软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006351]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于词相关性特征的社交网络突发事件检测方法*

蒋伟进[1,2], 王扬[2]

[1] (湖南工商大学计算机信息与工程学院,长沙,410205)

[2](湖南工商大学大数据与互联网创新研究院,长沙 410205)

通讯作者: 王扬, E-mail: 987505730@qq.com

摘要: 如何在社交媒体上检测数据流中的突发事件是自然语言处理中的一个热门研究主题,但是,当前用于提取突发事件的方法存在精度低和效率低的问题.为了解决这些问题,本文提出一种基于词相关性特征的突发事件检测方法,能从社会网络数据流中快速地检测出突发事件,以便相关的决策者可以及时有效地采取相关措施进行处理,使突发事件的负面影响能够被尽量降低,维护社会的安定.首先,通过噪声过滤和情绪过滤,我们得到了充满负面情绪的微博文本.然后,根据时间信息,对微博数据进行时间切片,计算每个时间窗口中该数据的每个单词的单词频率特征、用户影响力和单词频率增长率特征,运用突发度计算方法来提取突发词;根据 word2vec 模型合并相似词,利用突发词的特征相似性构成突发词关系图.最后,运用多归属谱聚类算法对单词关系图进行最优划分,并在时间窗滑过时关注异常词语,通过子图中词语突发度的变化而引起的结构变化对突发事件进行判断.由实验结果知,突发事件检测方法在实时博文数据流中具有很好的事件检测效果,与已有的方法相比,本文提出的突发事件检测方法可以满足突发事件检测的需求,不仅能检测到子事件的详细信息,而且事件的相关信息也能被准确地检测出来.

关键词: 突发事件;检测;词相关性特征;单词关系图;多归属谱聚类;

中文引用格式: 蒋伟进,王扬. 基于词相关性特征的社交网络突发事件检测方法.软件学报. http://www.jos.org.cn/1000-9825/6351.htm

英文引用格式: Jiang WJ, Wang Y. Social network emergency detection method based on word correlation characteristics. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/6351.htm

Social network emergency detection method based on word correlation characteristics

JIANG Wei-Jin[1,2], WANG Yang[2]

[1](Institute of Big Data and Internet Innovation, Hunan University of Technology and Business, Changsha 410205, China)

[2](Key Laboratory of Hunan Province for New Retail Virtual Reality Technology, Changsha 410205, China)

Abstract: How to detect sudden events in data streams on social media is a popular research topic in natural language processing. However, current methods for extracting emergencies have problems of low accuracy and low efficiency. In order to solve these problems, this paper proposes an emergency detection method based on the characteristics of word correlation, which can quickly detect emergency events from the social network data stream, so that relevant decision makers can take timely and effective measures to deal with , making the negative impact of emergencies can be reduced as much as possible to maintain social stability. First of all, through noise filtering and emotion filtering, we get microblog texts full of negative emotions. Then, based on the time information, time slice the Weibo data to calculate the word frequency characteristics, user influence and word frequency growth rate characteristics of each word of the data in each time window, and use the burst calculation method to extract the burst word. According to the word2vec model, similar words are merged, and the characteristic similarity of the burst words is used to form a burst word relationship graph. Finally, the multi-attribute

* 基金项目: 国家自然科学基金面上项目(61472136,61772196); 湖南省自然科学基金面上项目(2020JJ4249); 湖南省社会科学基金重点项目(2016ZDB006); 湖南省社会科学成果评审委员会课题重点项目(湘社评 19ZD1005); 湖南省学位与研究生教育改革研究项目(2020JGYB234)

收稿时间: 0000-00-00; 修改时间: 0000-00-00; 采用时间: 0000-00-00; jos 在线出版时间: 0000-00-00

spectral clustering algorithm is used to optimally divide the word relationship graph, and pay attention to abnormal words when the time window slides, and to judge the sudden events through the structural changes caused by the sudden changes of the words in the sub-graph. It is known from the experimental results that the emergency event detection method has a better event detection effect in the real-time blog post data stream. Compared with the existing methods, the emergency detection method proposed in this paper can meet the needs of emergency detection. Not only can it detect the detailed information of sub-events, but also the relevant information of events can be accurately detected.

Keywords: emergencies; detection; word relevance features; word diagram; multi-attribute spectral clustering

随着互联网技术的广泛应用,互联网逐渐转变为信息共享、互动交流的动态平台.43rd 统计报告显示,中国网民规模达 9.89 亿, 其中 99.1%的人通过手机上网[1].社交网络已经成为信息交流和意见表达的主流平台,吸引了很多人来分享包含其观点和感受,在社交虚拟网络上实时议论真实生活中发生的焦点、热度高的事情成为许多用户的一种趋向性消遣,并且对事情发表带有主观性、影响力较强的评论,使得现实生活中的突发事件在社交虚拟网络爆发的时间往往比官方通知的新闻要靠前[2].用户不仅是信息的接受者,也是在社交网络上发表文字评论的创造者. 社会舆论的传播会随着社会网络而滚雪球般扩大,突发事件的发展也会失控.例如,2017 年红黄蓝幼儿园事件、2018 年狮航客机失事事件、2019 年澳大利亚大火灾及 2020 年新的冠状病毒事件,这些事件在微博和微信等社交网络上迅速发酵,是网民的热议点[3-4].由于对社交网络监管不足而导致的连锁事件会带来不良影响,近年来突发事件的频度和危害性呈明显上升趋势.这些突发事件使得国家和人民的财产以及生命安全遭受了很大的损失,并且突发性影响了政府后续的应急处理,包括舆论以及救援等.通过从紧急灾难等大量来源发出及时、准确的警报,事件检测模型可以帮助人们迅速采取行动,以减轻巨大的生命和经济损失.因此,在各种突发事件发生后在社交网络上实时监测事件的演变情况,并采取相应措施控制其发展对舆论指导具有重要的研究意义.

突发事件是指通过网络传播工具和网络平台对即将发生或已经发生的事件发表个人意见,随着时间的推移控制突发事件的进一步扩大将有助于决策者分析整体情况,并根据演变过程做出正确的决策.在这种情况下,有必要确定关键事件并通过时间表对其进行控制.尽管一般很难使公众了解现实生活中的事件,但可以通过在社交媒体上提取和分析与社交事件相关的微博来获取时间信息[5].微博平台可以充当信息源,使个人、公司和政府组织可以随时了解“当前情况”和“人们对它们的看法”.检测突发事件和用户对其的看法至关重要,因为它们可以带来前所未有的宝贵信息.例如,公司可以使用这些信息来分析用户对其产品（或竞争对手）的看法,以回应客户的投诉并改善决策.对于政府而言,他们可以使用这些信息来预测犯罪或监视恐怖活动.与传统的信息传播渠道相比,在社会网络上检测获得的突发事件能更快的了解到事件的详细发展情况,以便相关部门能迅速采取对应策略,将会对社会的稳定和公众的利益产生重要的现实意义.

目前由于微博文本含有大量的口语单词, 网络短语, 广告 , 链接和其他垃圾邮件信息, 在对数据信息进行聚类分析和计算词语相关突发特征的时候, 引入过多无用信息对其造成噪音干扰.另外, 对微博文本进行聚类分析时, 需要对一些参数阈值进行调试以期达到最好的实验效果, 但一般都是以研究的相关经验设定参数阈值, 并且选择的质量会直接影响聚类的结果, 从而对检测的准确性产生影响.

为了解决这些问题, 本文提出了一种结合词语相关特征和多归属谱聚类算法检测突发事件.首先按时间先后对爬取的微博数据进行分段, 利用连续时间划分数据切片; 计算每个时间片段的数据信息的各词语的词频特征、用户影响力和词频增长率特征, 运用突发度计算方法来提取突发词.然后, 利用特征相似性对提取突发词进行矩阵构建, 转化为词语关系图.最后, 运用多归属谱聚类算法对单词关系图进行最优划分, 并在时间窗滑过时关注异常词语, 通过子图中词语突发度的变化而引起的结构变化对突发事件进行判断.

本文的其余部分安排如下.第 1 节对数据进行预处理,提出突发词的提取方式.第 2 节正式提出了社交事件检测算法,并提出了多归属谱聚类的图划分的方法以检测突发事件.第 3 节通过相关实验评估本文所提突发事件检测方法的有效性.第 4 节概述了社交事件检测中的相关工作,提出了本文突发事件检测的流程图.最后,第 5 节对本文工作进行总结.

