基于多种词特征的微博突发事件检测方法

张仰森[1,3],段宇翔[1],王 建[1],吴云芳[2]
(1. 北京信息科技大学智能信息处理研究所,北京 100101;2. 北京大学计算语言学研究所,北京 100871;3. 国家经济安全预警工程北京实验室,北京 100044)

摘 要: 近年来,各领域内频频发生各类突发事件,对社会稳定发展产生了一定程度的影响. 本文提出了一种基于多种词特征的微博突发事件检测模型,可以在海量微博数据中对突发事件进行检测,便于相关决策者进行微博监控和舆论引导,尽可能减少突发事件给社会带来的危害. 首先根据时间信息对微博数据进行时间切片,对每一个时间窗口内的数据分别计算各个词语的词频特征、话题标签特征和词频增长率特征;然后基于D-S证据理论和层次分析法,确定词的各个特征权重,并进行加权融合得到词的突发特征值,将突发特征值大的词挑选出来构成突发特征词集,构建基于共现度和结合紧密度的突发事件特征词集的耦合度矩阵;最后将该耦合度矩阵作为凝聚式层次聚类算法的输入,生成一棵由突发词为叶子节点的二叉树,并采用内部相似度的二叉树剪枝算法对聚类结果进行划分,即可实现对相应时间窗口突发事件的检测. 实验结果表明,基于突发词的事件检测模型在簇内部相似度阈值等于1.1时效果最好,正确率达到0.8462、召回率达到0.8684、F值为0.8571,表明了本文所提方法的有效性.

关键词: 微博; 突发事件; 突发特征词; D-S证据理论; 凝聚式层次聚类
中图分类号: TP393.2 **文献标识码:** A **文章编号:** 0372-2112 (2019)09-1919-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.09.015

Microblog Bursty Events Detection Method Based on Multiple Word Features

ZHANG Yang-sen[1,3], DUAN Yu-xiang[1], WANG Jian[1], WU Yun-fang[2]
(1. *Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100101, *China*;
2. *Institute of Computational Linguistics, Peking Universit, , Beijing*, 100871, *China*;
3. *Beijing Laboratory of National Economic Security Early-warning Engineering, Beijing* 100044, *China*)

Abstract: In recent years, a wide variety of bursty events have been occurring frequently in many fields, impacting both the stability and the development of our society. This paper proposes an event detection model based on multiple word features, which is intended to detect bursty events in the massive microblog data. The model will assist decision-makers to monitor microblogs and guide public opinions and will minimize the negative effect of bursty events to society. Firstly, the model slices the microblog data according to the time information. In each time window, the word frequency feature, the topic tag feature and the word frequency growth rate feature of each word are calculated separately. Then, the D-S evidence theory and the analytic hierarchy process are utilized to determine each word's feature weights, which are then merged to obtain the bursty feature value of the word. Words with large bursty feature value are selected to form the bursty feature word set and to construct a coupling degree matrix of bursty feature word set based on co-occurrence degree and tightness. Finally, the coupling degree matrix is used as the input of the hierarchical agglomerative clustering algorithm to generate a binary tree with bursty words being leaf nodes, and the internal similarity binary tree pruning algorithm is used to divide the clustering results. In this way, the detection of the corresponding time window's bursty events can be realized. The experimental results show that the event detection model based on bursty words has the best effect when the intra-cluster similarity threshold is 1.1, the correct rate is as high as 0.8462, the recall rate reaches 0.8684, and the F value is 0.8571, indicating the effective-

收稿日期:2018-08-13;修回日期:2018-11-20;责任编辑:梅志强
基金项目:国家自然科学基金(No.61772081);科技创新服务能力建设-科研基地建设-北京实验室-国家经济安全预警工程北京实验室项目(No.PXM2018_014224_000010)

ness of the proposed method.

Key words: microblog; bursty events; bursty feature words; D-S evidence theory; hierarchical agglomerative clustering

1 引言

近年来,各类突发事件频发,例如2008年南方大面积雪灾、2009年乌鲁木齐打砸抢烧事件、2010年玉树地震、2014年MH370航班失联、2015年英国脱欧事件、2017年红黄蓝幼儿园事件,无一不让人为之震撼.这一系列突发事件的发生对受难地和受难人造成经济和生命的双重损失,产生难以估量的影响,因此在各类突发事件后采取相应的措施进行及时的网络监测和舆论引导有着重要研究意义.以报纸和电视新闻为传播途径的传统媒体不能为用户带来及时的信息,同时,用户与用户之间无法实时互动参与,使得事件的传播速度和广度受到极大的限制.而随着互联网媒体的出现和快速发展,类似微博这种以用户和用户之间的交互关系为核心的分布式社交媒体使得信息的传播变得更加快速和广泛,同时互动性也大大增强.

微博用户在突发事件发生且产生相关的微博后进行转发和评论,而其微博好友可以进行信息的二次加工并继续传播,最终呈现出"一传十、十传百"的指数级别增长.如何在海量的微博动态数据中及时准确的检测出突发事件逐渐成为研究热点,这也是话题检测与追踪技术(Topic Detection and Tracking,TDT)中的重要分支,如果可以及时地发现突发性社会事件并进行合理的控制,将会对社会的稳定和公众的利益产生重要的现实意义.

2 国内外研究现状

近年来,国内外相关学者在网络社交媒体的突发事件检测领域已经投入了大量的研究,并取得了阶段性的成果.目前的核心问题与难点是如何从指数级增长的数据中迅速并准确地检测出突发事件.现有的突发事件检测方法主要可以归为"以文本为中心"[1~3]、"以突发特征词为中心"[4~8]和"以局部地域标签特征为中心"[9~15]三类.

以文本为中心的突发事件检测方法是一种基于文本语义之间的距离进行文本聚类的方法.该方法首先对时间进行切片,然后根据文本的发布时间将其划分到对应的时间窗口内,再对每个时间切片内的微博文本进行聚类,并在得到的每一个簇中抽取突发特征,对满足相应突发规则的类进行突发事件的识别.但由于微博文本中包含大量口语化词语、网络用语、广告、链接等垃圾信息,因此,在聚类并提取突发特征时会引入很多噪声信息.此外,在进行微博文本聚类时会涉及到很多参数阈值的选取,并且参数阈值的选取大多是根据学者的经验来设定,而选取的好坏会对聚类的结果产生较大的影响,从而影响检测的准确率.

以突发特征词为中心的突发事件检测方法是从微博文本中抽取具有突发性的特征词,针对得到的突发特征词进行聚类,从而实现了微博突发事件检测.该方法的核心工作在于突发词的特征选取而不是特征词的聚类,因此参数阈值设定并不会大幅度的影响以突发特征词为中心的检测方法的检测效果.该方法避免了以文本为中心的检测方法中参数阈值设定的问题,但是微博文本中存在大量与事件无关的文档,因此去除噪声并准确提取突发词是提高检测率的重要因素.

