突发情况下,词频的分布是在整个文本流中的均匀分布.在突发状态下的分布和自动机的状态转移概率分别用 2 个参数解析度和状态翻转代价来控制.He 等人[4]提出采用物理动力学原理对突发话题建模,通过植物学领域学术论文的阅读数量、引用数量、发表期刊影响力等来表示话题的重要性,引入话题的速度、动量、加速度等来描述话题的能量,采用股票市场的趋势统计方法来发现能量突发的话题.上述方法主要用于新闻、论文等规范的长文本,对于非正式用语频繁的微博数据将不太适用.

近年来,也出现了一些针对 Web2.0 媒体的突发话题检测研究.Zhu 等人[5]把代表性的 2 个模型 TF-IDF 和 UF-ITUF 结合起来,从内容特征和用户参与度 2 方面来计算主题和话题的相似度,并且由此来更新原话题和产生新话题.Chen 等人[6]用词的突发性作为噪音过滤的重要指标,基于突发词和核心用户设计无监督的突发话题发现算法来发现突发话题.Du 等人[7]将用户影响力、信息的点击数、回复数、收藏数综合表示关键词的能量,计算时间窗口内的平均能量来发现突发关键词.Diao 等人[8]结合微博的时间信息和用户个人兴趣,提出一种基于泊松状态机的突发话题检测方法.Petrovic 等人[9]提出了根据“Hash”的位置来检测 Twitter 中突发事件的方法.张晨逸等人[10]提出了改进的 LDA 模型,综合考虑了微博联系人关联信息和文本关联信息,来辅助进行微博的主题挖掘.Hong 等人[11]利用 Twitter 中的转发特征预测微博中的流行信息,检测突发新闻.Li 等人[12]先通过频繁模式方法发现微博中的突发语言片段,然后根据频次分布和内容相似度聚类得到事件,最后通过维基百科过滤产生有价值的事件描述片段.Phuvipadawat 等人[13]针对微

博中短文本带来的相似性计算问题,提出了一种方法来收集、分组、排序以及跟踪 Twitter 中的突发新闻.

上述方法虽然是针对论坛、微博的话题发现方法,但是大部分采用传统的词语空间作为特征空间,未考虑信息中的新词.而微博数据中新词大量涌现,这些新词经常是突发事件的关键特征,随着事件的发展而产生、消亡,微博的特征空间也将跟随事件动态变化.本文针对微博的这一特点,对 He 等人[4]的方法提出了改进,在微博数据上对有意义串重新建模,提出了实时检测微博突发特征的方法,最终发现微博突发话题.

2 微博突发话题检测方法

2.1 微博有意义串检测

本文作者前期提出了有意义串的概念[14],指具有统计意义、包含具体语义、能够独立灵活使用的语言单元.有意义串既包含未登录的新词和命名实体,又包含有意义的词组和短语.如"博鳌亚洲论坛"、"波士顿爆炸"、"周丹龄"和"中国式过马路"等.

有意义串提取[14]是一种回顾性检测,检测方法与数据规模大小无关,主要提取表现出重复特性的有意义串,对于在检测数据中只出现一次的有意义串,则无法提取.有意义串提取的一个重要依据是字符串的邻接类别,表示字符串的上文或者下文相邻不同词语的类别数量,反映了字符串语言环境的灵活性.具体的提取过程为:1)通过重复串发现得到具有统计意义的候选字符串;2)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;3)通过语言模型来判断字符串的语义完整性,经过 2 层过滤得到有意义串.

微博平台上的瞬时信息数量非常大,设定时间窗口 T,将时间窗口 T 内的微博信息作为检测数据,能够提取在时间窗口 T 内表现出重复特性的有意义串.突发话题一般在短时间内大量爆出,其关键信息在短时间内也将大量重复出现,所以,在时间窗口 T 内爆发的突发话题关键信息应该包含在时间窗口 T 的微博有意义串中.