1 突发词提取模型

现实中, 当一个事件发生时, 很多与该事件相关的消息文本都是在短时间内产生的, 因此很多与该事

件相关的词语在这段时间内都有突发频率.本文将词频增长率和用户的社会影响力与词权重相结合，提出突发词模型检测突发词.

1.1 文本预处理

在进行事件检测之前，对文本进行预处理能够使检测的结果更加准确.文本预处理步骤如下:

(1) 进行噪声过滤，采用 NLPIR (Natural Language Processing and Information Retrieval) 分词系统过滤掉无用文本，包括去除不含事件三要素[6-7]的博文、粉丝数在某一阈值以下的用户，以及文本中包含的图片网址链接、表情符号、英文等.

(2) 使用 BosonNLP 情感词典[8-9]过滤掉含情感的词语，公式(1)所示，

$$Se(n)=\sum_{\omega_i=positive} positive\_word(\omega_i)+\sum_{\omega_j=negative} negative\_word(\omega_j) \tag{1}$$

*Se(n)*为词语的情感度，*n* 表示第 n 个文档，*positive_word*(ω_i)为积极正面的情感词语数量；*negative_word*(ω_j)为消极负面的情感词语数量.

(3) 对文本进行规范.

1.2 突发词特征的分析与表示

(1) 计算词频增长率

在一个时间窗口内，词频特征在单词频率特性中考虑了高频单词，但没有考虑单词频率的变化趋势.如果某个事件刚刚发生，突发的单词只是在T_i时间窗口涌动，就不能通过单词频率以及引入的增长率来重新提取突发正确的单词，以识别意外单词.本文综合一些研究方法，计算词语在某段时间T_n的频率与之前的平均历史频率 $A_{n-1}(\omega)$ 之和.

$$A_n(\omega)=A_{n-1}(\omega)+\frac{f_n(\omega)-A_{n-1}(\omega)}{n} \tag{2}$$

其中，$f_n(\omega)$ 表示词 ω 在时间窗 T_n 下的词频.根据公式(2)对多个连续时间段的词语计算平均增长率，能够显示出单词频率的波动趋势.

(2) 计算用户的社会影响力

在微博上使用过该词的用户也会影响该词的爆发.一般来说，拥有众多粉丝的用户发布的微博会更具影响力，相应地这些用户讨论的事件有很大的潜力能成为突发事件，这会使我们在计算突发度时不够准确，少数高影响力的用户会成为主导因素，一些普通用户的影响力会被大幅度减弱.综上所述，本文采用归一化的方法计算用户的影响力，定义用户 U = (Rep, Com, Fan, Type, Update)，如公式(3)所示.Rep、Com 表示用户一个月之内转发和评论微博数量之和；Fan 表示用户的粉丝数量；Type 表示用户的类型，不同的类型权重不同，即官方认证的微博权重为 1、大 V 即粉丝数量多的微博权重为 0.7、普通用户权重为 0.5；Update 表示用户一个月之内的更博数，最小值不能为零.

$$B_u=\frac{(\mathrm{Re}\, p_u+Com_u)\times Fan_u\times Type_u}{Update_u} \tag{3}$$

在社交网络上，用户的粉丝数量越多，则影响力越大，如流量明星所发布的微博在几分钟内就有可能被几十万人看到.因此，影响力越高的用户对事件传播速度的贡献越大，其中出现词语描述突发事件的可能性也越高.

(3) 计算词权重

在突发事件中，与事件有关的微博会呈井喷式爆发，突发词会频繁的出现在同一事件的不同文本中[10].在微博短文本中，传统 TF-IDF 方法难以衡量关键词与普通词语的差异性，因此采用文献[11]中的文档频率-倒文档频率(DF-IDF)词权重算法.对于网络热议的话题，单词的 DF 会上升；若发生突发事件，单词的 IDF 会呈指数形式上升.该算法弥补了 TF-IDF 方法的缺点，能准确的计算词权重.

$$W_{j,t} = dfidf_j^{t,\tau} = df_j^t \, log\left(1 + \frac{1}{df_j^{t,\tau}}\right) \tag{4}$$

该公式为单词 j 第 t 天的词权重，与传统 TF-IDF 不同，本文 IDF 只限于近期微博（不超过一个月），为第 t $-\tau\sim$ t 天内词语 j 的平均 DF，其 DF 是 $\frac{\left|Y_j^t\right|}{Y^t}$，$Y_j^t$ 表示当天包含单词 j 的博文.由于一般社会事件的关注度都会随着时间而降低，不会超过两周，因此单词的时间段τ被设置为 14.

1.3 计算突发词的方法

为了能更好地得到一个突发词，综合用户影响力和突发词的重要性，突发度的计算公式如下:

$$word_{j,t} = \frac{1}{N}\sum_{k=t-N}^{t-1}(W_{j,t} \times A_t(\omega) \times \sum_{Pn \in Pj,t} lb(B_{Pn}) - W_{j,k} \times A_k(\omega) \times \sum_{Pn \in Pj,k} lb(B_{Pn})) \tag{5}$$

$word_{j,t}$ 越高，说明该词更有可能是突发词.其中 $word_{j,t}$ 代表单词 j 在时间段 t 内的突发度；B_{P_b} 是包含词语 j 的一条微博 p_n 的发布者的影响力； $P_{j,t}$ 是在时间窗 t 内包含词语 j 的所有微博；N 是时间窗的总数.

2 突发事件检测模型

在社交网络上每天都有大量的事件出现和传播.事件的传播和相应的评论会对舆论产生很大的影响.因此，需要一种有效的检测突发事件的方法.

2.1 构建词语关系图

当检测到突发词时，可以通过对这些突发词进行聚类来检测事件.每组突发词代表一个事件.为了对突发词进行聚类，构造一个词关系图，图中的每个顶点表示一个突发词，每个边表示两个对应词之间的关系.根据上述突发词的提取方法，按突发度的高低排序，选择突发度高的 n 个词语，过滤了含大量与事件无关的词语.

假设我们从文本流中连续获取边缘序列，词关系图是无向的，定义为

$$G = (V, E) \tag{6}$$

其中 V 代表从文本流中提取的词语集合，E 是在文本滑动窗口中与词语相对应的边缘集合.具体来说，V 中一个节点上具有相同含义的多个实体或动词.由于图形随着时间的变化，G 中节点之间的边缘权重将发生显着变化.边缘节点 g_i, g_j 在时间 t_s 边缘权重定义为 $R = (g_i, g_j, t_s)$.

给定两个词语矩阵 ω_i 和 ω_j，我们通过余弦距离定义它们之间的语义相似性:

$$sim(\omega_i, \omega_j) = \vec{v}_{\omega_i} \cdot \vec{v}_{\omega_j} \tag{7}$$

其中 $\vec{v}_\omega$ 是从 word2vec 模型计算出的单词的单位向量 ω .

归一化将具有表达式的维数转换为无量纲的表达式后将成为标量，将计算量简化.归一化交叉相似度 $D_{cc}(\omega_i, \omega_j)$ 定义如公式(8)所示，其中 s_i 表示单词 ω_i 的矩阵形式.

$$D_{cc}\left(\omega_i, \omega_j\right) = \frac{S_{\omega_i}^{\mathrm{T}} S_{\omega_j}^{\mathrm{T}}}{\sqrt{S_{\omega_i}^{\mathrm{T}} S_{\omega_i}^{\mathrm{T}}} \sqrt{S_{\omega_j}^{\mathrm{T}} S_{\omega_j}^{\mathrm{T}}}} \tag{8}$$

通过该公式计算，得到词语关系图的相似矩阵，且维度为 n（单词 ω_i 与单词 ω_j 的相似度），相似度高的即为同义词.然后使用 word2vec 模型将多个同义词合并到一个节点中.对于每个词语，将遍历词语关系图上的每个节点，如果相似度超过阈值 θ_{sim}^{mc}，则会将该词语与存在的节点进行比较，并按字典顺序用前一个短语表示.