以局部地域标签特征为中心的事件检测方法主要针对含有地域信息的微博数据,包括用户信息中附带的地理标签、微博内容中所包含的地点等.该方法可以检测出在全网微博文本中并不突出,但是发生在某一局部地域的热点突发事件.其核心问题主要集中在两个方面,其一是如何在微博文本中抽取地域性的突发特征词,其二在于如何在小区域范围内计算微博的热度.

微博突发事件检测的核心问题包括数据噪声、参数阈值的设定和突发性特征的抽取.本文基于突发词对微博的突发事件进行检测,争取从算法准确性以及执行效率上有所突破,其流程图如图1所示.

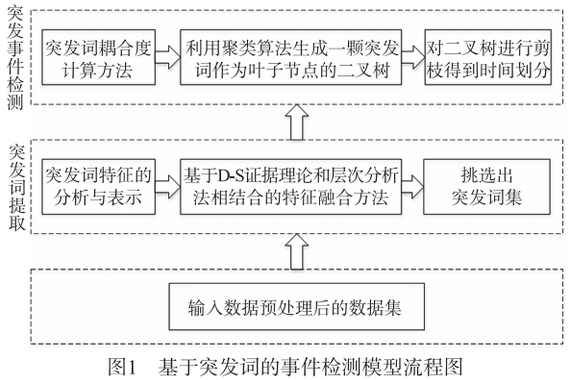

图1 基于突发词的事件检测模型流程图

3 基于多突发特征的突发词提取模型

在进行突发词提取之前,首先进行一些形式化定义,以便简化后续公式的描述.首先对所有微博数据按照时间进行划分,T_1 表示第1个时间窗口,则 T_n 表示第 n 个时间窗口.由此构成时间窗口序列 $T_1, T_2, \cdots, T_{n-1}, T_n$,且本模型中时间窗口的选取大小为一天.时间窗口

T_k 中包含微博数据集 D_k,因此时间窗口序列构成的微博数据集序列为 $D_1, D_2, \cdots, D_{n-1}, D_n$. 微博数据集 D_k 包含已经预处理好的微博数据 $D_k = \{d_1, \cdots, d_i, \cdots, d_{n-1}, d_n\}$,其中 d_i 表示预处理后的一条微博数据,由于预处理已经包含了中文分词的处理工作,因此 d_i 又可表示为 $d_i = \{w_1, w_2, \cdots, w_j, \cdots, w_{n-1}, w_n\}$,其中 w_j 为微博数据中的第 j 个词.

3.1　突发词特征的分析与表示

微博文本中突发事件的出现往往会产生一些相应的特性,如词频、词频的增长率以及形成的话题等. 例如,某些事件在某个时间窗口 T_i 的影响力很小,但是在时间窗口 T_{i+1}内受到广泛的关注,影响力突增,与此同时,与事件相关的词语、话题也会出现激增. 因此本文从词频、话题标签和词频增长率等多种特征来获取突发特征词.

(1)词频特征

词频特征是最能够直观反映一个词在整个时间窗口的数据集内的重要程度. 若某一个词频繁地在某个时间窗内出现,就意味着该时间窗口内可能出现了与该词相关的突发事件. 因此,将词频作为突发词的特征之一. 对于词频特征的计算,通常采用经典的 TD-IDF 方法,该方法能够找出文档集中具有高区分度的高频词语,并为其赋予的一定的权重. 然而,在面向微博数据的突发事件检测时,由于文档集数量大且文档长度很短,若直接采用 TF-IDF 方法,会把一些在大量微博中多次出现而没有区分度的词语赋予较低的权值,从而无法检测到部分突发词. 因此,本文提出了一种改进的 TF-IDF 计算方法. 在时间窗口 T_n 所有微博数据集中,某个词 w 的词频权重 $C_n(w)$ 计算方法如式(1)所示.

$$C_n(w) = \alpha + (1-\alpha) \times \frac{f_n(w)}{f_n^{\max}} \tag{1}$$

其中,$C_n(w)$ 表示词 w 在时间窗 T_n 下的词频权重,$f_n(w)$表示词 w 在时间窗 T_n 下的词频,$f_n^{\max}$ 表示在时间窗 T_n 中的词的最大频率,α 为词频权重的初始值. 这种词频权重的计算方法在进行突发词抽取时避免了传统 TF-IDF 方法对微博这种长短不一且均为短文本的数据处理时造成的干扰,更适合于基于微博的突发事件检测.

(2)话题标签特征

话题标签是新浪微博的核心功能之一,能够让用户自行选择其所发布文本的主题,即高度概括用户文本内容的短语. 因此,与突发事件相关的突发词就很可能会出现在相应的微博话题标签中. 本文将话题标签作为突发词抽取时的特征之一. 在时间窗口 T_n 所有微博数据集中某词 w 的话题标签权重 $T_n(w)$ 计算方法如式(2)、式(3)所示.

$$T_n(w) = \frac{N_{\text{tag}}(w)}{N_{\text{tag}}} \times if(w) + \frac{N_{\text{t\_blog}}(w)}{N_{\text{t\_blog}}} \times f(w) \tag{2}$$

$$if(w) = \begin{cases} 1, \text{至少有一个话题包含词 } w \\ 2, \text{没有任何的话题包含词 } w \end{cases} \tag{3}$$

其中,$N_{\text{tag}}(w)$表示出现词 w 的话题标签个数,N_{tag}表示总的话题标签个数. 同样地,$N_{\text{t_blog}}(w)$表示词 w 在包含话题标签的微博中出现的次数,$N_{\text{t_blog}}$表示包含话题标签的微博个数. $if(w)$是判断因子,用于判定话题标签中是否包含词 w. 这种话题标签权重计算方法,考虑了词语所处的位置,对于处于话题中的词语或处于带有话题的微博中的词语赋予较高的权重.

(3)词频增长率特征

词频特征考虑到一个时间窗口内的高频词语,但是没有考虑到词频的变化趋势. 若某一突发事件刚刚发生,其突发词仅在 T_i 时间窗内剧增,就无法通过词频权重进行突发词的提取,因此引入词频增长率特征来识别突发词很有必要. 本文结合历史数据,首先计算某词 w 在时间窗 T_n 以及其以前的历史平均词频 $A_n(w)$,计算方法如式(4)所示.

$$A_n(w) = A_{n-1}(w) + \frac{f_n(w) - A_{n-1}(w)}{n} \tag{4}$$

其中,$f_n(w)$表示词 w 在时间窗 T_n 下的词频. 利用上式计算连续多个时间窗口内的平均词频,用于反应某词的词频所产生的动态变化. 而词频增长权重可以根据历史平均词频和当前时间窗口词频来计算,表示某一词语当前处于爆发、平稳还是骤减状态,$B_n(w)$表示某词 w 在时间窗 T_n 的词频增长权重,计算方法如式(5)所示.

$$B_n(w) = \frac{f_n(w) - A_n(w)}{A_n(w)} \tag{5}$$

其中,$f_n(w)$表示某词 w 在时间窗 T_n 的词频,词频增长权重 $B_n(w)$反映了词相较于历史情况的活跃程度. 若 $B_n(w)$大于 0 表示该词处于增长阶段,值越大说明越有可能属于突发词. 反之如果小于 0 表示该词属于衰减阶段,基本不可能属于突发词.

