社交网络中隐式事件突发性检测

介 飞[1] 谢 飞[2] 李 磊[1] 吴信东[1,3]

摘 要 社交网络与人们的生活息息相关, 其上的用户行为可用于检测社交网络中的事件突发性, 进而准确定位事件的发生区间. 但用户行为易受主观及外部因素的影响, 有时会出现隐式事件突发性, 给事件突发性检测带来困难. 本文针对社交网络中的隐式事件突发性问题, 在以社交行为特征进行事件突发性检测的基础上, 引入关键词特征, 动态调整各个时间窗口的候选关键词, 将不同事件与不同的关键词特征绑定, 避免事件之间及噪音带来的干扰, 实现对隐式事件突发性的准确识别. 相关实验表明, 本文提出的算法可有效改善现有社交网络中事件突发性检测任务的效果.

关键词 突发性, 事件, 检测, 社交网络

引用格式 介飞, 谢飞, 李磊, 吴信东. 社交网络中隐式事件突发性检测. 自动化学报, 2018, **44**(4): 730−742

DOI 10.16383/j.aas.2017.c160564

Latent Event-related Burst Detection in Social Networks

JIE Fei[1] XIE Fei[2] LI Lei[1] WU Xin-Dong[1,3]

Abstract Social networks are closely bound up with our daily life, in which behaviors of users can be used for detection of event-related bursts and further for determination of the time period for each event. But latent event-related bursts, which result from internal or external impacts on users′ behaviors, will be difficult to identify. In this paper, in order to solve the detection problem of latent event-related bursts in social networks, on the basis of event burst detection via social behavior features, we introduce the features of keywords and dynamically change the keyword candidates for each time window, so as to bind different events with different keywords, aiming to avoid interferences from inter-events or noise and discover latent event-related bursts more accurately. Experimental results show that our proposed method can improve the performance of event-related burst detection in social networks compared with existing algorithms.

Key words Burst, event, detection, social network

Citation Jie Fei, Xie Fei, Li Lei, Wu Xin-Dong. Latent event-related burst detection in social networks. *Acta Automatica Sinica*, 2018, **44**(4): 730−742

社交网络深刻影响着大众的日常生活[1], 人们习惯将感兴趣的事件通过社交媒体与他人进行分享和交流. 伴随着事件的发生, 社交网络中相关文本的发布、转发及评论等行为会形成一个密集期, 即表现为行为特征的一个突发性. 突发性背后往往蕴含着事件信息, 可用来发掘潜在的市场需求和隐含的政治倾向, 进而为商业推广或舆情监控提供指导. 相较于传统媒体, 社交网络的公众参与度更高. 因此, 发现社交网络中的事件突发性具有更为重要的现实意义.

突发性即被观测目标的频数等特征值陡然上升的现象. 随着事件的发生, 某些特征值, 例如文档频数, 会急剧上升, 形成事件相关突发性 (Event-related bursts), 简称事件突发性. Kleinberg 首先构建了基于自动机理论的突发性检测模型[2], 用于描述电子邮件中的事件信息. 突发性检测最初是应用在新闻、电子邮件和科研论文等传统媒介中[2−5], 而随着社交网络的兴起, 为突发性检测提供了新的应用环境. 在传统的突发性检测中, 通常以关键词词频信息等文本型特征作为依据, 即考虑了内容信息; 而在社交网络中, 可以利用行为、链接和情感等非文本型特征进行事件突发性检测[1,6−8]. 但据我们所知, 还未有研究人员开展文本型特征与社交行为特征结合的相关研究. 其中, 文本型特征 (例如关键词) 可从语义上直接反映事件发生情况, 能准确判断事件是否发生, 但以其作为突发性检测的特征, 存在如何筛选的问题, 一般只能根据用户意图进

收稿日期 2016-07-31 录用日期 2017-03-21
Manuscript received July 31, 2016; accepted March 21, 2017
国家重点基础研究发展计划 (973 计划) (2013CB329604), 国家自然科学基金 (61503114, 61503116) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329604) and National Natural Science Foundation of China (61503114, 61503116)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 合肥工业大学计算机与信息学院 合肥 230009 中国 2. 合肥师范学院计算机科学与技术系 合肥 230601 中国 3. 路易斯安那大学拉菲特分校计算与信息学院 拉菲特 LA 70503 美国
1. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China 2. Department of Computer Science and Technology, Hefei Normal University, Hefei 230601, China 3. School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette LA 70503, USA

行人工选择, 再按选定的特征变动情况, 判断突发性, 自动化程度较低; 而社交行为特征用于事件突发性检测时, 由于其与事件发生的关系不明确, 可能由于事件交错, 事件突发性程度较低等原因导致漏检或错检. 根据对具体数据的分析, 当前利用社交行为特征进行事件突发性检测的方法不能准确发现图 1 中所示的事件突发性.

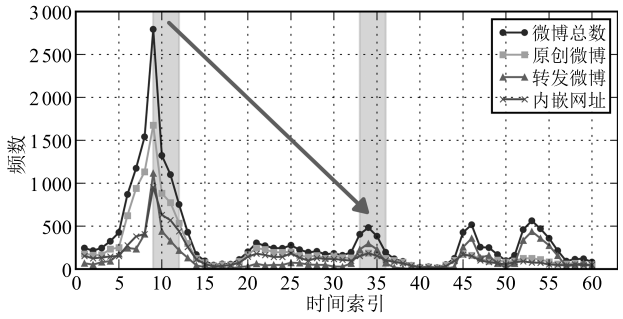


图 1　隐式事件突发性示例

Fig. 1　An example of latent event-related burst

图 1 中的数据爬取自新浪微博, 对应的时间段为 2015 年 10 月 21 日 12 时～24 日 0 时, 共 60 小时. 图中第一段标注区间 (9～12) 内进行了一场亚冠比赛, 恒大 0:0 战平日本柏太阳神队; 第二段标注区间 (33～36) 对应事件为恒大集团与英国相关机构签署协议, 开展项目合作. 由于该事件发生在夜晚 (22 日 21:00 左右, 对应图中索引 33), 因此与之相关的微博活动在事件发生后短时间内上升, 随后迅速下降, 第二天, 又呈现突发状态势 (对应区间 45～48 与 51～57). 图 1 中展示四种社交行为, 微博总数对应用户的发布行为, 原创微博对应用户的原创发布行为, 转发微博对应转发行为, 内嵌网址对应引用外部信息行为. 对比两个事件, 两者发生在连续两天的同一时间段 (相差 24 小时), 从不同行为频数特征的变动情况来看, 第一个事件引起的突发性远大于第二个事件的突发性, 表现为频数值的骤降 (图中箭头所示), 此时, 第二个事件对应区间就易被判别为非突发状态, 造成该事件突发性的漏检. 由图 1 可知, 第二段标注区域所示的事件突发性本身突发模式较为显著, 但由于邻近远高于自身突发性事件的影响, 易被其他事件 "掩盖"[1] 其突发性, 本文称此类事件突发性为隐式事件突发性. 上述类型的隐式事件突发性的发生是由于外部事件的干扰, 还有一类隐式事件突发性, 则是由于事件本身引起, 例如事件发生时, 关注该事件的用户数量不足, 则相应的用户行为 (例如转发、评论、点赞等), 不会发生明显变化, 但用户讨论内容具有明显倾向性, 如某些词语反复出现, 此时再单纯以社交行为进行事件突发性检测, 则会由于相关行为突发性不足造成漏检, 引入内容信息成为解决该问题的选项之一.

本文主要研究事件突发性中的非常规类型 — 隐式事件突发性, 该类事件突发性由于事件本身或外部因素的影响易被漏检, 成为现有事件突发性检测算法的瓶颈. 针对隐式事件突发性, 本文在当前基于行为特征的事件突发性检测方案基础上, 引入关键词特征, 伴随时间的推进, 动态改变各个时间窗口的关键词候选, 实现不同时间区间与不同关键词特征绑定, 进而将不同事件突发性映射到不同特征空间上, 以此剔除噪音及事件之间的互相影响; 随后, 将由关键词特征与行为特征得到的突发性结果关联, 以二者的突发性情况共同决定社交文本流的突发性, 从而更为准确地检测事件突发性. 本文的贡献主要有两点: 1) 首次将文本型 (关键词) 特征与非文本型 (社交行为) 特征结合, 开展事件突发性检测研究; 虽然已有相关文献[9−10] 开展多特征事件检测研究, 但与本文发现事件突发性区间的目标有所区别, 例如, 文献 [9] 只考虑结果是否处于事件发生时间前后的一定范围, 并不关注事件发生区间的确定问题; 2) 在进行以关键词为特征的事件突发性检测时, 本文提出了各时间窗口内候选关键词的筛选方案及多关键词突发性结果关联决定当前时间窗口突发性的策略. 在两个不同类别真实数据集上开展的相关实验表明, 上述方案可以有效提升社交网络中事件突发性检测算法的性能, 对事件检测等相关领域研究具有一定的参考价值.