对于微博文本中多个词语同时出现，我们通过最大化而非累积来更新该词语的权重.遍历所有文本后，通过将权重加在一起来合并它们.热门话题的影响会随着时间的流逝而逐渐消失，因此单词共现度在很长一段时间内都不会稳定下来.为了模拟时间效应，引入衰减因子 λ 来调节单词共现度随时间衰减的速率.公式如

下:

$$C\left(\omega_i,\omega_j\right)=2^{-\lambda}\left(\frac{f\left(\omega_i,\omega_j\right)}{f(\omega_i)}+\frac{f\left(\omega_i,\omega_j\right)}{f(\omega_j)}\right) \tag{9}$$

其中，$f(\omega_i,\omega_j)$ 表示单词 ω_i 和 ω_j 单词在某时间段内微博文本中同时出现的次数，$f(\omega_i)$ 表示词语 ω_i 与 ω_j 在时间窗内出现的总次数. 共现度显示了单词共同出现的频率，数值越高，描述同一事件的概率越大.

2.2 划分词语关系图算法

本文采用基于多归属谱聚类的图划分（Multi-attribute spectral clustering algorithm, MASCA）将词语关系图划分为表示突发事件的子图.谱聚类算法从数据的亲和力矩阵（即相似性矩阵）得出的拉普拉斯矩阵的特征向量，并将数据转换为新的维度，然后可以使用其他最小化失真度量的算法对其进行图划分.在这种情况下，亲和矩阵证明了数据点之间的成对相似性，并用于克服由于数据分布缺乏凸度而带来的困难.具体而言，与 K 均值不同，谱聚类不会在数据上施加超球形聚类，并且在大多数情况下，甚至在数据点不对应于凸区域时，也可以获得令人满意的聚类结果.

(1) 建立目标函数

为了对单词关系图进行最优划分，本文首先运用子图归属度向量表示词语对划分子图的归属程度，使子图内部的单词尽量相似.定义如下:

$$u_r=\left[u_{1,r},u_{2,r},u_{i,r},\cdots,u_{L,r}\right] \tag{10}$$

其中，$u_{i,r}$ 表示单词 ω_i 对第 r 个子图的归属程度，$0\le u_{i.r}\le 1$，L 表示词语的数量.每个子图包含一个事件的突发词，而一个突发词能对应多个事件，即对应多个子图，则不同子图会包含同一个单词.

NJW 方法[12]使用归一化相似度矩阵作为图拉普拉斯矩阵，并通过考虑对应于最大特征值的特征向量，基于归一化割准则优化分区建立目标函数 P 如公式(11)所示.公式的目标是同时考虑最小化 cut 边和划分平衡，即优化不同子图的归属度向量 u_r，以免 cut 出一个单独的词语.W 是词语关系图顶点之间的相似度矩阵，D 是相应的度矩阵.

$$u_r,\forall r:\min P=\min\sum_{r=1}^{M}\frac{u_r^{\mathrm{T}}\left(D-W\right)u_r}{u_r^{\mathrm{T}}Du_r} \tag{11}$$

目标函数 P 的最小化可转化为拉普拉斯矩阵 $D^{\frac{1}{2}}WD^{-\frac{1}{2}}$ 特征值的最大化，使用 U 表示所有子图的归属度矩阵，其定义见式(12)所示:

$$U=\left[u_1,u_2,\cdots,u_M\right]$$

$$\max_{U^{\mathrm{T}}\cdot U}\left\{\mathrm{trace}(U^{\mathrm{T}}D^{\frac{1}{2}}WD^{-\frac{1}{2}}U\right\}=\sum_{i=1}^{M}\lambda_i \tag{12}$$

其中，$U=U(U^TU)^{-\frac{1}{2}}$ 矩阵 U_e 包含拉普拉斯矩阵 $D^{\frac{1}{2}}WD^{1\frac{1}{2}}$ 的 M 个最大的特征值对应的特征列向量，维度为 L×M.

(2) 近似优化归属度矩阵

向量矩阵 U_e 按数学方法进行旋转变换，在不改变向量大小的情况下转换向量原有的方向，保持原矩阵的特性.转换之后得到单词的最优归属度矩阵 U_{opt}，即 $U_{opt}=U_eR$，其中 R 为旋转矩阵，属于单位正交矩阵.由于在连续域空间中优化 U_{opt} 无法得到最优结果，属于 NP 难问题，因此本文运用近似的方法在离散域中对其优化以期得到最好的结果，近似矩阵 $U^a=\left[u_1^a,u_2^a,...,u_M^a\right]$.

近似方法通过衡量近似矩阵 U^a 与最优归属度矩阵 U_{opt} 的误差进行优化，即在约束条件下如何使误差

最小的问题.U^a 与 U_{opt} 通过弗罗贝尼乌斯范数(Frobenius norm)进行表示，公式如下:

$$min R = arg\,min \left\| U^a - U^e R \right\| \;\; subject\ to\ R^{\mathrm{T}} = R^{-1} \tag{13}$$

$$R = \Xi\ \Pi^T\ and\ \ \mathrm{U}^{a^{\mathrm{T}}} U^e = \Pi\ \Omega\ \Xi^T \tag{14}$$

其中 (Π, Ω, Ξ) 是矩阵 U^{a^T} ,U^e 的奇异值分解矩阵，矩阵 Π 和 Ξ 均是正交矩阵.使用迭代的方法进行求解，具体算法伪代码如算法 1.

算法 1. 优化归属矩阵

输入: *n, m, U*

输出: U_{opt}

1: $R \leftarrow ortho\left[U, m\right]$

2: $value \leftarrow 1000$

3: **For** $k \leftarrow begin\ of\ 1\ to\ 10$ **do**

4: $U^I \leftarrow U[:, :m]R$

5: $U^a \leftarrow arg\,min\left(Frobenius\left(U^a - U^I\right)\right)$

6: $U_{svd}, s, V_{svd} \leftarrow SVD\left(U^{a^T}, U^I\right)$

7: $R \leftarrow V_{svd}^T U_{svd}$

8: $v \leftarrow Frobenius\left(U^a - U^I\right)$

9: **if** $v < value$ **Then**

10: $value > v$

11: $U_{opt} \leftarrow U^a$

12: **end if**

13: **end for**

14: **renturn** U_{opt}

(3) 确定聚类个数

谱聚类划分将微博文本数据聚类转换为单词关系图的多向划分问题，解决图划分的关键是找到准确的聚类个数.当确定了聚类的个数时，能够优化通过近似方法求出的近似矩阵值，并进一步精确.在本文中，为了使算法更适用于突发事件检测的实时应用场景，最优聚类个数由特征值的下降程度决定，由于下降程度无法精确，因此是近似估计.

算法 2 给出了确定聚类个数的伪代码.使用该方法计算最优聚类个数的线性时间复杂度为 O(L)，可以及时地检测出实时事件，避免时间延误.运用归属度矩阵优化的方法划分单词关系图，由算法得出的最优聚类个数是多少，则划分子图的个数就是多少.

算法 2. 使用特征值向量优化聚类个数

输入: D, W

输出: M_{opt}

1: $M_{opt} \leftarrow 0$

2: $U \leftarrow Eigenvalue\_Decomposition\left(D^{-\frac{1}{2}} W D^{\frac{1}{2}}\right)$

3: $U_{diff} \leftarrow Diff(U)$

4: $M_{opt} \leftarrow Index\left(U_{diff}\right)$

5: **return** M_{opt}

(4) 识别突发事件

子图划分之后，每个子图包含若干个突发词，这些突发词组成一个事件，即每个子图代表一个事件的集合.判断事件是否为突发事件由对应的单词关系图结构是否发生变化决定，即突发事件发生时，短时间内会出现与该事件有关的大量微博文本，而这些文本中会包含高突发度的词语，并出现在构建关系图的单词集合中.此时，发生变化的词语会显示出突发性，构成新的单词关系图.因此，在关系图中单词突发度发生了

改变，代表突发事件产生，算法伪代码如算法 3.

算法 3. 判定突发事件

输入: *Graph G, Graph S,* μ

输出: *true/false*

```
Get events set μ by searching event index
    If cos (s Graph G, Graph S) < μ Then
        return true
    end if
end for
return false
```

算法 4 说明了突发事件与文本聚类簇的映射关系，比较了事件关键词集合和聚类簇的关系，通过循环，找出与事件关键词集合相似度最大的文本聚类簇，即为突发事件的具体信息.