词频特征能够使我们在时间窗口内挑选出频率高的词,话题标签特征能够使我们挑选出时间窗口内具有代表性的词语,而词频增长率特征能够在时间的推移过程中快速的发现与突发事件相关的词语. 因此,突发词判断权重 $W_n(w)$将词语 w 的词频特征 $C_n(w)$、话题标签特征 $T_n(w)$和词频增长率特征 $B_n(w)$进行加权得到,计算方法如式(6)所示.

$$W_n(w) = \alpha C_n(w) + \beta T_n(w) + \gamma \frac{B_n(w) - B_n\min}{B_n\max - B_n\min} \tag{6}$$

其中,$W_n(w)$表示某词 w 在时间窗 T_n 的突发词权重,将加权结果大于某一阈值并取TopN作为该时间窗口内

的突发词. 由于词频增长权重 $B_n(w)$ 的取值范围是 $(-\infty,+\infty)$,因此在加权之前需要对其进行归一化,$B_n\max$ 和 $B_n\min$ 分别表示在时间窗口 T_n 内词频增长权重的最大值和最小值. α、β、γ 分别表示词频特征、话题标签特征和词频增长率特征的权值,且要求 $\alpha+\beta+\gamma=1$. α、β 和 γ 的取值直接影响突发词的抽取效果,因此下面将讨论权重 α、β 和 γ 的计算方法.

3.2 基于 D-S 证据理论和层次分析法相结合的特征融合方法

为了更好的抽取微博数据中的突发词,本文将词频特征、话题标签特征和词频增长率特征进行融合,得到突发词判断权重来抽取突发词,并作为下一步突发事件检测的输入. 微博突发事件具有突发性、不确定性,是处于一种"未知的"状态. 由于 D-S 证据理论是一种不确定性的推理方法,能够处理由未知性引起的不确定性,同时,层次分析法能够将定性问题转化为定量计算,并且能够对最终的定量计算结果进行一致性检验. 因此本文利用 D-S 证据理论对专家确定的初始权重矩阵进行推理,采用层次分析法构建各个特征的判断矩阵,并使用层次分析法中的特征矩阵一致性检验方法对上面得到的判断矩阵进行一致性检验,验证整个特征矩阵不确定性推理过程的有效性,以此得到相对准确的特征向量,并将该特征向量作为各个特征的权重. 特征融合过程如图 2 所示.

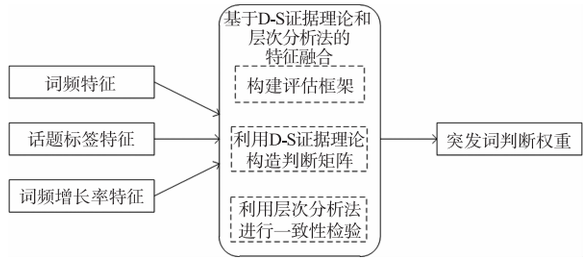

图2 特征融合权重计算方法流程图

(1)构建评估框架

评估框架是所解决问题的状态空间集合及其推理的证据体系,本文需要解决的问题是判断由微博数据中提取出的词是否为突发词,其状态主要分为突发词和非突发词两种. 因此,本文定义异常评估框架 $\Theta=\{Y,N\}$,其 Y(Yes)状态表示该词为突发词,N(No)状态表示该词不是突发词,则有 $Y\cap N=\varnothing$. 在突发词判定模型中主要考虑了词频特征、标签话题特征以及词频增长率特征,因此构建证据三元组 E(C,T,B),则词 w 在时间窗 T_n 下的三元组的取值为

$$E_x\left(C_n(w),T_n(w),\frac{B_n(w)-B_n\min}{B_n\max-B_n\min}\right)$$

(2)利用 D-S 证据理论构造判断矩阵

在 D-S 证据理论中,基于证据理论的不确定性推理出信任分配函数是最关键的一步. 由于突发词的抽取很难有标准的数据集去界定,因此结合多位专家意见,构造判断矩阵,最大程度减少个人对整个评估结果的影响,使结果更客观.

首先定义一个 3×3 的判断矩阵 $\boldsymbol{P}$ 来表示突发词判定用到的三个证据之间的关系,矩阵中的值 w_{ij} 代表第 i 个证据和第 j 个证据相对于突发词判定时重要程度的比较结果,w_{ij} 越大,代表证据 i 比证据 j 更重要. 首先需要 N 位专家给出第 i 个证据相比于第 j 个证据重要的概率 $m_1(A),m_2(A),\cdots,m_n(A)$,然后依据 D-S 证据理论合成规则,计算 N 位专家合成后的值 $m(A)$ 即 w_{ij},计算方法如式(7)、式(8)所示.

$$W_{ij}=m(A)=(m_1\oplus m_2\cdots\oplus m_n)(A)$$
$$=\frac{1}{K}\sum_{A_1\cap A_2\cdots A_n=A}m_1(A_1)\cdot m_2(A_2)\cdots m_n(A_n) \tag{7}$$
$$K=\sum_{A_1\cap A_2\cdots A_n\neq\varnothing}m_1(A_1)\cdot m_2(A_2)\cdots m_n(A_n) \tag{8}$$

(3)利用层次分析法对判断矩阵进行一致性检验

由上一步基于 D-S 证据理论合成来的判断矩阵 $\boldsymbol{P}$ 如式(9)所示.

$$\boldsymbol{P}=\begin{bmatrix}1 & w_{12} & w_{13}\\ w_{21} & 1 & w_{23}\\ w_{31} & w_{32} & 1\end{bmatrix} \tag{9}$$

为了把判断矩阵作为层次分析法的输入,将矩阵 $\boldsymbol{P}$ 中对角线以上的数据进行转换,将上三角部分转换为下三角部分的倒数,即:若 $i<j$ 则 $w_{ij}=\frac{1}{w_{ji}}$. 一般来说,若判断矩阵 $\boldsymbol{P}$ 中的元素满足 $a_{ik}+a_{kj}=a_{ij}$,则称矩阵 $\boldsymbol{P}$ 为一致矩阵. 如果判断为非一致矩阵,则使用最大特征根方法计算不一致程度指标 CI,如式(10)所示.

$$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1} \tag{10}$$

其中,CI 为不一致程度指标,$\lambda_{\max}$ 为判断矩阵 $\boldsymbol{P}$ 的最大特征根,n 为抉择因素的数量. 结合突发词影响因素,此处 $n=3$. 计算得到 CI 后需要查找一致性指标表(如表 1)得到随机一致性指标 RI=0.58,最后计算相对的一致性指标 CR,计算方法如式(11)所示.

$$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}} \tag{11}$$

一般来说,当相对的一致性指标 CR<0.1 时,表示判断矩阵 $\boldsymbol{P}$ 的不一致程度在允许的范围内,其对应的特征向量 $\{w_1,w_2,\cdots,w_n\}$ 可以作为权重向量.