本文结构如下: 第 1 节对研究的问题进行形式化表述; 第 2 节详细介绍综合两类特征的事件突发性检测算法的步骤; 第 3 节展示在两个真实数据集上的实验结果, 并对结果进行详细分析; 第 4 节介绍事件突发性检测研究领域的相关工作; 第 5 节对本文进行总结并指出未来可能的研究方向.

1　问题表述

本文主要研究社交网络中的事件突发性检测问题, 即在社交网络数据中, 确定由真实事件发生引起的突发性对应的时间区间, 包括确定事件突发性的开始与结束时间窗口, 着重解决现存算法对于隐式事件突发性的漏检问题.

事件突发性 (Event-related bursts), 是由某一真实事件引起的相关特征突发性对应的一段时间区间 $[t_s, t_e]$, t_s 与 t_e 分别表示事件突发区间的开始时间窗口与结束时间窗口. 与特定主题相关的事件突发性一般不止一个, 因此这里用集合表示为 $Busrts$

[1] "掩盖", 指当前突发性判定受临近事件突发性的影响, 并不表明二者时间上有重叠; 当事件重叠时, 相关算法会识别为一次突发性, 并不会影响突发性检测的准确性, 因此不必区分重叠事件.

$= \{[t_s, t_e] | t_s, t_e \in T, s \le e\}$, 其中, T 表示时间窗口序列, s, e 表示突发区间开始与结束对应的时间窗口索引值. 事件突发性与事件并非一一对应关系, 与事件内容、用户行为等因素有关, 一次事件可能引起多次事件突发性.

隐式事件突发性, 指具有以下两类特点之一的事件突发性. 1) 突发模式不明显, 突发程度绝对值较低; 2) 突发程度相对较低, 突发性被邻近突发程度更高的事件 "掩盖". 这两类事件突发性分别根据其特点称为真隐式事件突发性与假隐式事件突发性, 合称为隐式事件突发性. 本文着力解决隐式事件突发性的检测问题, 以提高现有事件突发性检测算法的效果.

本文涉及的其他概念与定义, 借用文献 [1] 中的相关表述, 描述如下:

行为 (Activity), 指话题或事件发生时用户进行的动作, 例如微博中的发布、转发、评论、点赞、嵌入网址链接等操作.

时间窗口序列 (Time window sequence), 一个长为 N 的时间窗口序列表示为 $T = (t_1, t_2, \cdots, t_N)$, t_i 表示第 i 个时间窗口. 将数据集按时间排序, 以等长时间粒度进行切分, 即可得到时间窗口序列.

行为流 (Activity stream), 用数字序列 $H = (n_1^m, n_2^m, \cdots, n_N^m)$ 表示, n_i^m 表示在第 i 个时间窗口内 m 类行为发生的总次数, N 表示时间窗口个数.

词语流 (Term stream), 用数字序列 $W = (n_1^w, n_2^w, \cdots, n_N^w)$, $n_i^w = df_{i,w}$ 表示在第 i 个时间窗口内词语 w 的文档频率, N 表示时间窗口个数.

状态序列 (State sequence), 每个时间窗口 t_i 对应状态 z_i, 由此构成状态序列 $Z = (z_1, z_2, \cdots, z_N)$, z_i 表示第 i 个时间窗口的状态索引值, $z \in \{0, 1, 2, \cdots, N_Z - 1\}$, z 取值为 0 时表示非突发状态, 非 0 表示突发状态, N_Z 表示不同状态数目. 状态索引值反映事件突发程度, 其值越大表示突发程度越高, 突发性检测即指定每个时间窗口的状态索引值, 连续状态索引值非零的时间窗口序列构成一个突发区间.

上述定义示例如图 2 所示, 横轴表示时间窗口, 纵轴表示 Activity 或 Term 特征频数值, 图中折线表示状态序列, 本文选用两种状态 ($z \in \{0, 1\}$), 即只区分突发状态与非突发状态.

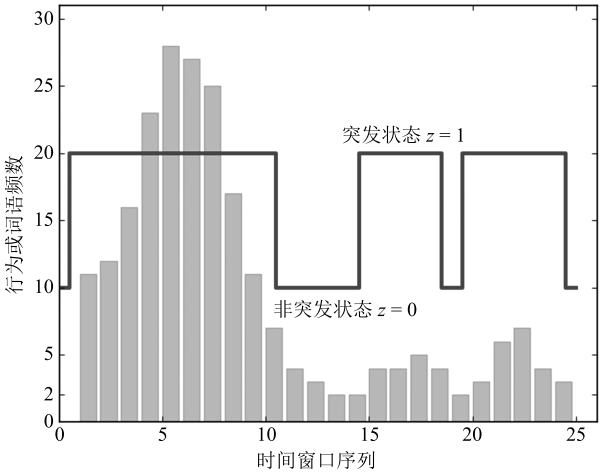


图 2 相关定义示意图

Fig. 2 A schematic diagram of related conceptions

2 方法设计

2.1 思路概述

由前文所述可知, 现有算法不易发现隐式事件突发性, 算法的召回率难以提升, 因此对这类非常规突发性必须提出针对性解决方案, 避免可能的漏检问题. 对于真隐式事件突发性, 事件突发性程度本身较低, 可以考虑引入新的特征表征事件; 在新的特征刻画事件时, 该事件能够表现出较高的突发性; 关键词特征与事件的发生直接相关, 事件发生, 则关键词出现频数大幅上升, 可以满足要求. 对于假隐式事件突发性, 若只关注行为特征, 事件发生时, 用户会产生相似的行为模式 (例如转发和评论等), 易造成时间上邻近的不同事件的 "掩盖" 问题, 而对于关键词特征, 不同事件对应的关键词集合重合度较低, 可将不同的时间窗口与对应的关键词集绑定, 则紧邻的事件由于关键词集的不同, 被映射到不同的关键词特征空间, 从而避免了邻近事件突发性的相互干扰. 综上, 为应对现有事件突发性检测算法对于隐式事件突发性的漏检问题, 文本型信息的引入是一个可选的方案, 本文提出的算法即基于此思路, 将社交行为特征与关键词特征结合, 解决隐式事件突发性的漏检问题, 从而提升事件突发性检测的整体效果.

2.2 突发性检测方法

本文使用文献 [1] 中提出的单目标序列与多目标[2] 序列突发性检测算法.

众所周知, 丰富的社交特征给我们提供了多样的数据来源, 但社交媒体普遍存在的噪音问题也阻碍传统方法直接应用在社交网络中. 因此, Zhao 等根据 Twitter 内容突发性的特点, 提出了适用于社交网络数据的单目标序列与多目标序列突发性检测算法, 构建了三类成本, 对社交网络中的消息生成进行建模, 包括生成成本、平滑成本以及跨目标流成本[1].

[2] 目标, 即指特征, 单目标序列表示算法输入为单一类别特征序列, 例如行为特征, 算法输入只有一种行为流时, 则为单目标, 当输入多种行为流时, 即为多目标.

生成成本 (Generating cost), 表示根据特定的概率分布, 当前时间窗口 i 在状态 z_i^m 下某个特定特征 m (例如社交行为) 出现次数 n_i^m 时的成本, 可取概率的对数负值, 此时概率越大, 对应成本越低. 概率分布可选用二项分布、泊松分布或指数分布. 使用泊松分布时, 概率分布函数 $f(n_i^m, i, z_i^m)$ 具体形式为 $(\mu_{z_i^m})^{n_i^m}\exp(-\mu_{z_i^m})/n_i^m!$, 其中 μ_0 表示一个时间窗口内特征频数的平均值, 如果处于突发状态, 目标特征会以更高的速率发生, 从而导致较高的期望 μ_1, 可以设置 $\mu_1 = \mu_0 \times \rho$, $\rho > 1$, 为参数.