算法 4. 将子图结果映射到文本聚类簇

输入: *subgraph,cluster*

输出: *eventcluster*

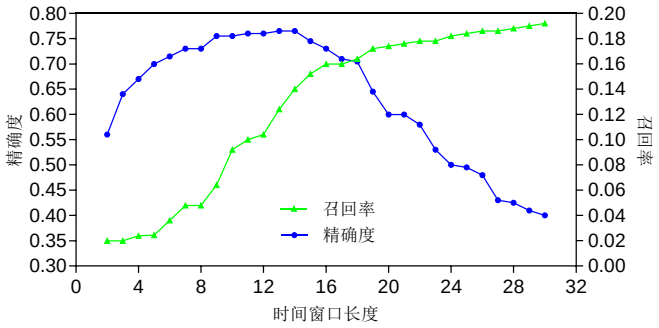
```
candidate_cluster ← [ ]
For word in subgraph do
    For subcluster in cluster do
        If word in subcluster Then
            candidate_cluster ← subcluster
        end if
    end for
end for
new_cluster ← ( Sort can)c
For subcluster in new_cluster do
    If Similarity (subgraph, subcluster)>sim Then
        sim ← Similarity(subgraph,subcluster)
        eventcluster ← subcluster
    end if
end for
return eventcluster
```

3 实验评估

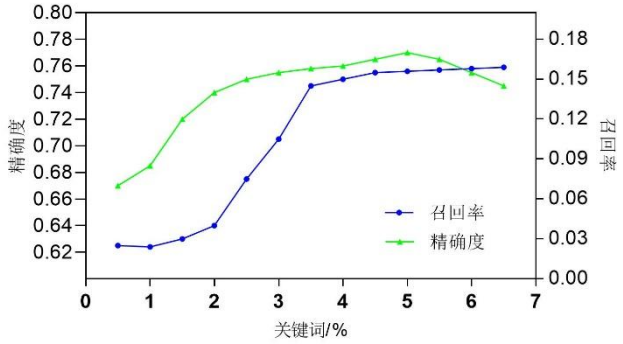
本文使用的数据集来自新浪微博，通过模拟微博登陆来爬取微博数据采集了 2019 年 11 月 1 日至 11 月 30 日间提取的 1000 万条微博，这些数据因为规模较大没有进行事件标注.由于微博不仅包含一些官方新闻事件，也包括娱乐新闻事件[13-17]，因此本文以官方新闻热议事件作为微博事件的参考.对于所有数据集，本文使用一种预处理技术，该技术在 3.1 节中进行了说明.表 1 给出了数据集所包含的内容，它描述了用户 ID、用户信息、转发量、评论数、粉丝数、发布时间和博文内容.所有实验均在具有 512GB 内存并在 Windows 10 上运行的 8.00 GHz Intel CPU 上进行.我们实现了该算法，以获取准确的突发事件并验证检测是否成功.

3.1 提取突发词效果测试

为了测试突发词提取模型的效果，从数据库中抽取 2019 年 11 月 20 日到 2019 年 11 月 30 日共计 10 天的数据，并对抽取的数据进行文本预处理.首先分析时间窗口参数对突发事件检测结果的影响，如图 1（a）所示；然后在分析提取突发词数量的多少是否会影响实验结果，如图 1（b）所示.



(a)



(b)

Figure 1 The impact of different burst word extraction parameters on event detection

图 1 不同突发词提取参数对事件检测的影响

如图 1(a)所示，时间窗口长度过小时，事件的准确率和召回率较小，IDF 仅在短期内被平均化，使关键词提取模型会受到干扰，并且容易获取到大量毫无关联的关键词.当时间窗口长度在 2 到 14 之间，准确率和召回率都呈逐渐上升趋势，无关联的关键词被剔除掉，对检测效果产生正面影响.当时间窗口长度继续增加，准确率继续上升，召回率下降较快.为使准确率和召回率都在一个大的数值范围上，时间窗口长度取 14.由图 1(b)可知，关键词数量较少，无法检测到突发事件，因此召回率和准确率都比较低.当关键词数量从 2%增长到 4.5%时，召回率和准确率都达到了顶峰，而当关键词数量继续增加时，太多的关键词容易引起混乱，使得检测效果变差(准确率下降).因此为了使检测效果最好，使用整个数据集 4.5%的词语来提取突发词.

按时间窗口长度为 14 对这 10 天的数据进行划分，根据公式得到突发词的突发度，经计算发现 11 月 27 号突发词存在突发度异常高的词语，如表 1 所示.

Table 1 The suddenness TOP table of some sudden words on November 27, 2019

表 1 2019 年 11 月 27 日部分突发词的突发度 TOP 表

词语	突发度	排名
高以翔	0.914037512	1
去世	0.680050085	2
意外	0.546576028	3
浙江卫视	0.344115381	4
追我吧	0.286655149	5
晕倒	0.20638858	6
心脏复苏	0.059574521	7

表 1 显示了突发度排名前七的词语,“高以翔”、“去世”、“意外”是排在前三的词语，与当天的新闻对比，高以翔去世事件是整个社交网络热议的话题.由此知道单词的突发性越强，更有利于缓解事件的早期发现问题.

3.2 图划分效果测试

(1) 词关系图参数测试

词关系图是进行谱聚类图划分的基础，因此分析基于图聚类的事件检测效果.如图 2 所示，分析关系图节点近邻数的大小对突发事件检测效果的影响.当节点近邻数较少时，即突发词之间的关系不足，极大地影响了事件的检测效果.直到数量达到 6 时，召回率和准确率都是最大值，事件检测的性能才最好.

图 3 显示了突发词相似度阈值的变化对突发事件检测的影响.可以发现，事件的准确率随着相似度阈值的增大而上升，表明突发词的相似度越高，越容易检测到突发事件.但阈值太大，会过滤掉一些相似度较小的突发词，导致事件的召回率较低.考虑到准确率和召回率的平衡，选择两者交点处的阈值，即 1.2.

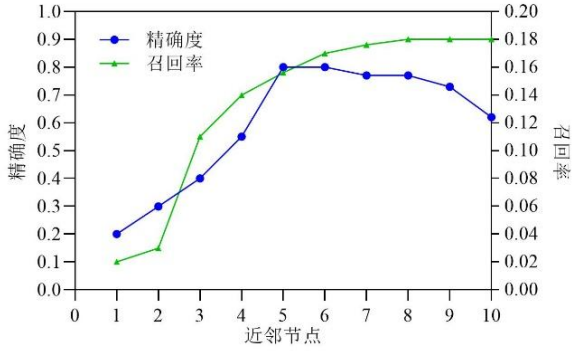


Figure 2 The impact of the number of neighbors of the word relationship graph on the performance of event detection

图 2 词关系图节点近邻数对事件检测性能的影响

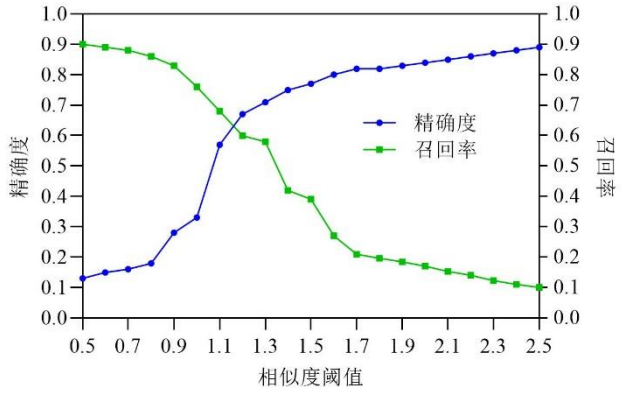


Figure 3 The impact of similarity threshold on event detection performance

图 3 相似度阈值对事件检测性能的影响

根据上述结果调好参数之后，选取突发度较高的 8 个单词，对单词集合为“高以翔、意外、去世、维 E 乳、协和、北京、鼠疫、确诊”按顺序构建单词关系图，八个单词的关系网络图如图 4 所示.实线表示两个词语之间相似度高（在 0.7 以上），细虚线表示词语之间相似度较低，粗虚线表示通过 word2vec 模型连接的边.其中第一个到第三个单词属于高以翔在录制浙江卫视一档节目中去世事件的关键词，第四个到第六个单词属于北京协和医院没出过协和维 E 乳事件的关键词，单词 6、7、8 是南开校长曹雪涛假论文事件的关键词.