表 1 随机一致性指标 RI 的取值表

矩阵阶数 n	1	2	**3**	4	5	6	7	8	9	10
RI	0	0	**0.58**	0.90	1.12	1.24	1.32	1.41	1.15	1.19

(4)各指标权重计算

通过上一步一致性检验得到的判断矩阵 $\boldsymbol{P}$,首先对其每一列进行归一化,得到新的判断矩阵 $\boldsymbol{P}'$,元素 w_{ij} 归一化为 W'_{ij} 的计算方法如式(12)所示.

$$W'_{ij} = \frac{w_{ij}}{\sum_{k=2}^{2} w_{kj}} \tag{12}$$

然后对判断矩阵 $\boldsymbol{P}'$ 中的每一行进行求和,得到各个特征权重对突发词抉择影响力大小的表征向量 $\boldsymbol{x}^{\mathrm{T}} = (x_1, x_2, x_3)$,其中 $x_i = \sum_{k=2}^{2} w_{ik}$ 表示某一行元素之和. 最后在对向量 $\boldsymbol{x}^{\mathrm{T}}$ 进行归一化,得到各特征对应的权重 $\boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_1, \alpha_2, \alpha_3)$,其中 $\alpha_i = \frac{x_i}{\sum_{k=0}^{2} x_i}$ 则为各特征对突发词判断的权重. 即式(6)中的 $\alpha = \alpha_1$、$\beta = \alpha_2$、$\gamma = \alpha_3$,根据式(6)即可计算得到某个词 w 的突发性权重来判断该词是否为突发词.

4　基于突发特征词的事件检测模型

首先,定义在时间窗口 T_n 抽取得到的突发词集合为 W_n,再基于 W_n 中的这些突发词对突发事件进行检测. 由于突发事件具有不确定的特点,且突发事件的准确个数也难以确定,因此我们采用机器学习中的聚类算法来构建基于突发特征词的事件检测模型.

在机器学习和数据挖掘领域,最常用的聚类算法有 K-means、层次聚类算法等. 由于 K-means 聚类算法需要预先给定簇的个数,因此,不适于本文这种未知个数的事件检测. 层次聚类算法的主要思想是将待聚类的数据集中的每一个元素均当成一个初始的中心,然后采用合并的算法自下而上地构建树型结构,树结构中的每一个子树均为聚类过程中的一个簇结果,通过设置距离阈值来控制迭代,本文拟采用层次聚类算法构建突发事件检测模型.

4.1　突发词耦合度计算方法

如果某一个突发事件发生,则可能同时伴随着大量微博的产生,且一些词语频繁地出现在这些微博中,则这些微博极有可能描述的是同一个突发事件. 为了防止所提取的突发词之间的语义差异而导致将描述同一突发事件的两个突发词分为两类,影响突发事件检测的准确率,本文引入突发词耦合度的概念. 所谓耦合度是指两个突发词之间共现度和结合紧密度的融合. 共现度表示了两个突发词同时出现在一个微博中的情况,而结合紧密度则反映的是两个词之间的语义相关度. 耦合度将两者结合,考虑到词语虽然同现但语义不相关的情况,将耦合度作为输入提供给层次聚类算法,得到一个以突发词为节点的树形结构,最后对树型结构进行拆分、剪枝,从而实现对突发事件的检测. 突发词耦合度的计算模型通过突发词之间的共现度和互信息融合来实现.

首先给出词语 w_i 相对于词语 w_j 的相对共现度 $R(w_i|w_j)$ 和词语 w_j 相对于词语 w_i 的相对共现度 $R(w_j|w_i)$ 的定义,如式(13)所示.

$$R(w_i|w_j) = \frac{\mathrm{tf}(w_i, w_j)}{\mathrm{tf}(w_j)} \quad R(w_j|w_i) = \frac{\mathrm{tf}(w_j, w_i)}{\mathrm{tf}(w_i)} \tag{13}$$

其中,$\mathrm{tf}(w_i, w_j)$ 和 $\mathrm{tf}(w_j, w_i)$ 表示同时包含词语 w_i 和 w_j 的微博中条数,$\mathrm{tf}(w_i)$ 和 $\mathrm{tf}(w_j)$ 表示包含词语 w_i 和 w_j 的微博条数. 显然 $\mathrm{tf}(w_i, w_j) = \mathrm{tf}(w_j, w_i)$ 恒成立,而大多数情况下 $R(w_i|w_j)$ 是不等于 $R(w_j|w_i)$ 的. 为此,定义了共现度 $C(w_i, w_j)$ 如式(14)所示.

$$C(w_i, w_j) = \frac{1}{2}R(w_i|w_j) + \frac{1}{2}R(w_j|w_i) \tag{14}$$

而两个词同时出现在一条微博中内的结合紧密度则采用互信息的计算方法来进行刻画,如式(15)所示.

$$\mathrm{MI}(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{15}$$

其中,$\mathrm{MI}(w_i, w_j)$ 表示词语 w_i 和词语 w_j 的互信息,$P(w_i)$ 和 $P(w_j)$ 分别表示词语 w_i 和词语 w_j 出现的概率,$P(w_i, w_j)$ 表示词语 w_i 和 w_j 一同出现的概率. 互信息越大表示两个词语的紧密程度越高. 最后融合词语的共现度和结合紧密度构建词语的耦合度计算模型,如式(16)所示.

$$S(w_i, w_j) = C(w_i, w_j) + \mathrm{MI}(w_i, w_j) \tag{16}$$

其中,$S(w_i, w_j)$ 即为词语 w_i 和词语 w_j 之间的耦合度. 由此构建突发词集 W_n 的耦合度矩阵 $\boldsymbol{U}'$ 如式(17)所示.

$$\boldsymbol{U}' = \begin{bmatrix} S_{\max} & S(w_1, w_2) & \cdots & S(w_1, w_{n-1}) & S(w_1, w_n) \\ S(w_2, w_1) & S_{\max} & \cdots & S(w_2, w_{n-1}) & S(w_2, w_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ S(w_{n-1}, w_1) & S(w_{n-1}, w_2) & \cdots & S_{\max} & S(w_{n-1}, w_n) \\ S(w_n, w_1) & S(w_n, w_2) & \cdots & S(w_n, w_{n-1}) & S_{\max} \end{bmatrix} \tag{17}$$

上述矩阵显然具有对称性,对角线元素 $S_{\max}$ 表示突发词之间耦合度最大的值. 将该矩阵中的每一个元素进行归一化操作,得到归一化后的耦合度矩阵 $\boldsymbol{U}$ 如式(18)所示.

$$\boldsymbol{U}=\begin{bmatrix} 1 & \frac{S(w_1,w_2)-S_{\min}}{S_{\max}-S_{\min}} & \cdots & \frac{S(w_1,w_{n-1})-S_{\min}}{S_{\max}-S_{\min}} & \frac{S(w_1,w_n)-S_{\min}}{S_{\max}-S_{\min}} \\ \frac{S(w_2,w_1)-S_{\min}}{S_{\max}-S_{\min}} & 1 & \cdots & \frac{S(w_2,w_{n-1})-S_{\min}}{S_{\max}-S_{\min}} & \frac{S(w_2,w_n)-S_{\min}}{S_{\max}-S_{\min}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{S(w_{n-1},w_1)-S_{\min}}{S_{\max}-S_{\min}} & \frac{S(w_{n-1},w_2)-S_{\min}}{S_{\max}-S_{\min}} & \cdots & 1 & \frac{S(w_{n-1},w_n)-S_{\min}}{S_{\max}-S_{\min}} \\ \frac{S(w_n,w_1)-S_{\min}}{S_{\max}-S_{\min}} & \frac{S(w_n,w_2)-S_{\min}}{S_{\max}-S_{\min}} & \cdots & \frac{S(w_n,w_{n-1})-S_{\min}}{S_{\max}-S_{\min}} & 1 \end{bmatrix} \tag{18}$$

将该矩阵 $\boldsymbol{U}$ 作为输入,进行层次聚类的过程会在下一节详细论述.