平滑成本 (Smoothness cost), 倾向于在标注时保持突发状态序列稳定, 实现剔除噪音, 处理数据随机波动的功能. 通常, 与事件相关的突发性会由于人们的持续关注而维持一段较长时间且波动较小, 而诸如广告等噪音信息带来的突发性, 更多时候出现时间较为短暂, 因此可以突发性延续的时间长短判断该突发性是由真实事件引起或由噪音引起. 其中一种衡量方案为

$$\Phi(z_1, z_2, \cdots, z_N) = \sum_{s_i < e_i} (e_i - s_i + 1)^2 \tag{1}$$

其中, s_i, e_i 分别表示第 i 个状态值相同的序列开始与结束时间窗口索引, 式 (1) 表示将状态序列中状态值相同的区间长度进行平方求和.

例如, 假设突发状态为二状态, 即只区分突发状态与非突发状态, 则一系列时间窗口对应一系列状态序列, 如 "0000100000" 与 "0000000000", 按式 (1) 计算平滑指标分别为 42 ($4^2 + 1^2 + 5^2 = 42$) 与 100 ($10^2 = 100$), 平滑指标取负值即可作为区别噪音与正常突发性的成本值, 在此例中, 如果指定第 5 个时间窗口出现突发状态, 其维持时间仅一个时间窗口, 时间较短, 显然为噪音的可能性较大, 因此其平滑指标较小 (取负值为 -42, 与没有突发性的序列的平滑成本 -100 比较, 成本较大).

跨目标流成本 (Cross stream cost), 借助上述思想, 在具有相关性的多目标序列中, 不同目标的突发模式类似, 因此多个目标序列的同一时间窗口的状态也应该趋同, 否则应给予一定的惩罚成本 (即跨目标流成本).

$$\sum_{i=1}^{N} \sum_{m_1, m_2} \Gamma(z_i^{m_1} \neq z_i^{m_2}) \tag{2}$$

其中, $\Gamma(\cdot)$ 为指示函数 (Indicator function), m_1 与 m_2 对应任意两类特征, 若其同一时间窗口内的状态值 $z_i^{m_1}$ 与 $z_i^{m_2}$ 不相等, 则取值为 1, 计入成本, 否则成本为 0.

由上述三类成本我们可以构建单目标序列与多目标序列突发性检测的成本模型 (分别记为 $SCost$ 与 $MCost$), 其中多目标序列成本模型比单目标序列成本模型额外考虑不同目标序列之间的成本, 具体为

$$SCost(\boldsymbol{z}) = \underbrace{-\sum_{i=1}^{N} \log f(n_i^m, i, z_i^m)}_{\text{generating cost}} + \underbrace{(-\Phi(z_1^m, \cdots, z_N^m) \times \gamma_1)}_{\text{smoothness cost}} \tag{3}$$

$$MCost(\boldsymbol{z}) = \sum_{m=1}^{M} \left\{ -\sum_{i=1}^{N} \log f(n_i^m, i, z_i^m) - \Phi(z_1^m, \cdots, z_N^m) \times \gamma_1 \right\} + \underbrace{\sum_{i=1}^{N} \sum_{m_1, m_2} \Gamma(z_i^{m_1} \neq z_i^{m_2}) \times \gamma_2}_{\text{cross stream cost}} \tag{4}$$

式 (3) 和式 (4) 中 M 和 N 分别表示特征类别与时间窗口数目, γ_1 和 γ_2 为参数, 用于调节不同类别成本之间的权重.

构建成本模型后, 利用动态规划算法可得总成本最小时文本流中各个时间窗口的突发状态, 具体算法可参考文献 [1−2], 处于突发状态的连续时间窗口即可构成突发区间, 由此实现突发性检测任务.

2.3　算法步骤

2.3.1　文本型特征筛选

在文献 [1] 的算法基础上引入文本型 (关键词) 特征, 词语的选择使用文献 [11] 中的关键词选择算法. 计算公式为

$$WScore_{i,w} = \frac{df_{i,w} + 1}{\log\left(\dfrac{\sum\limits_{j=i-L}^{i} df_{j,w}}{L} + 1\right) + 1} \tag{5}$$

$WScore_{i,w}$ 表示词语 w 在第 i 个时间窗口的 $WScore$ 值, $df_{i,w}$ 表示词语 w 在第 i 个时间窗口的文档频率, L 表示所考虑历史时间窗口个数, 为可调参数. 本文中, 一篇文档指时间窗口内的一条微博, 故文档频率 $df_{i,w}$ 即第 i 个时间窗口内包含词语 w 的微博条数.

式 (5) 中分子表示词语在当前时间窗口的文档频率, 分母计算词语在历史时间窗口的出现情况, 只有在当前窗口出现较多, 历史窗口出现较少的词语 $WScore$ 值较大, 故该值可较好地反映一个词语的

权重, 选出对于当前时间窗口最有代表性的词语.

在计算得到每个词语的 $WScore$ 值后, 递减排序, 抽取每个时间窗口 Top n 个词语中的名词作为关键词候选. 随着时间推进, 事件发生, 每个时间窗口对应的关键词候选集随之变化, 关键词与时间窗口的绑定, 将不同事件映射到不同关键词特征上, 消除噪音及事件之间的互相干扰, 从而提高识别效果. 具体效果如图 3 所示.

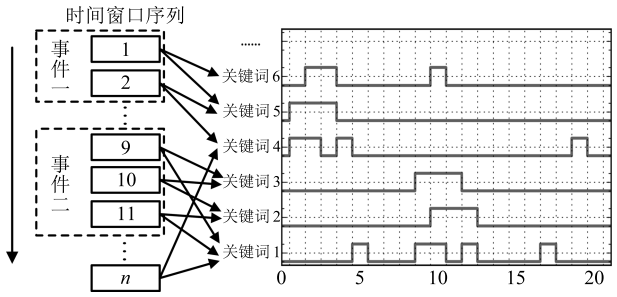


图 3 关键词特征作用示意图

Fig. 3 The schematic diagram of keyword feature relations

在得到各时间窗口的候选关键词后, 应用前述突发性检测算法, 可以发现每个候选词的突发区间.

2.3.2 关键词区间关联

在得到时间窗口内各个候选关键词的突发情况后, 需将多个关键词的突发区间关联, 共同决定当前时间窗口是否处于突发状态, 最终得到文本型特征突发区间. 为实现关键词突发区间的关联, 本文采用阈值法, 即当前时间窗口内关键词处于突发状态的比例超过阈值时, 则判定该时间窗口处于突发状态. 使用的计算公式为

$$z_i = \Gamma\left(\frac{\sum\limits_{w\in KW_i} z_{i,w}}{|KW_i|} \geq \lambda\right) \tag{6}$$

其中, z_i 表示第 i 个时间窗口的突发状态, $z_{i,w}$ 表示词语 w 在第 i 个时间窗口的突发状态, 其值为 0 或 1, λ 为阈值, 超过此值则代表当前时间窗口处于突发状态, $\Gamma(\cdot)$ 为指示函数, 决定是否处于突发状态, KW_i 表示当前时间窗口 i 的候选关键词集合.

2.3.3 突发区间优化

关键词作为目标时, 突发性检测算法得到的突发区间结果会发生碎片化现象, 原本完整的突发区间被分割为数段小区间, 造成这种现象的原因为关键词候选较多, 较之行为特征易受噪音 (非相关词) 影响. 为应对此现象, 提出两点假设: 1) 若候选词与特定事件相关度高, 则该词语会被反复提及, 因此其突发状态会维持一段时间, 否则, 对应突发区间为噪音的可能性较大, 应予以舍弃; 2) 若临近的两个被判定为处于突发状态的区间具有较为相似的关键词集合, 则表明这两个时间区间表现出的突发性与同一事件相关, 应予以合并, 构成新的突发区间.

上述两点假设符合对于事件发生时用户发布内容行为的基本判断. 对于第一点, 人们在相关事件发生时, 会以较高频率提及一些词语并持续一段时间, 因此, 当词语的突发性区间过短时, 可能只是数据的随机波动或噪音, 而突发性维持较长时间的词语, 则更有可能与用户关注的事件相关. 对于第二点, 在事件发生时, 人们讨论事件往往有特定的关注方面, 这样, 同一事件在连续数个时间窗口的关键词集应该具有较高重复性, 反之, 连续几个关键词集具有较高重复性的区间为讨论同一事件的概率亦大增, 可以进行合并. 基于以上两点假设, 可得区间优化算法.

输入区间集合 $inputIntervals$, 由关键词得到的突发区间组成, 按时间排序, 输出集合 $outputIntervals$ 为空, 每个时间窗口对应的关键词集合为 KW_i, i 为时间窗口索引, 突发区间对应的关键词集合由突发区间对应的时间窗口关键词集合取并集生成, 对于 $inputIntervals$ 集合中的突发区间按顺序逐个处理, cur, $next$, $third$ 分别指向 $inputIntervals$ 中当前第 1, 2, 3 个待处理的突发区间.