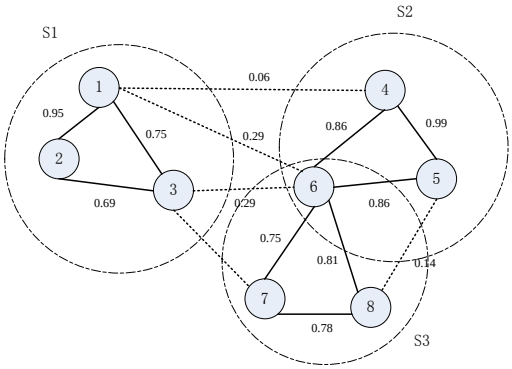


Figure 4 Schematic diagram of the effect of the word relationship diagram

图 4 词关系图效果示意图

(2) 多归属谱聚类效果测试

利用 2019 年 11 月 1 日至 11 月 30 日的微博数据根据提出的词的突发度计算公式得到了词的突发度，得到排名前几位的关键词以及与之相关的突发事件热度如下图所示，本文对 11 月份的突发事件进行分析.与该事件相关的突发关键词如图 5 所示.在此图中，这些关键词的趋势是相同的.同样，与不同事件相关的相同关键字也具有此特征，如图 6 所示.这两个图揭示了关于不同事件的关键字彼此之间具有某些语义相关性，并且相互影响.

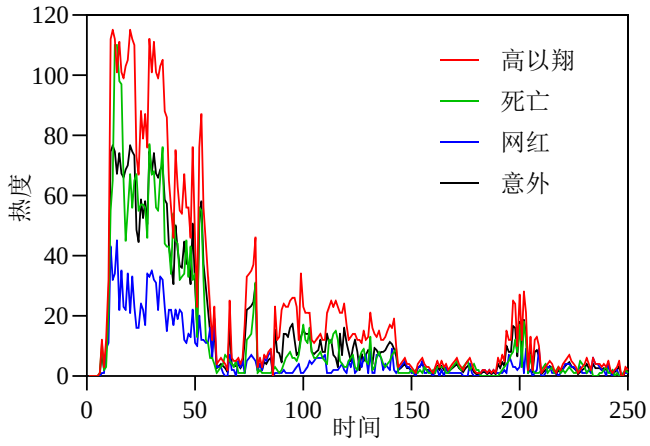


Figure 5 Popularity frequency of sudden keywords

图 5 突发关键词的热度频率

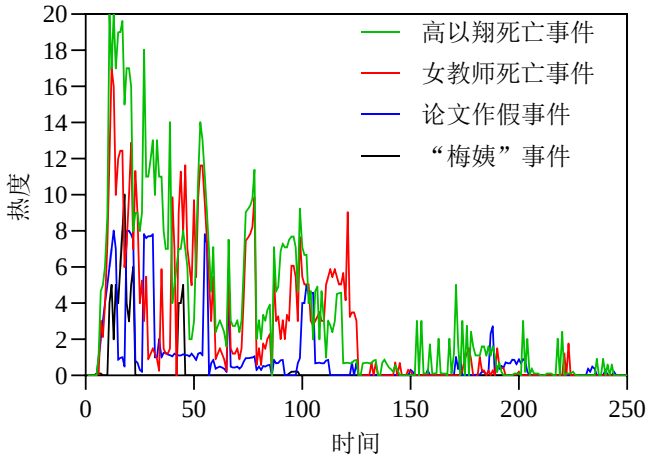


Figure 6 The frequency of emergencies

图 6 突发事件的热度频率

因此，最终选取突发度排名前 70 的突发词构建词关系图，得到 58 个词语组成的关系图.再利用 MASCA 算法对关系图进行划分，并且给出了图划分的最优个数，为 7，具体如表 2 所示.

Table 2 Schematic table of graph division results

表 2 图划分结果示意表

子图编号	词语集合
1	高以翔，意外，死亡，浙江卫视，追我吧，心脏，复苏，医院，抢救
2	梅姨，人贩子，拐卖，儿童，画像，落网，死亡，9 起
3	曹雪涛，南开，校长，论文，作假，PS，数据，图像
4	四川，女教师，坠楼，死亡，26 岁，巴中，丈夫，家暴，测谎
5	维 E 乳，北京，协和，侵权，网红，带货，作假，不实，消费者
6	鼠疫，北京，病例，确诊，锡林郭勒盟，内蒙古四子王旗，公共卫生
7	大唐，不夜城，不倒翁，真人版，网红，抖音，冯佳晨，模仿

表 3 给出了每个子图中词语集合信息，词语“死亡”同时属于事件编号为 1、2、4，单词“作假”同时属于事件编号 3、5，词语“网红”都包含在事件 5 和 7 里面，词语“北京”都包含在事件 5 和 6 里.根据词语子图与文本的映射算法，显示出与突发事件有关的微博文本信息.7 个子图与高以翔意外去世事件、“梅姨”事件、南开校长曹雪涛假论文事件、四川巴中 26 岁女教师坠亡事件、北京协和医院没出过协和维 E 乳事件、鼠疫事件和大唐不夜城不倒翁刷屏事件依次对应，表明 MASCA 算法能准确的划分词关系图，有效的解决词语的重复问题，因此可以更早地嗅探到突发事件，并且取消引用知识库可以节省时间.

3.3 突发事件检测效果测试

表 3 显示了突发事件检测算法中事件相似度阈值参数μ的各项指标，它能衡量检测突发事件的难易程度，与参数值成反比，即参数值越高，检测到的突发事件数量就越多.为了选择最佳的参数值，设置了阈值参数μ为 0.5、0.6、0.7、0.8、0.9 时，相对应的指标大小，并对其进行比较.

Table 3 Influence of μ value on experimental results

表 3 μ值对实验结果的影响

阈值	Precision(%)	Recall(%)	F1(%)
0.5	88.92	77.23	81.94
0.6	84.98	82.24	84.02
0.7	82.57	87.95	85.11
0.8	75.1	90.07	82.01
0.9	63.33	90.12	73.99

相似度阈值参数μ的 Precision、Recall 和 F1 在不同阈值下的变化趋势如图 7 所示.Precision 随着阈值参数μ的增加而逐渐下降，0.7 到 0.9 的区间下降幅度较大；与之相反，μ越大，Recall 也随着增大，0.8 至 0.9 区间基本保持不变；而 F1 的变化趋势是先增大然后减小，在$\mu = 0.7$时，F1 值最大，此时突发事件检测算法达到最优的效果，与之对应的 Precision、Recall 分别为 82.57%、87.95%.因此在检测突发事件时，事件相似度阈值参数μ取 0.7.

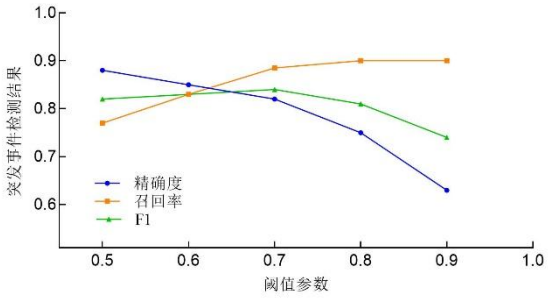


Figure 7 The effect of emergency detection

图 7 突发事件检测效果图

在国内微博爆发检测中，尚未识别手动标记的语料库[18-20].因此，结合使用 Search Billboard 中的微博和微博数据本身，可以手动注释 30 天的紧急情况，包括高以翔意外去世事件、“梅姨”事件、南开校长曹雪涛假论文事件、四川巴中 26 岁女教师坠亡事件、北京协和医院没出过协和维 E 乳事件、鼠疫事件和大唐不夜城不倒翁刷屏事件等大大小小 32 件事.近一个月内社交网络上热议最多的 7 个突发事件在表 4 显示，包含了事件的基本信息——事件编号、描述事件的微博文本、与事件相符的子图词语数量、单词重合率.与单词重合率代表子图中包含了多少突发事件的关键词不同，子图单词重合率是衡量子图与事件是否相符的指标.该值越大，子图与事件的相符程度越高，包含事件关键词的数量就越多.从突发事件检测的 Recall 值来看，子图单词都能描述出对应事件的发展经过，同时子图单词重合率平均值为 0.8929 ，表明本文提出的算法能准确地划分单词关系图，并且被划分的子图内单词集合能对事件进行简单的表达.