4.2 基于凝聚式层次聚类的事件检测方法

层次聚类算法的策略一般有自底向上的和自顶向下两种. 自底向上的聚类方法将每个元素初始都看作一个簇中心,然后计算任意两个簇之间的相似度并对最相似的两个簇进行融合,重复上述步骤,直到满足迭代次数阈值或合并成了一个簇停止计算,最终得到一个二叉树形结构. 而自顶向下的聚类方法恰恰相反,是将所有元素初始都看成一个簇内的,然后根据元素与元素之间的相似度进行拆分从而构成二叉树形结构.

本文采用自底向上的凝聚式层次聚类,并使用平均距离算法来计算簇与簇之间的距离. 首先确定簇间距离的计算方法,假设 $y_a=\{w_{a1},w_{a2},\cdots,w_{am}\}$ 与簇 $y_b=\{w_{b1},w_{b2},\cdots,w_{bm}\}$ 是由不同个数的突发词构成,簇间平均距离受突发词间的相似度影响,将两簇内两两元素间相似度的倒数平均值作为两簇的平均距离,构建簇与簇的距离计算方 $D(y_a,y_b)$,如式(19)所示.

$$D(y_a,y_b)=\begin{cases} \dfrac{|y_a|\times|y_b|}{\sum\limits_{w_i\in y_a,w_j\in y_b}\dfrac{S(w_i,w_j)-S_{\min}}{S_{\max}-S_{\min}}}, & \sum\limits_{w_i\in y_a,w_j\in y_b}\dfrac{S(w_i,w_j)-S_{\min}}{S_{\max}-S_{\min}}>0 \\ +\infty, & \sum\limits_{w_i\in y_a,w_j\in y_b}\dfrac{S(w_i,w_j)-S_{\min}}{S_{\max}-S_{\min}}\leqslant 0 \end{cases} \tag{19}$$

其中,$D(y_a,y_b)$ 则为两个簇之间的距离,$|y_a|$ 和 $|y_b|$ 分别表示簇 y_a 和簇 y_b 中突发词的个数,w_i 和 w_j 分别为两个簇中的突发词元素. 通过该平均距离公式来更新自底向上的凝聚式层次聚类算法中簇间距离,实现对突发词的聚类. 具体基于平均距离的自底向上凝聚式层次聚类算法如算法 1.

算法 1 自底向上凝聚式层次聚类算法

输入:时间窗 T_n 下的突发词集合 W_n、突发词集合 W_n 构成的耦合度矩阵 $\boldsymbol{U}$

输出:由突发词作为节点的二叉树型结构

步骤 1:将突发词集合 $W_n=\{w_1,w_2,\cdots,w_k\}$ 中的每一个突发词看作一个簇,因此当前簇集合为 $\text{Cluster}_n=\{y_1,y_2,\cdots,y_k\}=\{\{w_1\},\{w_2\},\cdots,\{w_k\}\}$

步骤 2:计算簇集合 Cluster_n 中任意两个簇之间的相似度,暂时缓存于集合 Temp 中,则 $\text{Temp}=\{D(y_1,y_2),D(y_1,y_3),\cdots,D(y_1,y_k),D(y_2,y_3),\cdots,D(y_{k-1},y_k)\}$

步骤 3:判断集合 Temp 的元素个数,若 $|\text{Temp}|=1$,则跳转步骤 5;若 $|\text{Temp}|>1$,则选取集合 Temp 中最小的距离值并获取对应的两个簇. 假设为簇 y_m 和 y_n,将这两个簇从集合 Cluster_n 中删除后,并将这两个簇合并成一个新的簇后加入到集合 Cluster_n,此时簇集合 $\text{Cluster}_n=\{y_1,y_2,\cdots,\{y_m,y_n\},\cdots,y_{k-1}\}$

步骤 4:根据合并后簇集合 Cluster_n 相应的更新 Temp 集合,跳转至步骤 3

步骤 5:输出对应的二叉树型结构.

经过算法 1 之后,可以得到一棵由突发词以及突发词集合构成的二叉树型结构. 二叉树所有叶子节点的集合是初始的突发词集合 W_n,而二叉树中的非叶子节点是突发词集合 W_n 的一个子集. 因此要对突发事件进行识别,则需要对这颗以突发词为叶子节点,突发词构成的簇为非叶子节点的二叉树进行分割,将分割后得到的子树中的突发词作为突发事件的关键词. 聚类算法步骤 3 中,取距离最小的两个簇进行合并的过程中,y_m 与 y_n 间的相似度是作为分割二叉树的标准. 若两个合并的簇内部相似度不够高,则需要对该簇进行分割. 具体基于内部相似度的二叉树剪枝算法如算法 2.

算法 2 二叉树剪枝算法

输入:凝聚式层次聚类输出的二叉树型结构、簇内部相似度阈值 θ

输出:突发事件的划分集合 E

步骤 1:判断根节点的左右子树的内部相似度 D,若 $D\geqslant\theta$ 说明根节点满足簇内相似度要求,则将整颗二叉树加入到集合划分 $E=\{\text{root}\}$ 并跳转到步骤 4,否则分别执行步骤 2 和步骤 3

步骤 2:若根节点的左子树不为空,则将其左子树和簇内部相似度阈值 θ 作为输入,进行基于内部相似度的二叉树剪枝

步骤 3:若根节点的右子树不为空,则将其右子树和簇内部相似度阈值 θ 作为输入,进行基于内部相似度的二叉树剪枝

步骤 4:输出突发事件划分集合 E

在最终得到的剪枝结果集合 E 中,所有的突发词都在其中且会被放在不同的簇内,且每个簇的大小不

一定完全相同. 本文将在下一节中以真实的数据进行实验来验证该模型的有效性.