步骤 1. 若 cur 与 $next$ 之间时间窗口间隔 $SEP(cur, next) \leq \lambda_1$, 转步骤 2, 否则转步骤 3;

步骤 2. 若区间 cur 的关键词集合 KW_{cur} 与下一个突发区间 $next$ 关键词集合 KW_{next} 重合度 $TOR(cur, next) \geq \lambda_2$, 转步骤 4, 否则转步骤 3;

步骤 3. 若当前突发区间长度 $LEN(cur) \geq \lambda_3$, 转步骤 5, 否则转步骤 6;

步骤 4. 合并 cur 与 $next$ 形成新的 cur, $next = third$, $third$ 指向随后的一个突发区间, 转步骤 1;

步骤 5. 将 cur 指向的突发区间移入 $outputIntervals$, 转步骤 6;

步骤 6. $cur = next$, $next = third$, $third$ 指向随后的突发区间, 若 cur 指向 $inputIntervals$ 中最后一个区间, 则整个算法结束, 此时 $outputIntervals$ 即为优化后的区间集合, 否则转步骤 1 继续执行.

区间优化算法流程图如图 4 所示.

上述步骤中对于突发区间之间的时间窗口间隔 SEP 与突发区间对应关键词集合重合度 TOR 的阈值限制保证合并的突发区间时间相近, 语义相关, 以满足第二点假设; 突发区间的长度 LEN 的阈值限制保证只有较长的突发区间才能成为事件突发性, 对应第一点假设. 关键词集合重合度 TOR 使用 $Jaccard$ 系数衡量.

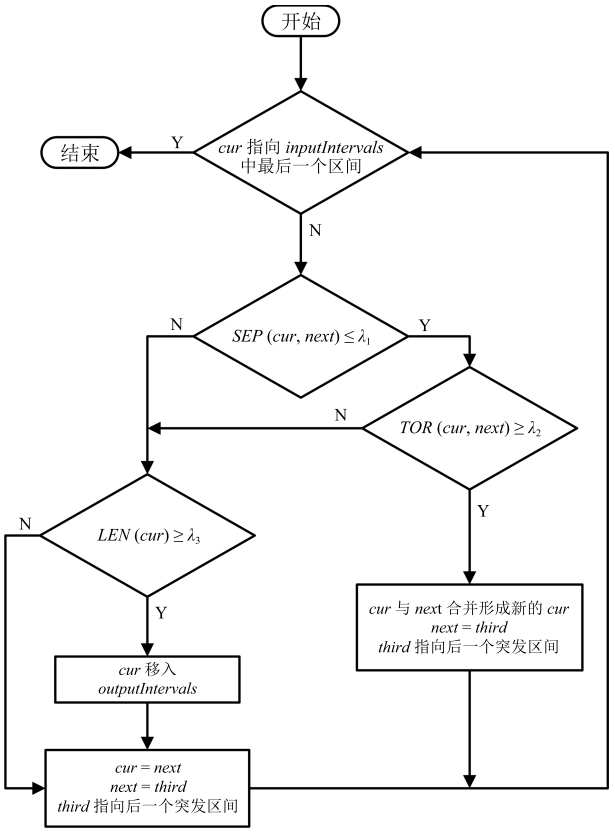


图 4　区间优化算法流程图

Fig. 4　The flow chart of interval optimization algorithm

下面介绍阈值 λ_1, λ_2, λ_3 的设定, 其中 λ_1 值表示由相同事件引起的突发区间的间隔大小 (以间隔时间窗口个数衡量), 其值过大会将不同事件突发区间合并, 造成错误, λ_3 反映事件突发性维持时间的最小值, 其值过小会引入随机波动, 上述两个参数均根据数据集取经验值 3 小时. λ_2 表示关键词重合度, 反映临近突发区间语义相似度, 本文取值为 0.5.

2.3.4　关键词区间与社交行为区间关联

至此, 得到分别根据社交行为与关键词特征得到的突发区间, 接下来介绍两组结果的关联策略, 以得到最终的突发区间. 本文提出三种方案以供实验.

交集策略 (Conjunct): 如果一个突发区间与其他突发区间不重合, 则忽略; 如果一个突发区间与某一区间重合, 则取交集.

$$[t_s, t_e] = [t_{\max(s^1,s^2)}, t_{\min(e^1,e^2)}] \tag{7}$$

其中, s^i, e^i 表示突发区间 i 的开始与结束时间窗口索引值.

并集策略 (Disjunct): 如果一个突发区间与其他突发区间不重合, 则其单独构成一个突发区间; 如果一个突发区间与某一区间重合, 则取并集.

$$[t_s, t_e] = [t_{\min(s^1,s^2)}, t_{\max(e^1,e^2)}] \tag{8}$$

混合策略 (Hybrid): 如果一个突发区间与其他突发区间不重合, 则其单独构成一个突发区间; 如果一个突发区间与某一区间重合, 则取交集.

例如, 现有关键词区间 (以窗口的突发状态序列表示, 0 值表示对应窗口不发生突发性, 1 表示发生突发性) "001111000000", 社交行为区间 "011110001110". 使用交集策略结果为 "001110000000"; 使用并集策略结果为 "011111001110"; 使用混合策略的结果为 "001110001110".

当两类特征发现的突发区间区别不大时, 交集策略与并集策略结果差异较小, 当两类特征发现的突发区间区别较大时, 交集策略与并集策略结果差异较大, 因此可以根据交集策略与并集策略的实验结果判断两类特征对于发现事件突发性的作用是否相同, 从而验证引入的文本特征是否可以弥补行为特征的缺陷, 发现隐式事件突发性.

经过上述步骤, 得到最终的事件突发区间集合. 完整的事件突发性检测方法流程如图 5 所示.

3　实验

本节介绍实验细节, 讨论不同算法的实验结果并分析原因; 针对本文提出算法, 对比使用不同关联策略时的实验结果, 分析原因; 指出单独使用文本特征时效果较差的原因; 解释综合文本与社交行为特征的算法改善事件突发性检测效果的机制, 并结合实例进行分析.

3.1　数据集

微博[3] 是一种通过关注机制分享简短实时信息的广播式的社交网络平台, 已成为目前最流行的社交平台之一[12]. 本文实验数据集以真实微博数据构建, 通过微博提供的搜索及高级搜索功能, 利用网络爬虫程序定时爬取微博数据, 构建实验数据集. 根据搜索关键字的不同, 共获得两个数据集.

3.1.1　数据集 1

以 "恒大"[4] 作为查询关键字, 利用爬虫程序爬取微博搜索页面结果, 定期 (10 分钟) 执行, 共获得微博 165 644 条, 时间跨度为 2015 年 9 月 16 日 0 时～2015 年 11 月 3 日 0 时, 共 48 天, 1 152 小时. 在获得的微博中, 原创微博占比 56.83 %; 转发微博占比 43.17 %; 内嵌网址微博占比 41.72 %.

[3] http://www.weibo.com/

[4] 中国职业足球队名称, 亦是企业恒大集团简称, 涉及地产、酒店、体育及文化等产业.