由事件检测结果知，本文提出的突发事件检测算法能准确的识别出突发事件，并且通过不同时刻单词关系图的变化反映事件在不同时间的演变趋势，说明本文提出的突发事件检测方法检测事件更全面.

Table 4 Detection results of some emergencies

表 4 部分突发事件检测结果

事件编号	事件描述博文	子图单词数量	单词重合率
1	2019 年 11 月 27 日凌晨 2 点左右，网络上有消息曝出艺人高以翔在录制浙江卫视一档夜晚城市实境追跑真人秀《追我吧》过程中晕倒，现场经过十几分钟的心脏复苏后被送往医院抢救.	9	1.0
2	2019 年 11 月 14 日起，全网的热议对象“梅姨”是一位作案十几年的人贩子，但是一直都没有落网.十几年间，经手交易了 9 名被拐儿童，目前为止仅有两名受害儿童得到解救.	9	0.95
3	2019 年 11 月 14 日，微博网友@长夜扁舟 发布一则消息“南开大学校长曹雪涛院士也是用 PS 代替做实验的高手，至少有 18 篇论文被发现数据造假”将 Elisabeth Bik 博士对曹雪涛论文“图像异常”问题的质疑带到了国内社交媒体.	10	0.9
4	2019 年 11 月 7 日，四川巴中 26 岁女教师坠楼身亡疑似因为丈夫家暴.	11	0.85
5	2019 年 11 月初，一款名为“协和维 E 乳”护肤品成为网友热议焦点.仅 7 天就卖出 51 万瓶，顶峰时期有超 100 个主播同时推广.有媒体发现，市场流通各种各样的 “维生素 E 乳”，仿照与协和研制标婷维生素 E 乳一样的包装有十几种，蹭了协和的知名度，但是产品与北京协和毫无关系.	10	0.9
6	2019 年 11 月 12 日关于“北京确认接诊鼠疫病例”的消息引发网络关注度最高； 11 月 17 日，“锡林郭勒盟一病人被确诊为腺鼠疫”的消息再次引发网络关注度；11 月 28 日，“内蒙古四子王旗确诊一例腺鼠疫”的消息网络关注度再度掀起.	9	0.8
7	抖音平台上西安大唐不夜城不倒翁小姐姐牵手视频在网络上得到许多网友的关注，迅速成为网红.真人版的不倒翁扮演者为冯佳晨，吸引了各地游客前来西安打卡，也使得许多商家以及网友纷纷开始模仿不倒翁.	8	0.85

3.4 事件检测评价

这一部分检测比较本文与其它文献的方法，使用标准指标 Precision(P)，Recall(R)，F1-Measure(F1)和评估量化模型的有效性，计算如下:

$$precision = \frac{B \cdot correct}{B \cdot outout}$$

$$Recall = \frac{B \cdot correct}{B \cdot number}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

其中: $B \cdot correct$ 为系统中识别正确的突发事件个数，$B \cdot number$ 为数据集中事件的总数量，$B \cdot outout$ 为数据集手动标注的突发事件个数.

(1) 准确率(Precision) 、召回率(Recall) 与 F 值(F-measure)对比

文献[37]中提到的基于词共现图的方法，将微博数据进行预处理，根据主题词间的共现度构建词共现图，把词共现图中每个不连通的簇集看成一个新闻话题进行突发事件检测，当共现度阈值选 0.6 时 F 值最高，达到 0.6615，正确率是 0.6454，召回率是 0.77.在文献[6]中，通过博文的转发关系，跟随关系和转发时间创建消息传递图，然后从图结构方面提取时间演化特征识别突发事件，当时间演化聚类距离阈值选取 0.8 时 F 值最高，达到 0.7668，正确率是 0.7364，召回率是 0.8050.与本文方法的正确率、召回率、F 值相比较，如图 8

所示.

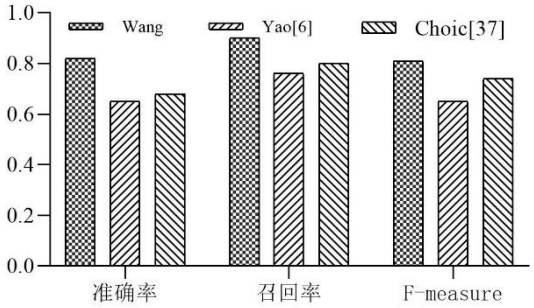


Figure 8 Comparison of experimental results

图 8 实验结果对比图

由图 8 可知, 本文提出的方法 在 Precision、Recall 与 F 值上都要优于其它两个方法, 这是由于本文为了解决微博的时间特性专门设计了一种新型词语突发度以及词语矩阵相似度的计算方法, 使得提取的突发词全面准确, 能够更好地对突发事件进行描述.并且本文采用的基于多归属谱聚类的图划分的事件检测方法能够使突发词构建的共现图包含较大较全的信息量, 提高检测的准确率.

(2) 事件检测时延

检测时间是指事件发生到检测到事件之间的时间间隔, 它反映了算法的效率[21-22].我们选择 30 种通过给定 5 种方法成功检测到的事件.MASCA 算法的计算时间复杂度为 O(L), 通过提取突发词再构建关系图, 是线性时间复杂度, 能够满足实时事件检测的要求, 由图 9 可知, 本文提出的方法在检测时延上有较好的结果.在所有方法中, 我们的方法花费最少的平均时间进行事件检测.由于此数据集中每个事件的稀疏分布, 因此所有方法比由预定义事件组成的其他数据集花费的时间更长, 说明本文提出的突发事件检测方法在较短的时间内能够检测到结果, 能使相关人员及时的采取措施进行控制.

值得注意的是, 本文发现实验中其他方法的召回率比 MASCA 低得多, 检查了真实数据, 发现关系图中最早和最新的事件不一定彼此相似.例如, 灾难性事件可能演变成政治事件, 因为政府官员可能会对事故负责.但是其他方法将它们视为无关事件, 因为它没有达到阈值.在我们的方法中, 我们获得了由最相似事件之前已经构造的旧关系图, 并将我们的候选事件放入其中, 因此事件不需要足够相似就可以放在一个图中, 这会增加召回率.

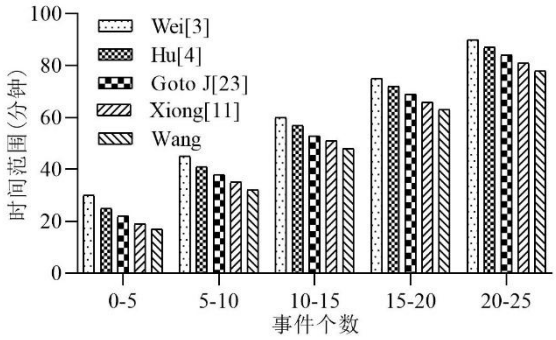


Figure 9 Delay comparison diagram of emergency detection

图 9 突发事件检测时延对比图

4 相关工作

我们主要从文本分析, 突发特征分析以及用户特征分析这 3 个方面介绍相关工作.

4.1 以文本为中心的突发事件检测方法

以文本为中心是将文本语义之间的相似程度通过相关方法度量为距离对文本进行聚类分析，根据聚类结果检测突发事件.该方法将单词的时间序列离散为一小组级别，记录每个单词和每个单词对的出现次数.然后通过滑动时间窗口将共现标记聚类，形成候选事件簇，对满足相应突发规则的类进行突发事件的识别[23-25].李莹莹[26]通过聚类定义了有关事件的隐式语义信息，以引入相关事件，对具有相同主题的意外事件进行聚类，该聚类是在监视事件演变的社交网络中进行的.张婧丽[27]通过计算事件检测标签的文本框架类型相似度方法来识别框架，从而检测出一种紧急情况.并改进紧急情况触发词的识别，能更正确地识别触发词，有效提高识别率.陆垚杰[28] 基于不确定的语言变量构建突发事件模型，减少了文字语言的干扰，从文本的语法和语义两个角度进行研究，使突发事件的检测模型更具准确性.Zhu Z[29]提出了一种改进的术语频率逆文档频率（TF-IDF）算法，称为 TA TF-IDF，用于根据时间分布信息和用户注意来查找热门术语，从而实现对新闻中热点话题的检测.Hossny A H[30]等提出了最佳关键词选择方法，包括单词计数，单词形式（例如 n-gram，skip-gram 和单词袋）以及包括关联技术和相似性技术的数据关联方法，从而检测突发事件.Comito C[31]提出了一种在线算法，可以将推文流逐步分组.该方法通过维护大量文本和时间特征，将经过检查的推文汇总到聚类中心，使得该方法能够有效地发现特定主题上的兴趣组.