5　实验结果分析

5.1　突发事件检测结果与分析

在微博突发事件检测领域,由于没有通用的标准数据集,本实验通过新浪 API 爬取从 2018 年 3 月 10 日到 3 月 29 日的微博数据,经相关预处理后保留微博文本 450799 条进行实验. 鉴于微博数据中存在的大量噪声,本文采取以下方法进行解决:(1)将少于 5 个字的微博文本或者文本为重复字母、重复汉字、重复符号的微博进行删除;(2)删除微博的非文本信息,如:URL 链接、表情符号、特殊字符等;(3)构建了一个面向微博文本的用户词表和停用词表,并基于 NLPIR 分词系统对文本进行分词处理和停用词去除. 经处理后的微博数据示例如图 3 所示,其存储结构如表 2 所示[17].

user_id	user_name	content	zan	zhuan	time
6372119568	歪果网红的日常	泪难将你重重摔在地上，不要去伤心，难过，崩溃，哭泣，你应该做的是当场假装去世。 歪果网红的日常的秒拍视频	844	377	03月12日 22:27
6367204564	水分子	富里府（省）的某寺庙里，大批信徒和百姓参加了知名高僧龙婆篇的“换衣仪式”。有缘见到高僧真身，望保佑我们	744	35	03月12日 11:45
6432839794	原来是主播咧	表白的时候！这次被气死的真然是主播？路人都这么皮的吗？结尾真的非常扎心了[允悲]“我喜欢你啊！”“呃”	1247	411	03月12日 13:21
5679434436	六月的雪meimei	！当得知夏杉杉去世！当场懵掉！为自己的无心之错感到深深自责！……童薇因杉杉的离去不肯原谅晓飞！二人不对	454	132	03月12日 22:52
1829578697	圈扒酱	蛮有看头的，尤其最后，简直看到心绞痛！彻底被毛林林的演技折服了～毛林林凭借《谈判官》脱颖而出，本身具有	875	431	03月12日 22:17
3739754765	影视剧透社	，以至于童薇和谢晓飞分手。商碧晨 秦天宇 有情人终成眷属。赵晨曦提出和谢晓飞比赛，如果谢晓飞先爬到山顶	956	461	03月12日 22:12
3942614076	H_桃西_T	，不是一方屡屡包容迁就。童薇把夏杉杉去世这件事也拿来怪谢晓飞，也是奇了怪了。谢晓飞当时面对父亲进医院，	450	231	03月12日 21:44
5592488502	鑫鑫与	感动中国2008年度候选人物吴玉兰是乌鲁木齐市某社区一位74岁的贫苦老人。9年来，她以拾荒、卖废品的方式还清	272	11	03月12日 16:28
6177412609	小仙女看剧	店定好了想给你一个惊喜…你是不是等我等太久…我只是希望我的事业能稳定一点，等你把孩子生下来我们一家	340	51	03月12日 22:59
5622004557	痞湿傅	是父亲生前饲养的狗狗，名叫zozo。他后来才知道，父亲去世后，zozo虽然有人照顾，但它每天都会来这里守着，陪	439	15	03月12日 04:46
2695071834	影视娱乐圈	子韬泪奔的剧情真的太虐了…妈妈去世，童薇离开…说好的甜宠呢谢晓飞？不过韬韬的哭戏还真是挺有感染力的电视	694	499	03月12日 21:15
2054853403	追剧情报社	两条人命？你把话说清楚薇薇：如果我能争取早一个航班回来，杉杉可能就不会死，就是因为你拦着我，你知道吗	259	164	03月12日 23:02
6065783763	经典影歌回忆珺	而她唯一孩子也没有抢救过来。童薇几乎要崩溃，这一集好虐，老齐赶到后看到的只是夏杉杉的遗体，相爱不要错	221	99	03月12日 23:25

图3　微博数据示例

表 2　微博数据存储结构

序号	字段名称	字段类型	描述
1	user_id	long	用户 ID
2	user_name	string	用户昵称
3	content	string	微博内容
4	zan	int	点赞数
5	zhuan	int	转发数
6	time	timestamp	发布时间

在本节中,通过真实的微博数据对突发词提取模型进行验证. 首先针对三个突发特征的权值 α、β 和 γ 进行确定,本文收集了两位专家对三种证据的不确定性并将他们的意见融合构建判断矩阵 $\boldsymbol{P}$,证据不确定性概率如表 3 所示,对其意见进行融合后得到 $w_{13}=0.539$、$w_{21}=0.281$ 和 $w_{23}=0.18$ 并构建判断矩阵 $\boldsymbol{P}$,并将矩阵 $\boldsymbol{P}$ 中对角线以上的数据进行转换得到层次分析法的输入矩阵 $\boldsymbol{P}'$如式(20)所示.

$$\boldsymbol{P}'=\begin{bmatrix}1 & 3.559 & 0.539\\ 0.281 & 1 & 0.18\\ 1.855 & 5.556 & 1\end{bmatrix} \tag{20}$$

表 3　证据不确定性概率表

证据	专家 1	专家 2
词频特征	0.3	0.45
话题标签特征	0.2	0.25
词频征率特征	0.5	0.30

对判断矩阵进行一致性检验,首先计算最大特征值 $\lambda_{\max}=3.003$,然后求得 CI = 0.002,当 $n=3$ 查表 1 得到 RI = 0.58,最后根据式(11)计算得到 CR = 0.003 < 0.1. 因此判断矩阵 $\boldsymbol{P}'$满足一致性检验,说明通过证据理论得到的判断矩阵 $\boldsymbol{P}'$是合理的. 最终对判断矩阵 $\boldsymbol{P}'$进行归一化操作得到的各特征权重向量 $\boldsymbol{x}^{\mathrm{T}}=(\alpha,\beta,\gamma)=(0.328,0.0976,0.5744)$.

为了测试突发词提取模型的效果,从数据库中抽取 2018 年 3 月 10 日到 2018 年 3 月 19 日共计 10 天的数据,按天进行时间窗口划分,计算每个词的词频特征、话题标签特征和词频增长率特征,得到突发词权重. 经式(6)计算发现 3 月 14 号突发词的权重存在突发词权重异常高的词语,如表 4 所示.

表 4　2018 年 3 月 14 日部分突发词权重 TOP 表

词语	词频权重	话题标签权重	词频增长率权重	突发词权重	排名
霍金	1	0.212151232	0.984212312	0.914037512	1
去世	0.265441992	0.190421223	1	0.680050085	2
76 岁	0.717050919	0.123412312	0.521132112	0.546576028	3
物理学家	0.265318703	0.032124221	0.442123123	0.344115381	4
享年	0.250955493	0.203123111	0.321234212	0.286655149	5
宇宙	0.158858341	0.102211321	0.251231231	0.20638858	6
再见	0.14264579	0.058321234	0.012351236	0.059574521	7

通过对上述表的分析可以明显看出词语“霍金”、“去世”、“76 岁”的突发词权重明显高于其他词语. 以“霍金”为例,我们将该词这十天的词频进行了统计,统计结果如图 4 所示,可以很明显的发现在 14 号当天发生了“突增”. 从人工角度来看显然当天发生了突发事件——物理学家史蒂芬·霍金去世.