从有意义串的提取过程看出,有意义串在当前时间的真实文本中具有一定流通度,能够在多种不同语言环境中使用.而且,有意义串的粒度可以比词语更大,能够更加具体完整地反映话题的关键信息.随着网络事件的事态发展,微博空间不断涌现出大量的新词和术语,用传统的静态词典中的词语来表示微博信息将会遗漏部分关键特征,而从微博流通文本中提取出来的有意义串,可以涵盖正在使用的微博新词和术语,能够更加准确有效地反映微博的实时内容.所以,与传统的词空间相比,有意义串空间更适合表示微博信息.本文将有意义串作为微博信息的基本特征,采用上下文邻接分析与语言模型相结合的方法[14]来动态检测微博信息中的有意义串,构成微博信息的动态特征空间.

2.2 微博有意义串的时变特性

微博信息涌现的大量有意义串具有生命周期,呈现出时间局部性和空间局部性的特点.有意义串在事件的生存周期内大量出现,事件结束后,有意义串在文本中的流通度大幅下降,甚至消失.如图 1 所示,有意义串"博鳌亚洲论坛"和"波士顿爆炸"在微博信息中的出现频次呈现出明显的局部性,"博鳌亚洲论坛"在 2013 年 4 月 6 日博鳌论坛开幕后的一段时间内,"波士顿爆炸"在 2013 年 4 月 16 日波士顿发生爆炸案发生的几天内,分别频繁出现,而其他时间则很少出现;"中国式过马路"这个有意义串 2012 年已经被提出并热烈讨论,在 2013 年 4 月的信息中频次较低.

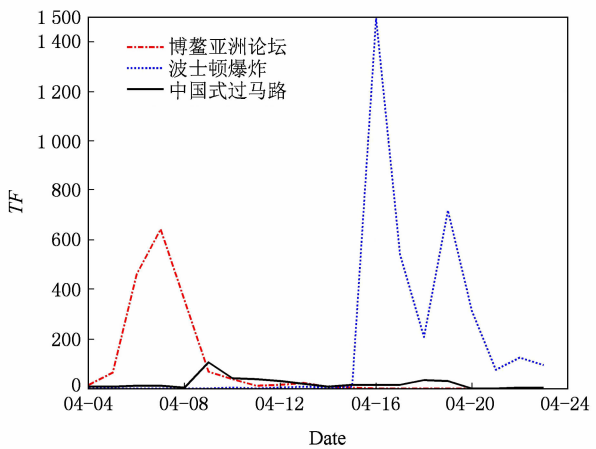


Fig. 1 Time distribution of meaningful strings.

图 1 有意义串时间分布图

所以,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,一方面可以显著降低该窗口内信息的特征维度,与静态的词空间相比,能够大幅缓解特征稀疏问题,降低计算复杂度;另一方面能够即时发现最新出现的关键特征,为实时突发特征检测创造了条件.

微博信息是时间序列上的文本流,设置观察时间窗口 T,将时间窗口 T 内的微博信息作为文档集

合 $D=\{D_1,D_2,D_3,\cdots\}$,提取 D 中的有意义串,形成窗口 T 内微博信息的特征空间 S. 随着时间窗口的推移,特征空间 S 将动态变化.

2.3　突发特征识别

突发特征识别是突发话题检测区别于其他话题检测的关键,一般通过提取在时间上具有突发性的词语作为突发特征. 本文将有意义串作为特征,借鉴 He 等人[4]的研究方法,采用物理动力学原理来构建微博的突发特征模型. He 的方法将某个学术领域的论文作为研究对象,发现论文中的突发话题. 通过话题相关论文的数量、阅读数量、引用数量、发表期刊的影响力等来衡量一个话题的重要性. 因为微博数据与学术论文相比,在内容特征和结构特征方面存在显著不同,内容不正式、随意性大、数量更加庞大、结构更加复杂,He 的方法不能直接应用于微博领域. 本文将结合微博的数据特点,采用动量模型来发现具有突发特性的微博有意义串.