4.2 以突发特征为中心的突发事件检测方法

在这类方法中，第一步获取与突发事件相关的微博内容特征，然后对得到的突发事件相关特征进行聚类分析，最后一步是根据聚类算法的结果获取突发事件的相关信息.张鲁民等[32]在微博上建立了一个情绪符号模型，以确定一般情况下网民的情绪可以控制事件传播的程度，紧急情况的发生导致相关事件的信息量迅速上升，网民的情绪也随着评论起伏不定.因此，对微博的原始文本和评论内容进行情感分析可以显着提高紧急事件检测的准确性，但只考虑网民的情绪变化还不够全面.仲兆满等[33]考虑到地域突发特征，构建了基于网络地域的突发事件检测方法，但是该方法不能检测到没有地域突发特征的内容，例如“仝卓往届生”事件. Dong 等[34] 提出了一种在线突发事件检测框架,该框架基于滑动时间窗和两级哈希表检测异常消息。结合事件特征，采用在线增量聚类算法对异常消息进行聚类，检测突发事件。在实时微博消息流环境下的实验结果表明，该框架可以用于在线突发事件检测，与其他方法相比具有更高的准确率。.Xiaomei Z 等[35]提出了一种结合情感和主题标签的模型，以在线检测微博流的中文突发事件，但在某些活动没有任何标签的情况下，这种方法将失败.张仰森等[36]引入网页排名的方法，对用户影响力的比值进行计算，并提取了突发词特征来发现突发事件.该方法引入了用户影响力因素，但是一些僵尸用户以及“水军”也被引入，增加了噪音信息.Choi H J 等[37]提出了一种基于高效模式挖掘（HUPM）的 Twitter 上新兴话题检测方法，通过 HUPM 查找具有高频率和高效用的单词组.Yang W 等[38]提出了一种突发事件检测方法，用于量化微博文本的影响.通过提高影响力微博的分析，挖掘突发词，构造潜在突发事件数据集，并采用 k-means 聚类分析方法对事件进行检测.Zhang Q[39]提出了一种基于突发项值计算和伪突发项识别的突发主题检测方法 BTDF，通过使用术语的基本权重和突发权重来提取突发项，并通过分析术语的新颖性来过滤伪突发项，但没有对无效突发项的过滤.

4.3 以用户行为特征为中心的突发事件检测方法

以用户行为特征为中心是对用户的社交媒体行为数据进行分析，在突发事件检测系统输入用户行为数据，判断系统检测的结果是否与现实事件基本相同.Gupta 等人[40] 对 10,350 条独特的 tweet 进行了特征分析，以了解伪造图像传播的时间、社会声誉和影响模式.并利用用户行为特征和文本特征构建分类器进行研究，结果显示，在 10,215 位用户中，排名前 30 位的用户（0.3％）导致了 90％的伪造图像转发.Wang Z 等人[41]研究用户转发行为，提出了一种基于多层个人信息（MII）和动态时间序列（DTS）算法的用于谣言事件检测的新型两层 GRU 模型，称为 MII–DTS-GRU.在新浪微博数据集上的实验结果表明，他们的模型达到了 96.3％的高精度.Nazir F 等[42]提出了一种使用来自 Twitter 数据的情感分析（推特数量，顶部主题标签和情感分析）进行突发事件信号检测的方法，但是无法检测出事件的具体内容.Ansah J 等[43]提出了一个名为 SensorTree 的新颖的突发事件检测框架，利用社区中用户之间的网络结构连接来进行突发事件检测.赵海林

[44]提出了一种基于用户行为特征的监督式机器学习事件确定方法，利用从推文文本和元数据中提取的统计特征，并在突发序列中将推文簇的特征对应于紧急情况确定，以实现分类器.但是有许多用户行为与国家安全无关，这将延迟紧急情况的判断时间.介飞[45]针对网络媒体的突发问题隐式事件，根据检测到的事件来分析突发社会行为特征，引入关键词功能，动态调整每个候选关键词的时间窗.不同事件具有不同的关键字功能绑定，避免了事件之间的干扰，准确地识别了隐性突发事件，但对于单词中的巨大语义变化，并不适用.

5 总结

随着社交网络的普及，每分钟都有大量的信息产生.因此，用户很难知道现在发生了什么事件，这个事件会有多受欢迎.在本文中，我们提出了一种结合词相关性特征和 MASCA 算法的模型，用于检测微博流的中文突发事件.在此模型中，引入了增量 word2vec 以在检测过程中合并同义词，以词语的基本特征为基础，通过使用 DF-IDF 和用户影响力提取事件的突发词，结合词语关系图和事件的相似性度量来进行图划分.当任务完成时，我们不仅可以检测突发事件，还可以提取人们对突发事件的把握.实验结果表明，本文的方法具有很高的性能和有效性.为了提高性能，我们的检测模型对相关参数进行调整，得到了最优检测性能，当共现度阈值取\mu= 0.7 时，Precision、Recall 与 F 值都有良好的效果.为了提高有效性，本文的方法在精度，召回率和时间延迟方面均优于其他方法.

由于社交网络不仅是文本信息，也有其他非结构数据.因此，在未来的工作中，会继续对突发事件的检测模型进行优化，并加入更多的其他模态数据，使检测更加准确，并能使用多方面的信息对事件进行描述.

References:

[1] CNNIC released the 47th "Statistical Report on China's Internet Development Status", 2021, http://www.gov.cn/xinwen/2021-02/03/content_5584518.htm.

[2] Liu Y, Peng H, Li J, et al. Event detection and evolution in multi-lingual social streams. Frontiers of Computer Science, 2020, 14(5): 1-15.

[3] Wen J, Wang HJ, Deng J, Liu PF. Abnormal Event Detection Based on Deep Learning. Chinese Journal of Electronics, 2020,48(02):308-313.

[4] Hu WB, Wang H, Yan LP, Qiu ZY, Nie C, Du B. Event Detection Method for Social Networks Based on Node Evolution Fluctuations. Journal of Software, 2017,28(10):2693-2703.

[5] Pradhan A K, Mohanty H, Lal R P. Event detection and aspects in twitter: A bow approach. International Conference on Distributed Computing and Internet Technology. Springer, Cham, 2019: 194-211.

[6] Yao ZY, Tu SZ, Huang ML, Zhu XY. A Semi-supervise method for Filtering Chinese Spam Tweets. Journal of Chinese Information Processing, 2016, 30(5): 176－186.

[7] Wang Y, Xiao SB, Guo YX, Lv XQ. Research on Chinese Micro － blog Bursty Topics Detection. New Technology of Library and Information Service, 2018(2): 57－62.

[8] Fei SD, Yang YZ, Liu PY, Wang J. Method of bursty events detection based on sentiment filter. Journal of Computer Applications, 2015, 35(5): 1320－1323.

[9] Chen GL. Micro － blog Emergencies Detection Approach Based on Burst Words Distinguishing. Journal of Intelligence, 2014(9):123－128.

[10] Guo YX, Lv XQ, Li Z. Bursty topics detection approach on Chinese microblog based on burst words clustering. Journal of Computer Applications, 2014, 34 (2) : 486 －490, 505.

[11] Xiong Y, Zhang YF, Feng S, Wang DL. Event detection and tracking in microblog stream based on multimodal feature deep fusion. Control and Decision,2019,34(7):1409-1416.

[12] Shi J, Malik J. Normalized cut and image segenmation. IEEE Trans Pattern Anal. Mach Intell, 2000, 22(8): 888-905.

[13] Zheng F R, Miao D Q, Zhang Z F, et al. News topic detection approach on Chinese microblog. Computer science, 2018, 39(1): 138－141.

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

[15] Li Y. Analysis of Non-Parametric Event Evolution in Social Networks. 2018.

[16] Hu L, Yu S, Wu B, et al. A Neural Model for Joint Event Detection and Prediction. Neurocomputing, 2020.

[17] Pekar V, Binner J, Najafi H, et al. Early detection of heterogeneous disaster events using social media. Journal of the Association for Information Science and Technology, 2020, 71(1): 43-54.