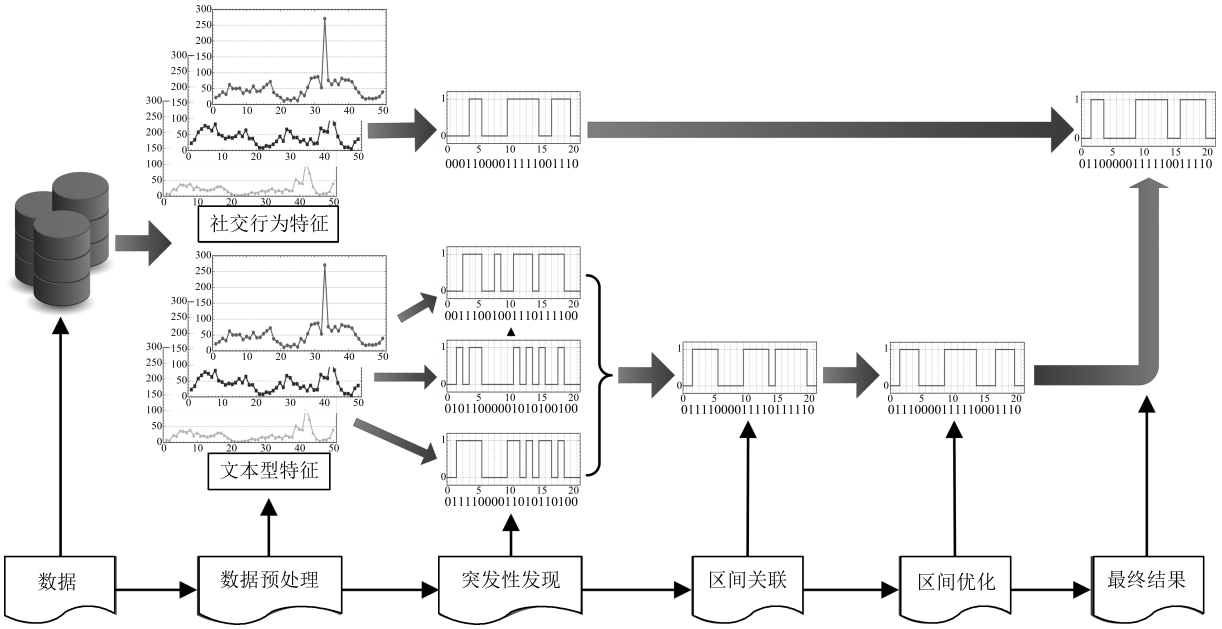


图 5 社交网络中事件突发性检测方案流程示意图

Fig. 5 The flow diagram of event-related burst detection in social networks

3.1.2 数据集 2

以 “爸爸去哪儿”[5] 作为查询关键字, 爬取微博, 共获得微博 154 366 条, 时间跨度为 2015 年 9 月 22 日 0 时 ∼ 2015 年 11 月 7 日 0 时, 共 46 天, 1 104 小时. 在获得的微博中, 原创微博占比 50.41 %; 转发微博占比 49.59 %; 内嵌网址微博占比 27.07 %.

以上数据集分别记作 HD 与 BA. 数据集 HD 与 BA 涵盖体育与娱乐类内容, 涉及人们关注的主要方面, 因此数据集选取具有普遍性. 关于数据集突发区间的确定, 本文爬取了主要门户网站 (包括新浪、搜狐、网易) 的相关专题新闻及搜索引擎 (百度) 特定关键词加时间的搜索结果, 根据这两类数据, 人工选择出较为热门的事件, 结合真实事件发生的时间区间, 确定事件突发区间的开始时间与结束时间, 作为实验评价时的真实突发区间集合.

3.2 评价指标

本文使用文献 [1] 中的评价指标, 突发区间重合率 (Bursty interval overlap ration, BIOR), 定义如下:

$$BIOR(f,\chi)=\frac{\sum\limits_{f'\in\chi}\Delta l(f,f')}{L(f)} \tag{9}$$

其中, f 是一个突发区间, $\Delta l(f,f')$ 是 f' 与 f 重合的长度, $L(f)$ 是突发区间 f 的长度. χ 是一组突发区间, BIOR 用于衡量一组突发区间 χ 对于突发区间 f 的覆盖比例. 由此可以定义准确率 (Precision)、召回率 (Recall) 和 F 值, 计算公式如下:

$$R=\frac{\sum\limits_{f\in B}\Gamma\left(\frac{1}{|M_f|}BIOR(f,M)>0.5\right)}{|B|} \tag{10}$$

$$P=\frac{1}{|M|}\sum_{f'\in M}(BIOR(f',B)) \tag{11}$$

$$F=\frac{2\times P\times R}{P+R} \tag{12}$$

其中, M 为通过相关候选算法发现的突发区间集合, B 是真实突发区间集合, M_f 是在集合 M 中与 f 重合的突发区间集合. $\Gamma(\cdot)$ 是指示函数, 当且仅当条件为真时函数值为 1.

3.3 对比算法

本文实验的三种算法, 依次为:

SingleBurstDetector: 使用概率分布及自动机理论构建突发性检测模型[1], 不同突发状态对应分布的参数不同, 因而不同状态时生成同一特征序列的概率大小不同, 即成本不同, 通过最小化成本可得最优的突发状态序列, 进而得到突发区间.

MultiBurstDetector: 原理同 SingleBurstDetector, 但输入为多类行为特征, 方法考虑了不同特征的突发情况, 可以应对噪音对单一特征的干扰[1].

CombinedDetector: 这是本文提出的方法, 综合了关键词特征与社交行为特征, 能够发现隐式事

[5] 一档亲子类综艺真人秀节目名称, 节目有中国版与韩国版.

件突发性, 提升算法性能.

上述三种方法分别简记为 Single、Multi 和 Comb.

3.4 实验步骤

3.4.1 数据预处理

预处理阶段包括分词、去停用词和词性标注等操作, 针对分词和词性标注在微博环境中效果不佳的问题, 可利用分词器提供的新词发现功能以及引入人工构建的外部字典解决. 然后对微博数据按原创、转发、内嵌网址、是否提到其他用户 (含 "@" 符号) 进行分类. 将微博按时间排序, 时间窗口设置为 1 小时, 统计每个窗口内各类别特征出现的频数, 构成前述行为流 (Activity stream). 本文使用 5 种行为流, 分别为: 微博总数 (代表用户发布行为)、原创微博 (代表用户原创内容发布行为)、转发微博 (代表转发行为)、内嵌网址微博 (代表引入网址行为)、含 "@" 符号微博 (代表提到其他用户行为), 此设置与文献 [1] 相同. 计算各时间窗口内词语的 $WScore$ 值, 提取 Top n 词语中名词性词语作为候选词, 构建各个候选词的 Term stream.

3.4.2 计算事件突发性

对以上 Activity stream, 分别应用 Single, Multi 和 Comb 算法, 其中, Single 算法使用单一行为特征, Multi 和 Comb 算法同时使用多种行为特征, 得到事件突发性结果; 对于 Term stream, 应用单目标突发区间检测算法 (Single)[6] 进行突发性检测, 得到各个词语的事件突发性区间, 再将其与多类别行为特征的事件突发性结果关联, 进行区间优化后, 最终得到 Comb 算法的突发性检测结果.

3.4.3 实验对比

根据前述评价指标计算各个算法的准确率、召回率和 F 值, 比较不同算法的结果, 分析各个算法的效果及特点.

3.5 结果及分析

运行前述算法, 实验测试不同参数设置, 此处仅列出最优结果, 如表 1 和表 2 所示. 各个突发性检测算法涉及的参数较多, 在此不再列出, 仅给出 Comb 算法最优结果时的参数设置, 以供参考, $n = 5/5$, $\gamma_1 = 1.9/1.9$, $\gamma_2 = 10.5/11.5$, $\rho = 3/8$, $L = 5/5$, $\lambda = 0.6/0.7$, $\lambda_1 = \lambda_3 = 3$, $\lambda_2 = 0.5$ (两个数据集的参数设置以 "/" 分隔). 针对 Single 算法, 本文测试了前述 5 种社交行为, 包括微博总数、原创、转发、内嵌网址、含 "@" 符号微博 (分别记为 all, post, repost, url, user), 这 5 种特征基本覆盖了典型的社交行为, 具有普遍意义. 对于 Multi 算法, 同时使用 3 种行为特征 (post, repost, url) 进行实验, F 值指标显示 Multi 算法优于前两种算法, 印证了文献 [1] 中的相关结论. 对于 Comb 算法, 我们在多特征的基础上测试前述 3 种关联策略. 实验中也验证了单独使用文本特征时的效果, 如表 3 所示.

表 1　数据集 HD 上各算法实验结果
Table 1　The experimental results of different algorithms on dataset HD

实验项目		实验结果		
Method	Feature/Strategy	P	R	F
Single	all	0.9000	0.3846	0.5389
	post	0.8352	0.3462	0.4894
	repost	**0.9902**	**0.5385**	**0.6976**
	url	0.6803	0.3846	0.4914
	user	0.6573	0.4615	0.5423
Multi	post + repost + url	**0.9525**	**0.6923**	**0.8018**
Comb	conjunct	1.0000	0.5385	0.7000
	disjunct	**0.8256**	**0.9231**	**0.8716**
	hybrid	0.9949	0.6923	0.8165

表 2　数据集 BA 上各算法实验结果
Table 2　The experimental results of different algorithms on dataset BA

实验项目		实验结果		
Method	Feature/Strategy	P	R	F
Single	all	**0.9662**	**0.4000**	**0.5658**
	post	0.9740	0.2000	0.3319
	repost	0.8640	0.3000	0.4454
	url	0.2574	0.1333	0.1757
	user	0.7346	0.3333	0.4586
Multi	post + repost + url	**0.8787**	**0.4667**	**0.6096**
Comb	conjunct	0.9554	0.2667	0.4170
	disjunct	**0.9030**	**0.5333**	**0.6706**
	hybrid	0.8051	0.5667	0.6652

表 3　单独使用关键词特征时实验结果
Table 3　The experimental results with only keyword features

数据集	实验结果		
	P	R	F
HD	**0.7709**	**0.7692**	**0.7701**
BA	**0.6327**	**0.3667**	**0.4643**

[6] 此处使用单目标算法, 是由于多目标算法基于假设: 在特定事件发生时, 不同行为具有一致的突发模式, 而词语由于候选集合较大, 语义多样, 相关性无法保证, 因此不适用多目标算法.