微博话题产生、发展、高潮、衰落、消失的过程,与动力学中的物体从静止开始运动、速度加快再到速度变缓、最终停止的过程类似. 在物理学中,动量是与物体的质量和速度相关的物理量,描述这个物体在它运动方向上保持运动的趋势. 有意义串动量模型就是指将有意义串作为微博的特征,借鉴动力学中的动量定义对特征建模,将特征在大规模统计文本中的流通度作为特征的"质量"m、将特征当前时刻的热度作为特征的"位置"x 来计算特征在当前时刻的动量,动量直接反映了特征在事件发展中的能量大小和变化趋势. 如下给出特征的物理学基本属性的具体定义:

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. 该值采用传统的方法 TF-IDF 来衡量,通过统计特征在大量信息中的 tf 和 idf 值计算得到. 特征 i 的质量 $m(i)=tf(i)\times idf(i)$.

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,计算如下:

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

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.

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 的一系列物理值:

速度:

$$v=\frac{\Delta x}{\Delta t},\quad(2)$$

加速度:

$$a=\frac{\Delta v}{\Delta t},\quad(3)$$

动量:

$$p=m\times v.\quad(4)$$

经过动量模型建模后,特征的动量 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.

微博的突发特征是与时间相关的,指在某一时刻突然爆发、大量涌现的特征. 如 2013 年 4 月 6 日的"博鳌亚洲论坛",在 2013 年 4 月 6 日以前也零星出现,但数量比较少,而 2013 年 4 月 6 日的出现数量呈集中爆发态势,属于 2013 年 4 月 6 日的一个突发特征. 2013 年 4 月 16 日的"波士顿爆炸",之前的一段时间内从未出现,2013 年 4 月 16 日第 1 次出现,而且数据巨大,也是典型的突发特征. 从这 2 个例子可以看出,突发特征具有 2 方面的特性:1)当前时刻的瞬时能量比较大;2)与历史情况比较,加速度比较大,有迅速增长的趋势. 这 2 方面正好与动量 p 以及加速度 a 相对应. 所以,基于特征的动量模型能够检测出突发特征,首先计算特征在时刻 t 的动量 p 和加速度 a;然后判断 p 和 a 是否大于一定阈值,当 p 和 a 都满足一定条件时该特征就是突发特征. 动量 p 和加速度 a 的阈值通过在一定规模的标注语料中训练学习而得到.

2.4　突发话题发现

2.3 节检测出的每个突发特征对应一个广义的话题,这些广义话题之间可能存在交叉重复现象. 所以,还需要对这些突发特征进行合并,多个突发特征共同来描述一个话题,形成具体明确的突发话题.

特征之间的互信息指特征在相同微博信息中的共现情况,体现了 2 个特征的依赖程度,互信息越高,特征的相关度越高,描述同一话题的可能性越大. 本文通过计算特征之间的互信息(式(5))来对特征进行合并. 考虑到话题的特征之间可能有交叉,一

个突发特征有可能描述多个不同的话题,特征合并时需要计算两两之间的互信息,互信息大于一定阈值时将特征合并.经过多轮层次合并后,最后得到突发话题.

$$MI(i,j)=\mathrm{lb}\frac{P(i,j)}{P(i)P(j)},\tag{5}$$

其中,$P(i)$代表特征 i 在观察时间窗口的文档中出现的概率;$P(i,j)$代表特征 i 和 j 在时间窗口内共现的概率.

综上所述,突发话题检测的算法描述如下:

算法 1. 突发话题检测算法.

输入:微博历史数据集、微博实时数据流;

输出:突发话题流.

Step1. 提取历史微博数据集中的有意义串,计算每个串的质量 m;

Step2. 接收时间窗口 T 内的微博数据流,记为数据集 D;

Step2.1. 提取 D 中的有意义串,记为集合 FS;

Step2.2. 对于 FS 中的每个有意义串 i

Step2.2.1. 计算动量 p 和加速度 a;

Step2.2.2. 如果($p>T_p$)且($a>T_a$),有意义串 i 是突发特征;

Step2.3. 突发特征合并,产生突发话题;

Step3. 转到 Step2.