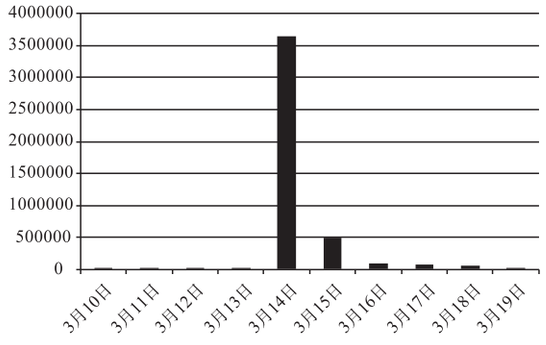

图4 词语“霍金”的频次统计图

接下来是确定基于突发特征词的事件检测模型中的参数,簇内部相似度阈值 θ 是在对突发事件进行划分时的核心参数. 利用从 2018 年 3 月 10 日到 3 月 29 日的数据集抽取到的突发词集合 $W_{3.10}$,$W_{3.11}$,…,$W_{3.29}$共计 20 个. 在中文微博突发事件检测中,国内尚且没有公认的人工标注语料. 因此,结合微博的热搜榜、风云榜和微博数据本身,本文人工标注了这 20 天内的突发事件,包括“物理学家史蒂芬·霍金去世”、“美国宣布对价值约 600 亿美元中国商品征收关税”、“李嘉诚正式宣布退休”、“奔驰车深夜定速巡航失控”、“政协第十三届全国委员会领导人选举产生”、“央视 315 晚会曝光大众途锐设计缺陷”、“普京当选俄罗斯第七届总统”、“Uber 自动驾驶车撞死路人”、“美团打车正式登陆上海”、“多国宣布驱逐俄外交官”、“教育部取消 5 项全国性高考加分”、“金正恩访华”、“高云翔被曝涉嫌性侵在悉尼被捕”等共计 38 个大大小小的突发性事件. 以式(21)~式(23)定义的正确率、召回率和 F 值三个评估指标对突发事件的检测模型进行检验.

$$\text{Precision}=\frac{\text{正确检测到的突发事件个数}}{\text{本模型检测到的突发事件个数}} \tag{21}$$

$$\text{Recall}=\frac{\text{正确检测到的突发事件个数}}{\text{数据集中实际的突发事件个数}} \tag{22}$$

$$F_{\text{measure}}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{23}$$

首先对数据集进行凝聚层次聚类后得到的突发词集,然后通过对簇内部相似度阈值 θ 取 0.5 至 1.5 不等,分别对数据集进行基于内部相似度的二叉树剪枝,得到如表 5 所示的实验结果.

表 5 簇内部相似度阈值 θ 取值的不同对实验结果的影响

阈值 θ	模型检测到的突发事件个数	正确检测到的突发事件个数	Precision	Recall	F_{measure}
0.5	3	1	0.3333	0.0263	0.0487
0.6	5	2	0.4000	0.0526	0.0930
0.7	14	8	0.5714	0.2105	0.3077
0.8	19	13	0.6842	0.3421	0.4561
0.9	30	19	0.6333	0.5000	0.5588
1.0	36	27	0.7500	0.7105	0.7297
1.1	39	33	**0.8462**	0.8684	**0.8571**
1.2	42	34	0.8095	0.8947	0.8499
1.3	46	34	0.7391	0.8947	0.8091
1.4	49	34	0.6939	0.8947	0.7816
1.5	51	35	0.6863	**0.9211**	0.7866

从表 5 中可分析得到,当簇内部的相似度阈值 θ 取 1.1 时,正确率和 F 值分别达到最高值 0.8462 和 0.8571. 而随着簇内部相似度阈值 θ 增大,虽然召回率 Recall 有了小幅度的提升,但是正确率 Precision 和 F 值 F_{measure}均有较大幅度的下降,因此,将簇内部相似度阈值 θ 的取值定为 1.1. 在 $\theta=1.1$ 时对数据集进行突发事件的划分,其中检测正确的突发事件部分结果如表 6 所示.

表 6 部分突发事件划分结果及其对应事件表

时间窗口	突发事件划分 E	与检测结果相关的事件
3 月 14 日	霍金、物理学家、去世、著名、蜡烛、76 岁、语言、事件、英国、逝世、星辰、史蒂芬、爱因斯坦、传奇、生前、悼念、轮椅、简史、宇宙、杰出……	物理学家史蒂芬·霍金去世
3 月 16 日	李嘉诚、退休、超人、商业、首富、正式、宣布、香港、华人、集团、长江、接棒、业绩、人生、时代、财富、帝国、李泽巨、长子、公司、回顾、股东、回应、家庭……	李嘉诚正式宣布退休
3 月 16 日	奔驰、巡航、失控、刹车、小时、高速、车主、收费站、河南、生死、薛先生、无法、后台、交警、车辆、行驶、狂飙、轿车、韩寒、质疑、车门、江苏……	奔驰车深夜定速巡航失控
3 月 22 日	中国、美国、进口、关税、加征、贸易、宣布、备忘录、奉陪、反击、商务部、利益、价值、央视、钢铁、影响、贸易战、跌幅、限制、签署、全球……	美国宣布对价值约 600 亿美元中国商品征收关税
3 月 23 日	俄罗斯、普京、选举、总统、祝贺、身份、目瞪口呆、获胜、选票、投票、宣言、悬念、通电话、支持率、民意、调查、再次、秘书、呼吁、关系、背后……	普京当选俄罗斯第七届总统

表 6 中与检测结果相关的事件名称是人工标注的. 不难发现,通过本文基于突发词的事件检测模型能够较为准确的发现微博中的突发事件,检测结果中也基本包含事件的时间、地点、人物三要素. 以“普京当选俄罗斯第七届总统”为例,事件发生事件为 3 月 23 日,

地点为俄罗斯,人物为普京,还包括“选举”、“总统”、“获胜”、“目瞪口呆”等等该事件的核心关键词.

5.2　正确率、召回率、*F* 值对比

根据文献[16]中提到的增量聚类方法,利用 2018 年 3 月 10 日到 3 月 29 日的微博数据,将得到的突发词耦合度矩阵输入增量聚类算法进行计算,当增量聚类距离阈值选取 300 时 *F* 值最高,达到 0.6315,正确率是 0.5454,召回率是 0.75. 在文献[8]中,采用基于突发词相似度和凝聚式层次聚类的方法实验,当凝聚式层次聚类距离阈值选取 500 时 *F* 值最高,达到 0.7368,正确率是 0.6364,召回率是 0.8750. 与本文方法的正确率、召回率、*F* 值相比较,如图 5 所示.