对比不同算法以及同一算法使用不同特征或关联策略时的实验结果, 可得到一系列有价值的结论.

1) Single 算法实验结果分析. 该算法引入了区分噪音与事件突发性的平滑成本等措施, 大幅提升了事件突发性检测的准确率, 在两组数据集上准确率均较高, 但其召回率最低, 并且算法准确率波动性很大. 造成此类结果的原因, 在于不同行为与事件突发性的关系不同, 当某些事件发生与某一行为关系紧密时, 则利用此行为特征检测到的突发性基本都与这些事件有关, 即算法发现的突发区间是真实事件的突发区间的概率较大, 此时算法的准确率 (P 值) 就会很高; 但当该行为与某类事件关系不紧密时, 此类事件发生, 对应行为变化不明显, 则利用该行为进行突发性检测, 就会造成漏检, 进而拉低召回率 (R 值). 因此, 基于单一行为特征算法的效果优劣很大程度上取决于使用的行为特征与事件的关系. 图 1 也可以证实此结论: 在 50 ∼ 55 区间内, 事件发生 (恒大集团与英国相关机构合作), 微博总数与转发微博都有明显的上升, 而原创微博与内嵌网址微博并无明显变化, 说明不同行为对事件的反应不同.

2) Multi 算法实验结果分析. 该算法的准确率较 Single 算法在两个数据集上均有所降低, 但其弥补了 Single 算法召回率过低的缺陷, 从而在衡量算法整体性能的 F 值指标上优于 Single 算法. 分析 Multi 算法召回率提升的原因, 在于多种行为特征加强了行为特征与事件的关系, 避免单一行为特征由于与事件相关性不足或随机波动造成的漏检, 因而召回率上升; 而准确率的下降是由于该算法在根据每个单一特征突发性检测结果生成最终的突发区间时使用了并集策略[1], 即只要一个特征将当前时间窗口标注为突发状态, 就认为这个时间窗口产生突发性, 因而多类特征的噪音都会引入到 Multi 算法结果中来, 使其准确率下降.

3) 不同关联策略实验结果分析. 针对 Comb 算法, 本文测试了 3 种关联策略. 由表 1 和表 2 可知, 在进行文本特征与社交行为特征融合发现事件突发性时, 采用并集 (Disjunct) 处理是进行区间关联的最优策略. 分析不同的关联策略, 可以看出, 交集 (Conjunct) 策略保留文本特征与社交行为特征共同的结果, 因此获得优于 Multi 算法的准确率, 但是由于忽略了仅由单一类别特征得到的结果, 召回率较差; 并集策略与混合 (Hybrid) 策略均保留仅由单一类别特征得到的结果, 因此实现了较高的召回率, 而上述两种策略的准确率取决于关联前两类特征分别的准确率, 因而准确率有升有降. 并集策略取得最优, 而交集结果较差说明, 两类特征在进行事件突发性检测时的作用并不相同, 后文给出具体分析.

4) 单独使用文本特征实验结果分析. 由表 1、表 2 和表 3 对比可知, 未进行融合, 单独使用文本特征时, 实验结果较使用行为特征的差, 这是因为词语候选集庞大, 噪音词较多, 造成使用文本特征发现的突发区间较短, 易被噪音信息割裂, 引入噪音区间, 发生前述的碎片化现象, 导致结果较差.

5) 文本与行为特征特点及融合效果分析. 通过对比单独使用文本特征与行为特征所发现的突发区间, 我们发现: a) 行为特征属于宏观特征, 对于引起较高关注的事件, 才会表现出较为明显的对应行为的突发性 (必须有大量的用户参与, 才能造成行为的突发表现), 即行为特征对于事件的弱突发性敏感度不够. 以用户行为作为特征时发现的突发区间对应的事件关注度普遍较高, 并且突发性维持的时间较长. b) 文本特征属于微观特征, 对在小范围内引起有限突发性的事件也会有所反映, 例如用户单位时间内发布微博的数目波动很小, 即发布行为突发性弱, 此时以该行为进行突发性检测容易失效, 但只要有部分微博集中讨论同一事件, 则也会表现出相关词语的突发性, 即文本特征对事件突发性更为敏感, 能够发现事件的弱突发性 (真隐式事件突发性). 另外由于本文提出的方法将不同事件与不同的关键词绑定, 消除了突发程度高的事件对于突发程度低的事件的影响, 从而解决假隐式事件突发性问题. 综上, 两类特征对于发现的事件突发性类型各有侧重, 社交行为特征容易忽略突发程度低的事件, 而文本特征会很好地弥补此缺陷, 因此本文提出的融合两类特征的综合方法具有较好的效果.

6) 案例分析. 结合上述分析, 回顾图 1, 具体展示本文所述方法的作用效果. 图 1 呈现了两个引起突发性的事件 A 和事件 B, 事件 A 是一场足球比赛, 事件 B 是恒大集团与英国相关机构合作, 关注同一行为特征时, 事件 A 的突发程度远高于事件 B, 如图 6 左侧所示[7]. 当使用关键词特征时, 由于事件 A 和事件 B 不同的关键词, 如表 4 所示 (删除线标注为查询词 “恒大”), 事件 A 的关键词在事件 B 发生时不会突发, 反之亦然, 如图 6 右侧所示, 关注文本特征时, 避免了事件之间的影响, 发现由事件 B 所引起的隐式事件突发性, 从而提高事件突发性检测的性能.

4 相关工作

突发性检测问题, 最早在文献 [2] 中提出, 作者根据电子邮件文本流中话题出现时邮件数量陡增的现象, 引出流式数据中突发性的形式化表述, 并探讨了流式数据中的层次结构问题. 作者借助自动机的思想, 将文本流数据根据时间切分为时间窗口, 根据

[7] 出于图表直观考虑, 图 6 仅为模拟图, 具体数值与真实情况并不对应.

突发程度及历史信息确定当前时间窗口的状态, 并对突发状态的生成与转换成本进行建模, 利用动态规划方法求解, 得到各个时间窗口的突发状态. 文献 [4] 使用卡方测试的方法, 进行词语的突发性检测, 再对发现的突发词语进行聚类, 获得数据集的事件话题. 文献 [13] 提出了无需调整参数的概率方法, 用于在报纸文章中寻找不同时间窗口的突发特征, 确定突发事件的热度区间. 文献 [14] 针对时序数据, 提出基于概率统计模型的变化节点发现方法. 以上研究均在传统文本 (电子邮件、新闻文本和科研论文等) 中进行.

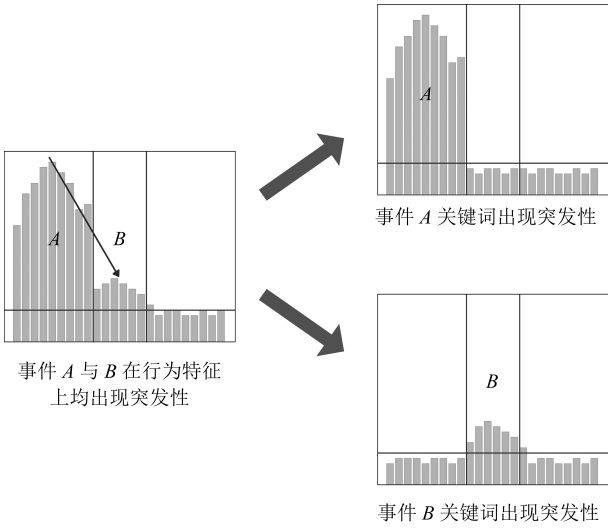


图 6　Comb 方法作用示意图

Fig. 6　The schematic diagram of method Comb

表 4　事件 A, B 的关键词提取结果

Table 4　Extracted keywords of event A and B

时间窗口	关键词 (Top 3)
2015-10-21 19 时	~~恒大~~、决赛、亚冠、广州
2015-10-21 20 时	~~恒大~~、决赛、亚冠、广州
2015-10-21 21 时	~~恒大~~、决赛、亚冠、进
2015-10-22 19 时	~~恒大~~、英国、峰会、工商
2015-10-22 20 时	~~恒大~~、集团、英国、峰会
2015-10-22 21 时	~~恒大~~、英国、峰会、工商