3 实验及结果分析

3.1 实验数据及评价标准

中文微博的研究还处于起步阶段,目前尚无公认的语料集和标注结果.本文通过互联网采集新浪微博 1 000 个加 V 活跃博主发表的,2013-02-23—2013-04-23 两个月的微博信息共 526 438 余条.由 2 名舆情分析领域的专业人员分别对每天的数据进行标注,产生每天的突发话题,取 2 人标注的交集作为评价实验结果的标准.

实验计算突发话题的准确率(precision, P)、召回率(recall, R)和综合指标(F-measure, F)来对算法评价.

3.2 实验结果

实验将全部的 52 万余条数据作为背景语料计算特征的"质量"m,将 2013-02-23—2013-04-23 的标注结果作为训练数据集,学习得到速度 v、加速度 a 以及动量 p 的阈值.以天作为时间窗口,将 2013-04-04—2013-04-23 的数据看作微博信息流检测每天的突发话题.

3.2.1 本文方法与 TF-IDF&UF-IUF 方法比较

TF-IDF&UF-IUF 方法[5]是网络论坛话题发现比较有代表性的模型,本实验选择该方法与本文方法进行对比,实验结果如表 1 所示.其中,速度模型采用速度 v 和加速度 a 来检测突发特征,动量模型采用动量 p 和加速度 a 来检测突发特征.

Table 1 Comparison of Experimental Results with Different Methods

表 1 本文方法与 TF-IDF&UF-IUF 方法实验结果对比 %

Method	P	R	F
TF-IDF&UF-ITUF	67.00	69.35	68.16
Velocity Modle	80.47	83.06	81.75
Momentum Model	87.60	85.48	86.53

从表 1 中看出,速度模型和动量模型明显优于 TF-IDF&UF-ITUF 模型.实验表明,基于有意义串动量模型的突发特征检测适用于微博数据,能够快速消除大量噪音,在第一时间准确发现突发特征.动量模型的结果优于速度模型,这是因为动量模型中结合了特征的"质量"m,将特征在通常情况下的流通程度和重要性考虑进来,更全面地描述了特征的突发态势.例如,速度模型未检测到 2013-04-05 的"微信收费"话题,而动量模型能够检测到就是因为"微信"的 m 值比较大.

图 2 是从 2013-04-04—2013-04-23 突发话题的数量以及动量模型(momentum modle)和 TF-IDF&UF-ITUF 模型 2 种方法的准确率分布情况,左边纵轴是准确率,右边纵轴是话题个数.从图 2 看出,在 4 月 9,20,21 日这 3 天的突发话题数量较少、话题较集中,2 种方法的检测准确率都比较高而且比较接近.而在 7,8,22 日这 3 天的突发话题数量较多,检

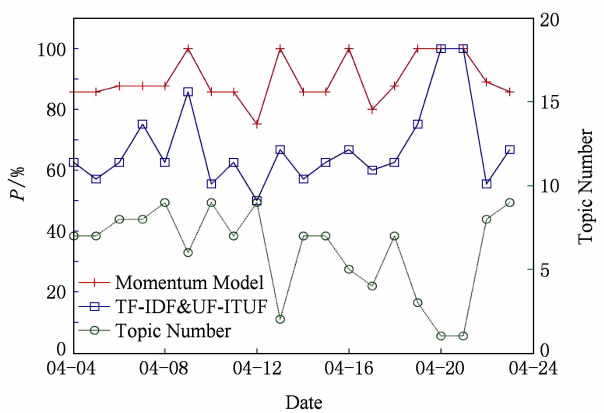


Fig. 2 Time distribution of experimental results.

图 2 实验结果时间分布图

测结果的准确率都相对较低,而且 TF-IDF&UF-ITUF 模型比动量模型的准确率下降幅度更大.产生这一现象的原因是,话题数量较少时,突发话题相对集中,如 2013-04-20 的"雅安地震",其他非突发话题讨论比较少、噪音较小,容易检测出突发话题,2 种方法表现接近.而当突发话题较多时,突发话题比较分散,有些突发特征与其他非突发话题的特征相互交叉、噪音较多,突发话题被淹没,而动量模型比 TF-IDF&UF-ITUF 模型去噪能力更强,准确率下降幅度更小.