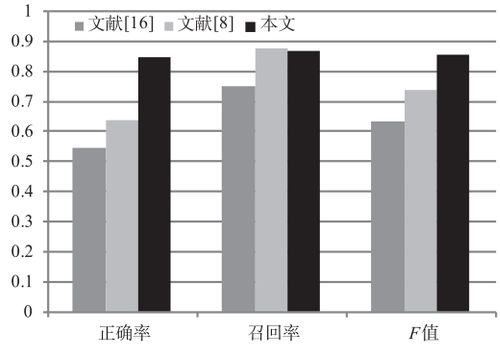

图5　实验结果对比图

从图 5 中可以看出,本文的方法与文献中的两种方法相比,在 *F* 值和正确率上都有很大的提升,分析原因可能有以下两点:

(1)在突发词的提取过程中,本文基于 D-S 证据理论和层次分析法实现了对突发特征词的提取,使得到的突发词能够更准确地对突发事件进行描述,为后面的事件检测提供了良好的数据,所以在准确率和 *F* 值上有更好的表现.

(2)文献[16]中采取的增量聚类方法对突发特征词的输入顺序要求严格,而在实际的实验中,很难找到一个最优的输入顺序,故增量聚类算法在事件检测中的准确性比较低. 本文采用的凝聚式层次聚类的事件检测方法就能有效的避免由输入数据造成的结果误差.

6　总结与展望

本文首先把微博数据集根据时间信息进行切片,对每一个时间窗口内的数据分别计算每一个词的词频特征、话题标签特征和词频增长率特征,然后基于 D-S 证据理论和层次分析法相结合的特征融合方法确定各个特征的权重,根据权重大小挑选出具有突发特征的词集. 接下来基于词语共现度和结合紧密度计算各突发词之间的耦合度,并构建相似度矩阵作为凝聚式层次聚类算法的输入,再采用基于内部相似度的二叉树剪枝算法对聚类结果进行划分,即可得到该时间窗口所对应的突发事件. 实验结果表明,基于突发词的事件检测模型在簇内部相似度阈值取 $\theta = 1.1$ 时,取得正确率 0.8462、召回率 0.8684、*F* 值 0.8571 的良好结果,相比于其他两种方法,在准确率和 *F* 值上有了很大的提升.

对于微博突发事件的检测,本文提出的检测模型还有许多地方需要做出改进:(1)突发事件检测中用到了聚类算法、特征加权融合等需要选择参数的算法及方法,而参数的确定方法是通过真实的数据进行测试得到的最优解. 在实际应用中,数据会随着时间流动态更新,预先设置好的参数不一定适合未来的每一批数据,因此建立参数的自适应模型对未来的系统优化至关重要. (2)在凝聚层次聚类后进行了二叉树剪枝操作,但是该算法计算量较大,算法效率偏低,需要进一步的优化改进. (3)由于数据源、硬件环境等因素影响,本文采用离线批处理计算模式对突发事件进行检测,处理时延在 24 小时之内. 但是突发事件检测的时效性直接决定着其可用性,因此引入流式计算框架以提高突发事件的检测效率是今后努力的方向.

参考文献

[1] Goto J, Miyazaki T, Takei Y, et al. Automatictweet detection based on data specified through news production[A]. Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion [C]. Tokyo: ACM, 2018. 1.

[2] 周刚,邹鸿程,熊小兵,等. MB-SinglePass:基于组合相似度的微博话题检测[J]. 计算机科学,2012,39(10):198-202.
ZHOU Gang, ZOU Hongcheng, XIONG Xiaobing, et al. MB-SinglePass: Microblog topic detection based on combined similarity[J]. Computer Science, 2012, 39(10): 198-202. (in Chinese)

[3] Qiu YF, Cheng L B. Research onsudden topic detection method for microblog[J]. Computer Engineering, 2012, 38(9): 288-290.

[4] Du Y, Wu W, He Y, et al. Microblog bursty feature detection based on dynamics model[A]. International Conference on Systems and Informatics [C]. Bandung: IEEE, 2012. 2304-2308.

[5] Salas A, Georgakis P, Petalas Y. Incident detection using data from social media[A] Proceedings of the 20th International Conference on Intelligent Transportation Systems [C]. Yokohama: IEEE, 2017. 751-755.

[6] Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing[A] Proceedings of the

Fifth International Workshop on Natural Language Processing for Social Media[C]. Boston:Association for Computational Linguistics,2017. 1 – 10.

[7] Kalden J P H. Dataanalysis within the netherlands coastguard:risk mapping, social network analysis and anomaly detection[A]. NL ARMS Netherlands Annual Review of Military Studies 2018[C]. The Hague:TMC Asser Press, 2018. 193 – 200.

[8] 郭跑秀,吕学强,李卓. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用, 2014, 34(02): 486 – 490,505.
GUO Yixiu,LYU Xueqiang,LI Zhuo. Bursty topics detection approach on Chinese microblog based on burst words clustering[J]. Journal of Computer Applications,2014,34(02):486 – 490,505. (in Chinese)

[9] Unankard S,Li X,Sharaf M A. Emerging event detection in social networks with location sensitivity[J]. World Wide Web,2015,18(5):1393 – 1417.

[10] Quezada M, Pe†a-Araya V, Poblete B. Location-aware model for news events in social media[A]Proceedingsof the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York:ACM,2015. 935 – 938.

[11] Krumm J,Horvitz E. Eyewitness:Identifying local events via space-time signals in twitter feeds[A]Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems[C]. Washington:ACM,2015. 20.

[12] Cecaj A,Mamei M. Data fusion for city life event detection[J]. Journal of Ambient Intelligence and Humanized Computing,2017,8(1):117 – 131.

[13] Jinhua L,Zhongjie A. Analyzing geographical coordinates data for micro-blog trending events[J]. Data Analysis and Knowledge Discovery,2016,32(2):90 – 101.

[14] Hua T,Chen F,Zhao L,et al. Automatic targeted-domain spatiotemporal event detection in twitter[J]. GeoInformatica,2016,20(4):765 – 795.

[15] Wakamiya S,Jatowt A,Kawai Y,et al. Analyzingglobal and pairwise collective spatial attention for geo-social event detection in microblogs[A]Proceedings of the 25th International Conference Companion on World Wide Web[C]. Montreal:WWW,2016. 263 – 266.

[16] Zheng F R,Miao D Q,Zhang Z F,et al. News topic detection approach on Chinese microblog[J]. Computer science,2012,39(1):138 – 141.

[17] 张仰森,郑佳,唐安杰. 基于多特征融合的微博用户权威度定量评价方法[J]. 电子学报,2017,45(11):2800 – 2809.
Zhang Yangsen,Zheng Jia,Tang Anjie. A quantitative evaluation method of micro-blog user authority based on multi-feature fusion[J]. Acta Electrinica Sinica,2017,45(11):2800 – 2809. (in Chinese)

作者简介

张仰森 男,1962 年 6 月出生于山西临猗,博士后,教授,研究方向为中文信息处理、人工智能.
E-mail:zhangyangsen@163.com

段宇翔 男,1992 年 3 月出生于山西太原,硕士研究生,研究方向为中文信息处理、突发事件检测.
E-mail:duanyx5173@163.com

王 建 男,1993 年 5 月生于浙江温州,硕士,研究方向为中文信息处理、信息安全.
E-mail:455858538@qq.com

吴云芳 女,1973 年 3 月生于山西,博士,副教授,研究方向为语义计算,智能问答.
E-mail:wuyf@pku.edu.cn