随着社交网络的兴起, 科研人员开始进行社交网络中突发性检测与应用的工作. 文献 [1, 6−10, 15] 基于 Twitter 文本流, 开展关于在社交网络中的突发事件与话题的发现. 其中, 文献 [15] 应用了词语的突发性, 但与本文的研究问题有所区别, 且其仅使用了词语, 并未综合社交行为信息; 文献 [6] 基于事件发生时频数特征的上升下降模式确定事件突发区间, 该方法易受噪音 (例如广告信息) 与多事件交错的影响, 整体效果有待提高; 文献 [7] 着重利用社交网络中的链接异常进行话题发现, 使用了概率方法进行事件发生的预测, 是基于非内容特征进行社交网络挖掘的典型示例; 文献 [8] 使用情感符号作为特征, 利用情感突发趋势发现事件; 引入链接与情感符号的行为与用户的使用习惯紧密相关, 但不是所有事件相关的社交文本都包含此类特征, 限定了这两类方法的使用范围; 文献 [9] 研究了异构网络挖掘问题, 使用社交网络中的多种信息构建异构网络, 将每个特征节点视作传感器, 利用统计方法统一不同属性的异常变化情况, 最后使用改进的图扫描算法发现变化最大的子图, 实现突发事件的检测, 发现事件的形式为 (地点、日期), 将研究任务分为预测 (日期在真实事件发生之前) 及发现 (日期在真实事件发生之后), 在事件发生前后 7 天内的结果都视作发现事件, 但并未强调对于事件突发区间的覆盖, 而本文研究的是如何发现事件发生时导致的突发区间, 目标是尽量使算法得到的区间覆盖真实事件发生的区间并找到足够多的真实区间, 因此与本文研究问题有所区别; 文献 [10] 研究突发性事件的热度预测问题, 考虑了内容特征、用户影响力与历史信息, 对突发事件进行流行度预测, 流行度预测是在已知事件发生的情况下进行的, 可以在任意时刻开始, 作者并不关注事件的开始与结束时间, 该论文方法不能直接应用到事件突发性发现任务中; 文献 [1] 改进了文献 [2] 中基于自动机与概率方法的突发性检测方法, 首次利用社交网络中的行为信息进行事件突发性检测, 引入去噪措施, 指出单一行为特征在发现事件突发性时的不稳定性, 提出利用多类行为特征的检测方法, 在大规模社交网络数据集上验证了其有效性. 本文基于文献 [1] 的算法, 添加文本型 (关键词) 信息, 实现了社交行为与内容两方面信息的综合, 有利于消除事件之间的相互影响, 可以发现前述的隐式事件突发性, 从而更为完整地发现突发事件对应的时间区间, 改善相关算法性能.

突发性检测算法最初用来挖掘文本流突发性背后蕴含的真实事件, 因此本文也涉及事件检测领域. 事件检测, 最早要追溯到美国国防部发起的 TDT (Topic detection and tracking) 项目[16]. TDT 项目中事件检测分为回溯事件检测与新事件检测[17], 主要处理文本和音频等传统新闻媒体. 最初使用的方法以文本聚类算法居多, 后来, 随着以 LDA[18] 为代表的主题模型的提出, 基于贝叶斯概率推断的话题发现算法成为事件检测领域研究的主流. 而随着 Facebook、Twitter 和微博等新型社交媒体的兴起, 以社交网络为研究对象的事件检测成为人们关注的热点. 文献 [19] 将突发性特征引入到传统的向量空间模型中, 使文本表示既包含语义信息又包含时间信息, 从而更好地进行事件检测, 但该模型仅在新闻

文本中进行了实验, 应用到社交媒体的效果未知. 文献 [20] 提出应用于 Twitter 类短文本的话题发现算法, 并利用该算法对传统新闻媒体与社交媒体进行话题分析, 比较二者之间的异同点, 但仅考虑文本内容分析, 未涉及时间信息与事件突发性问题. 文献 [21] 将事件发生时的突发性特点融入一个变形的概率图模型中, 实现对突发事件的发现, 侧重于对所发现事件的语义描述. 文献 [22] 考虑社交网络中提供的地理标注服务, 借助统计主题建模与稀疏编码技术, 构建带位置信息的话题发现模型, 探索事件、话题的发生与地理位置的关系. 文献 [23] 利用信号处理中的小波分析方法筛选词语, 再应用基于模块度的图切割方法聚类词语, 用于发现社交网络中的事件. 文献 [24] 提出了一种新的数据结构, 处理不断到来的在线式数据, 并成功应用于 Twitter 趋势发现及总结中; 作者综合数量与内容变化信息, 构建话题切换的检测模型, 用来跟踪话题的演化情况, 此处的话题切换仅关注话题发生变化的起始时间节点, 并未探讨如何确定话题的结束时间节点. 文献 [25−26] 均采用监督分类模型区分事件信息与非事件信息, 从而发现目标事件, 但此类方法需要人工创建训练数据集, 这在一定程度上限制其应用领域的扩展. 文献 [27] 使用文本挖掘及网络分析技术, 挖掘事件发生时的重点要素 (例如时间和地点等), 为舆情监控提供指导. 文献 [28] 基于在线 LDA 模型分析各时间片内子话题的关联, 定义话题的产生、消亡、继承、分裂、合并等演化类型, 构建了话题的内容与强度演化模型.

本文工作也属于社交网络挖掘范畴. 在该方向, 除了进行事件检测的研究之外, 科研人员也开展了其他各式各样的挖掘工作. 文献 [29] 分析事件中公众的情感走向; 文献 [30] 利用社交网络的情感分析预测股市走势. 文献 [11] 探讨各类话题发现算法的优劣, 并分析数据预处理等阶段对话题发现最终结果的影响. 文献 [31] 关注社交网络中影响力分析领域, 详细介绍各种影响力度量方法, 以及影响力分析在意见领袖和影响力最大化问题中的应用. 文献 [32] 提出一种新型的社交网络节点表示形式, 可以有效提高各类社交网络挖掘任务的效果. 文献 [33] 借助 LDA 模型构建语义社会网络, 使用标签传播算法进行社区发现, 较好地解决了语义重叠社区的发现问题. 文献 [34] 提出半监督算法, 融合先验信息, 解决数据缺失与噪音环境中的社区发现问题. 文献 [35] 利用基于线性回归的混合算法分析内容在社交网络中的传播过程.

5 结论

通过对相关方法的分析与实验可得, 单纯依靠社交行为特征, 不足以区别事件交错与噪音对于事件突发性检测带来的干扰, 会引起隐式事件突发性的漏检问题, 因此在多次实验的基础上, 本文引入文本型 (关键词) 信息, 提出了一个综合方案, 将每个时间窗口与不同的关键词集合绑定, 间接将事件映射到不同的关键词特征空间, 从而避免事件交错及噪音的影响, 在得到由关键词特征确定的突发区间后, 将其与由社交行为特征得到的突发区间关联, 得到最终的事件突发性. 在真实数据集上的实验结果表明, 加入关键词信息的事件突发性检测算法能有效改善相关算法的性能, 提升事件突发性检测任务的效果, 验证了该算法的有效性.

最后, 指出一些当前工作有待改进与提高之处, 供各位读者参考.

1) 在进行突发区间计算时, 本文借用了前人提出的算法, 但算法并不完全适合, 会出现区间碎片化问题. 在多目标序列建模时, 其他作者仅假设所选目标之间具有相关性, 对于语义变化巨大的词语, 并不适用, 因此, 在进行多词语序列突发性关联时, 可以尝试构建考虑词语语义关系的突发性检测模型.

2) 在方法设计部分, 本文探讨了多事件紧邻带来的检测困难, 而对于可能的重叠事件突发性, 现有算法仅视作一次突发性, 无法区别不同事件以及分析事件之间的相互影响, 因此有必要进行语义分析, 构建统一内容特征与非内容特征的事件模型, 以便开展事件检测与跟踪工作.

3) 本文仅利用行为与文本特征发现事件相关突发性, 但对于行为、文本与事件关系的研究不够充分, 需要细化, 例如事件与行为的关系, 事件发生时的群体行为反应, 行为与文本关联策略的选择等问题均值得进一步研究.