3.2.2　有意义串特征与词语特征对比

有意义串是比词语粒度更大的特征,为了对比有意义串特征与词语特征的差别,本文也实现了以词典中词语作为特征,采用动量模型来发现突发话题的方法,与基于有意义串动量模型的方法对比结果如表 2 所示:

Table 2　Comparison of Experimental Results with Meaningful String Features and Word Features

表 2　有意义串特征与词语特征实验结果对比　%

Feature	*P*	*R*	*F*
Word	78.74	80.65	79.68
Meaningful String	87.60	85.48	86.53

从表 2 看出,有意义串特征与词语特征相比,对于提高突发话题发现的准确率、召回率有明显的贡献.表 3 是 2 种方法发现的几个突发话题特征对比:

Table 3　List of Bursty Topic Features

表 3　突发话题特征列表

Date	Meaningful String Features	Word Features
2013-04-12	永州市劳教委 中级人民法院 唐慧败诉	永州市 劳教委 上访
	考驾照 使用教练车 驾校培训	驾校 驾照 公民
	直接考驾照 法律法规 学车	公安部 培训
	卖猪肉 陆步轩 讲创业 杀猪	猪肉 母校 北大
	羽泉 林志炫 我是歌手	歌手 总决赛 歌王
2013-04-13	巴厘岛机场 冲出跑道 狮航	乘客 飞机 巴厘岛 机场 跑道
	凤凰古城 检票口 门票	女友 凤凰 古城
	直系亲属 凤凰副县长	黄田 父母
	北京市卫生局 北京地坛医院	禽流感 病例 确诊
	北京首例 疑似 北京发现 禽流感	感染 北京
2013-04-16	波士顿爆炸 马拉松 终点线 奥巴马	波士顿 起爆 爆炸
	中国女留学生 肯尼迪图书馆 周丹龄	美国 马拉松
	复旦大学 黄洋 室友 复旦研究生 林某	黄洋 投毒 短信
	遭投毒 中山医院 寝室饮水机	研究生 复旦
	赵某 武汉大学 考试作弊	

从表 3 看出,有意义串特征比词语特征更加明确地反映了话题的特征,如 2013-04-12 的"直接考驾照"、"冲出跑道"、2013-04-13 的"北京首例"等直观地描述了话题的关键信息,而从词语特征无法获取这些关键信息.而且,有意义串作为特征检测出了"赵某 武汉大学 考试作弊"这一突发话题,而词语作为特征时未发现.所以,在微博突发话题检测过程中,有意义串比词语更适合表示微博的突发特征.

4　结束语

本文提出了一种基于有意义串动量模型的微博突发话题检测方法.该方法针对微博特征空间动态变化、信息噪音大的特点,首先在时间窗口内提取有意义串作为动态特征;然后根据动量模型对特征建模,从特征能量大小、变化趋势以及二阶变化率方面检测突发特征;最后通过特征互信息合并特征,形成突发话题.实验表明,该方法适用于在线的大规模微博数据流中检测突发话题,准确率和召回率都取得了令人满意的结果.动态提取意义串既能准确发现话题的关键特征,又能对局部微博数据降维.动量模型较好地反映了话题发展过程中特征的能量变化和发展趋势,能够有效地过滤噪音信息,准确检测出特征的突发点.

基于能量模型的微博突发话题发现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)突发特征识别的策略进一步优化,采用学习方法或产生式策略加以整合识别;2)微博富特征的应用,通过充分利用好友关系、链接关系、转发关系等丰富复杂的关联关系,提升突发特征检测的准确性.