[18] Wang Z, Guo Y. Rumor events detection enhanced by encoding sentimental information into time series division and word representations. Neurocomputing, 2020.

[19] Guo L, Li BC, Zhao JL. Topical Word Embedding Cluster Based New Event Detection within Topics. A Semi-supervise method for Filtering Chinese Spam Tweets. Journal of Chinese Information Processing, 2019, 33(6): 64-71, 79.

[20] Wang K, Hong Y, Qiu YY, Yao JM, Zhou GD. Combining Context Dependency and Sentence Semantic Representation for Event Nugget Detection. Journal of Frontiers of Computer Science and Technology, 2018, 12(3): 423-431.

[21] Dai X, Huang XF, Tang R, Jiang MT, Chen XS, Wang HZ, Luo L. Subtopic Detection Algorithm Based on Hierarchical Clustering. Journal of South China University of Technology（Natural Science Edition）, 2019, 47(8): 84-95.

[22] Janani R, Vijayarani S. Text document clustering using spectral clustering algorithm with particle swarm optimization. Expert Systems with Applications, 2019, 134: 192-200.

[23] 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. Tokyo: ACM, 2018. 1．

[24] Zhou G, Zou HC,Xiong XB, Huang YZ. MB-SinglePass: Microblog Topic Detection Based on Combined Similarity. Computer Science, 2017, 39(10):198－202.

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

[26] Li YY, Ma S, Jiang HY, Liu Z, Hu CM, Li X. An Approach for Storytelling by Correlating Events from Social Networks. Journal of Computer Research and Development,2018,55(9):1972-1986.

[27] Zhang JL, Zhou WX, Hong Y, Yao JM, Zhou GD, Zhu QM. Frame Semantics Based Training Data Expansion for Supervised Event Detecting. Journal of Chinese Information Processing, 2019,33(5):82-92+131.

[28] Lu YJ, Lin HY, Han XP, Sun L. Linguistic Perturbation Based Data Augmentation for Event Detection. Journal of Chinese Information Processing, 2019,33(7):110-117.

[29] Zhu Z, Liang J, Li D, et al. Hot topic detection based on a refined TF-IDF algorithm. IEEE Access, 2019, 7: 26996-27007.

[30] Hossny A H, Mitchell L, Lothian N, et al. Feature selection methods for event detection in Twitter: A text mining approach. Social Network Analysis and Mining, 2020, 10(1): 1-15.

[31] Comito C, Forestiero A, Pizzuti C. Bursty event detection in Twitter streams. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(4): 1-28.

[32] Zhang LM, Jia Y, Zhou B, Zhao JH, Hong F. Online Bursty Events Detection Based on Emoticons. Chinese Journal of Computers, 2013, 36(8) : 1659－1667.

[33] Zhong ZM, Guan Y, Li CH, Liu ZT. Localized Top-k Bursty Event Detection in Microblog. Chinese Journal of Computers, 2018, 41 (7) : 1504－1516.

[34] Dong GZ, Gao J, Huang L, et al. Online burst events detection oriented real-time microblog message stream. Computers, Materials and Continua, 2019, 60(1): 213-225.

[35] Xiaomei Z, Jing Y, Jianpei Z. Sentiment-based and hashtag-based Chinese online bursty event detection. Multimedia Tools and Applications, 2018, 77(16): 21725-21750.

[36] Zhang YS, Duan YX, Wang J, Wu YF. Microblog Bursty Events Detection Method Based on Multiple Word Features. Chinese Journal of Electronics,2019,47(09):1919-1928.

[37] Choi H J, Park C H. Emerging topic detection in twitter stream based on high utility pattern mining. Expert systems with applications, 2019, 115: 27-36.

[38] Yang W, Li D, Liang F. Sina Weibo Bursty Event Detection Method. IEEE Access, 2019, 7: 163160-163171.

[39] Zhang Q, Du J, Kou F, et al. Bursty Topic Detection Based on Bursty Term Detection and Filtration. Chinese Intelligent Systems Conference. Springer, Singapore, 2019: 211-219.

[40] A. Gupta, H. Lamba, P. Kumaraguru, et al. Faking Sandy:characterizing and identifying fake images on Twitter during Hurricane Sandy[C]. Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, 2013, 729-736

[41] Wang Z, Guo Y. Empower rumor events detection from Chinese microblogs with multi-type individual information. KNOWLEDGE AND INFORMATION SYSTEMS, 2020.

[42] Nazir F, Ghazanfar M A, Maqsood M, et al. Social media signal detection using tweets volume, hashtag, and sentiment analysis. Multimedia Tools and Applications, 2019, 78(3): 3553-3586.

[43] Ansah J, Liu L, Kang W, et al. Leveraging burst in twitter network communities for event detection. World Wide Web, 2020: 1-26.

[44] Zhao HL. Research on the Twitter Event Detection Method Based on Users' Behaviors. University of Electronic Science and Technology of China,2018.

[45] Jie F, Xie F, Li Lei, Wu XD. Latent Event-related Burst Detection in Social Networks. Acta Automatica Sinica, 2018,44(04): 730-742.

附中文参考文献：

[1] CNNIC 发布第 47 次《中国互联网络发展状况统计报告》, 2021, http://www.gov.cn/xinwen/2021-02/03/content_5584518.htm.

[3] 闻佳,王宏君,邓佳,刘鹏飞.基于深度学习的异常事件检测.电子学报,2020,48(2):308-313.

[4] 胡文斌,王欢,严丽平,邱振宇,聂聪,杜博.面向节点演化波动的社会网络事件检测方法.软件学报,2017,28(10):2693-2703.

[6] 姚子瑜,屠守中,黄民烈,朱小燕.一种半监督的中文垃圾微博过滤方法.中文信息学报,2016,30(05):176-186.

[7] 王勇,肖诗斌,郭跑秀,吕学强.中文微博突发事件检测研究.现代图书情报技术,2013(2):57-62.

[8] 费绍栋,杨玉珍,刘培玉,王健.融合情感过滤的突发事件检测方法.计算机应用,2015,35(5):1320-1323.

[9] 陈国兰.基于爆发词识别的微博突发事件监测方法研究. 情报杂志, 2014(9):123－128.

[10] 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. 计算机应用, 2014, 34 (2) : 486 －490, 505.

[11]熊宇,张一飞,冯时,王大玲.基于多模态特征深度融合的微博流事件检测与跟踪.控制与决策,2019,34(7):1409-1416.

[13]郑斐然,苗夺谦,张志飞,高灿.一种中文微博新闻话题检测的方法.计算机科学,2012,39(1):138-141.

[19]郭磊,李弼程,赵军磊.基于主题词向量聚类的话题内新事件检测. 中文信息学报, 2019, 33(6): 64-71, 79.

[20]王凯,洪宇,邱盈盈,姚建民,周国栋.融合上下文依赖和句子语义的事件线索检测研究.计算机科学与探索,2018,12(3):423-431.

[21]代翔,黄细凤,唐瑞,蒋梦婷,陈兴蜀,王海舟,罗梁.基于层次聚类的子话题检测算法.华南理工大学学报(自然科学版),2019,47(8):84-95.

[24] 周刚,邹鸿程,熊小兵,黄永忠.MB-SinglePass:基于组合相似度的微博话题检测.计算机科学,2012,39(10):198-202.

[25]邱云飞,程亮.微博突发话题检测方法研究.计算机工程,2012,38(9):288-290.

[26] 李莹莹,马帅,蒋浩谊,刘喆,胡春明,李雄.一种基于社交事件关联的故事脉络生成方法.计算机研究与发展,2018,55(9):1972-1986.

[27]张婧丽,周文瑄,洪宇,姚建民,周国栋,朱巧明.基于框架语义扩展训练集的有监督事件检测方法.中文信息学报,2019,33(5):82-92+131.

[28]陆垚杰,林鸿宇,韩先培,孙乐.基于语言学扰动的事件检测数据增强方法.中文信息学报,2019,33(7):110-117.

[32]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.

[33]仲兆满,管燕,李存华,刘宗田.微博网络地域 Top-k 突发事件检测.计算机学报,2018,41(7):1504-1516.

[36]张仰森,段宇翔,王建,吴云芳.基于多种词特征的微博突发事件检测方法.电子学报,2019,47(9):1919-1928.

[44]赵海林. 基于用户行为的推特事件检测方法研究[D].电子科技大学,2018.

[45]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.