References

1 Zhao W X, Shu B H, Jiang J, Song Y, Yan H F, Li X M. Identifying event-related bursts via social media activities. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA, USA: ACL, 2012. 1466−1477

2 Kleinberg J. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 2003, **7**(4): 373−397

3 Swan R, Allan J. Extracting significant time varying features from text. In: Proceedings of the 8th International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 1999. 38−45

4 Swan R, Allan J. Automatic generation of overview timelines. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, 2000. 49−56

5 Mei Q Z, Zhai C X. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York, NY, USA: ACM, 2005. 198−207

6 Marcus A, Bernstein M S, Badar O, Karger D R, Madden S, Miller R C. Twitinfo: aggregating and visualizing microblogs for event exploration. In: Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2011. 227−236

7 Takahashi T, Tomioka R, Yamanishi K. Discovering emerging topics in social streams via link-anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 120−130

8 Zhang Lu-Min, Jia Yan, Zhou Bin, Zhao Jin-Hui, Hong Feng. Online bursty events detection based on emoticons. *Chinese Journal of Computers*, 2013, **36**(8): 1659−1667
(张鲁民, 贾焰, 周斌, 赵金辉, 洪锋. 一种基于情感符号的在线突发事件检测方法. 计算机学报, 2013, **36**(8): 1659−1667)

9 Chen F, Neill D B. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2014. 1166−1175

10 Zhang X M, Li Z J, Chao W H, Xia J L. Popularity prediction of burst event in microblogging. In: Proceedings of the 15th International Conference on Web-Age Information Management. Macau, China: Springer, 2014. 484−487

11 Aiello L M, Petkos G, Martin C, Corney D, Papadopoulos S, Skraba R, Goker A, Kompatsiaris I, Jaimes A. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 2013, **15**(6): 1268−1282

12 Feng Chong, Shi Ge, Guo Yu-Hang, Gong Jing, Huang He-Yan. An entity linking method for microblog based on semantic categorization by word embeddings. *Acta Automatica Sinica*, 2016, **42**(6): 915−922
(冯冲, 石戈, 郭宇航, 龚静, 黄河燕. 基于词向量语义分类的微博实体链接方法. 自动化学报, 2016, **42**(6): 915−922)

13 Fung G P C, Yu J X, Yu P S, Lu H J. Parameter free bursty events detection in text streams. In: Proceedings of the 31st International Conference on Very Large Data Bases. New York, NY, USA: ACM, 2005. 181−192

14 Urabe Y, Yamanishi K, Tomioka R, Iwai H. Real-time change-point detection using sequentially discounting normalized maximum likelihood coding. In: Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg, Germany: Springer-Verlag, 2011. 185−197

15 Mathioudakis M, Koudas N. TwitterMonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM, 2010. 1155−1158

16 Allan J, Carbonell J G, Doddington G, Yamron J, Yang Y M. Topic detection and tracking pilot study final report. In: Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop. Lansdowne, Virginia, USA: DARPA, 1998. 194−218

17 Atefeh F, Khreich W. A survey of techniques for event detection in twitter. *Computational Intelligence*, 2015, **31**(1): 132−164

18 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

19 Zhao W X, Chen R S, Fan K, Yan H F, Li X M. A novel burst-based text representation model for scalable event detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2012, **2**: 43−47

20 Zhao W X, Jiang J, Weng J S, He J, Lim E P, Yan H F, Li X M. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. Berlin, Heidelberg, Germany: Springer-Verlag, 2011. 338−349

21 Diao Q M, Jiang J, Zhu F D, Lim E P. Finding bursty topics from microblogs. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2012, **1**: 536−544

22 Hong L J, Ahmed A, Gurumurthy S, Smola A J, Tsioutsiouliklis K. Discovering geographical topics in the twitter stream. In: Proceedings of the 21st International Conference on World Wide Web. New York, NY, USA: ACM, 2012. 769−778

23 Weng J S, Lee B S. Event detection in twitter. In: Proceedings of the 2011 International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI, 2011. 401−408

24 Wang Z H, Shou L D, Chen K, Chen G, Mehrotra S. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 2015, **27**(5): 1301−1315

25 Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. New York, NY, USA: ACM, 2010. 851−860

26 Becker H, Naaman M, Gravano L. Beyond trending topics: real-world event identification on twitter. In: Proceedings of the 2011 International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI, 2011. 438−441

27 Fu Ju-Lei, Liu Wen-Li, Zheng Xiao-Long, Fan Ying, Wang Shou-Yang. Analyzing the characteristics of "east Turkistan" activities using text mining and network analysis. *Acta Automatica Sinica*, 2014, **40**(11): 2456−2468
(付举磊, 刘文礼, 郑晓龙, 樊瑛, 汪寿阳. 基于文本挖掘和网络分析的"东突"活动主要特征研究. 自动化学报, 2014, **40**(11): 2456−2468)

28 Hu Yan-Li, Bai Liang, Zhang Wei-Ming. Modeling and analyzing topic evolution. *Acta Automatica Sinica*, 2012, **38**(10): 1690−1697
(胡艳丽, 白亮, 张维明. 一种话题演化建模与分析方法. 自动化学报, 2012, **38**(10): 1690−1697)

29 Thelwall M, Buckley K, Paltoglou G. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 2011, **62**(2): 406−418

30 Bollen J, Mao H N, Zeng X J. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011, **2**(1): 1−8

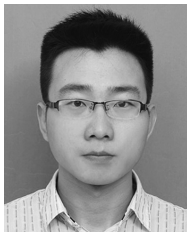
31 Wu Xin-Dong, Li Yi, Li Lei. Influence analysis of online social networks. *Chinese Journal of Computers*, 2014, **37**(4): 735−752
(吴信东, 李毅, 李磊. 在线社交网络影响力分析. 计算机学报, 2014, **37**(4): 735−752)

32 Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2014. 701−710

33 Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping semantic community structure detecting algorithm by label propagation. *Acta Automatica Sinica*, 2014, **40**(10): 2262−2275
(辛宇, 杨静, 谢志强. 基于标签传播的语义重叠社区发现算法. 自动化学报, 2014, **40**(10): 2262−2275)

34 Huang Li-Wei, Li Cai-Ping, Zhang Hai-Su, Liu Yu-Chao, Li De-Yi, Liu Yan-Bo. A semi-supervised community detection method based on factor graph model. *Acta Automatica Sinica*, 2016, **42**(10): 1520−1531
(黄立威, 李彩萍, 张海粟, 刘玉超, 李德毅, 刘艳博. 一种基于因子图模型的半监督社区发现方法. 自动化学报, 2016, **42**(10): 1520−1531)

35 Tsur O, Rappoport A. What′s in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, Washington, USA: ACM, 2012. 643−652



介 飞 合肥工业大学计算机与信息学院博士研究生. 2014 年获得合肥工业大学工学学士学位. 主要研究方向为数据挖掘与社交媒体分析.
E-mail: hfut_jf@163.com
(**JIE Fei** Ph. D. candidate at the School of Computer Science and Information Engineering, Hefei University of Technology. He received his bachelor degree from Hefei University of Technology in 2014. His research interest covers data mining and social media analytics.)

谢 飞 合肥师范学院计算机科学与技术系副教授. 2007 年和 2011 年获得合肥工业大学硕士和博士学位. 主要研究方向为数据挖掘与自然语言处理.
E-mail: xiefei9815057@sina.com
(**XIE Fei** Associate professor in the Department of Computer Science and Technology, Hefei Normal University. He received his master and Ph. D. degrees from Hefei University of Technology in 2007 and 2011, respectively. His research interest covers data mining and natural language processing.)

李 磊 合肥工业大学计算机与信息学院副研究员. 2012 年获得澳大利亚麦考瑞大学计算专业博士学位. 主要研究方向为数据挖掘, 社会计算, 图计算.
E-mail: lilei@hfut.edu.cn
(**LI Lei** Associate professor in the Department of Computer Science and Information Engineering, Hefei University of Technology. He received his Ph. D. degree in computing from Macquarie University, Australia in 2012. His research interest covers data mining, social computing, and graph computing.)

吴信东 长江学者, IEEE Fellow, AAAS Fellow, 合肥工业大学计算机与信息学院教授, 美国路易斯安那大学拉菲特分校计算与信息学院教授. 1993 年获得英国爱丁堡大学人工智能博士学位. 主要研究方向为数据挖掘, 知识库系统, 万维网信息探索. 本文通信作者.
E-mail: xwu@hfut.edu.cn
(**WU Xin-Dong** The Yangtze River Scholar, IEEE Fellow, AAAS Fellow, professor at the School of Computer Science and Information Engineering, Hefei University of Technology, professor at the School of Computing and Informatics, University of Louisiana at Lafayette, USA. He received his Ph. D. degree from the University of Edinburgh, UK in 1993. His research interest covers data mining, knowledge based systems, and Web information exploration. Corresponding author of this paper.)