参 考 文 献

[1] Fung G, Yu J, Yu P, et al. Parameter free bursty events detection in text streams [C] //Proc of the 31st ACM Int Conf on Very Large Data Bases (VLDB'05). New York: ACM, 2005: 181-192

[2] He Q, Chang K, Lim E. Analyzing feature trajectories for event detection [C] //Proc of the 30th ACM Annual Int Conf on Research and Development in Information Retrieval (SIGIR'07). New York: ACM, 2007: 208-214

[3] Kleinberg J. Bursty and hierarchical structure in steam [C] //Proc of the 8th ACM Int Conf on Knowledge Discovery and Data Mining (SIGKDD'02). New York: ACM, 2002: 91-101

[4] He D, Stott P. Topic dynamics: An alternative model of 'bursts' in streams of topics [C] //Proc of the 16th ACM Int Conf on Knowledge Discovery and Data Mining (SIGKDD'10). New York: ACM, 2010: 443-452

[5] Zhu M, Hu W, Wu O. Topic detection and tracking for threaded discussion communities [C] //Proc of 2008 IEEE/WIC/ACM Joint Conf on Web Intelligences and Intelligent Agent Technology (WI-IAT'08). Piscataway, NJ: IEEE, 2008: 77-83

[6] Chen Y, Cheng X, Yang S. Bursty topics extraction for Web forums [C] //Proc of the 18th ACM Int Conf on Information and Knowledge Management (CIKM'09). New York: ACM, 2009: 55-58

[7] Du Y, He Y, Tian Y. Microblog bursty topic detection based on user relationship [C] //Proc of the 6th IEEE Information Technology and Artificial Intelligence Conf. Piscataway, NJ: IEEE, 2011: 260-263

[8] Diao Q, Jiang J, Zhu F, et al. Finding bursty topics from microblogs [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2012: 536-544

[9] Petrovic S, Osborne M, Lavrenko V. Streaming first story detection with application to twitter [C] //Proc of the 11th Annual Conf of the North American Chapter of the Association for Computational Linguistics (NAACL'10). Stroudsburg, PA: ACL, 2010: 181-189

[10] Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic mining for microblog based on MB-LDA model [J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802 (in Chinese)
(张晨逸, 孙建玲, 丁轶群. 基于微博 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10): 1795-1802)

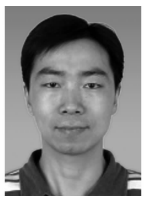
[11] Hong L, Dan O, Davison B D. Predicting popular messages in twitter [C] //Proc of the 20th ACM Int Conf Companion on World Wide Web (WWW'11). New York: ACM, 2011: 57-58

[12] Li C, Sun A, Datta A. Twevent: Segment-based event detection from tweets [C] //Proc of the 21st ACM Int Conf on Information and Knowledge Management (CIKM'12). New York: ACM, 2012: 155-164

[13] Phuvipadawat S, Murata T. Breaking news detection and tracking in twitter [C] //Proc of the 9th IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Technology (WI-IAT'10). New York: ACM, 2010: 120-123

[14] He Min. Web-oriented Chinese meaningful string mining [D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2007 (in Chinese)
(贺敏. 面向互联网的中文有意义串挖掘[D]. 北京: 中国科学院计算技术研究所, 2007)

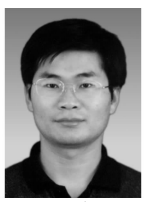
He Min, born in 1982. PhD candidate. Her main research interests include natural language process, Web mining and information security.



Du Pan, born in 1981. PhD. His main research interests include Web search and mining, machine learning and social network (dupan@software.ict.ac.cn).

Zhang Jin, born in 1978. PhD. His main research interests include topic analysis and distributed system (zhangjin@software.ict.ac.cn).

Liu Yue, born in 1971. Associate professor. Her main research interests include information retrieval and Web mining (liuyue@ict.ac.cn).



Cheng Xueqi, born in 1971. PhD. Professor and PhD supervisor in Institute of Computing Technology, Chinese Academy of Sciences. His main research interests include network information security, large-scale information retrieval and knowledge mining (cxq@ict.ac.cn).