基于词相关性特征的多归属谱聚类突发事件检测

蒋伟进[1,2,3,4]，王扬[1,2]，刘晓亮[2,3]，吕斯健[2,3]

（1. 湖南工商大学大数据与互联网创新研究院，湖南 长沙 410205；2. 新零售虚拟现实技术湖南省重点实验室，湖南 长沙 410205；
3. 湖南工商大学计算机与信息工程学院，湖南 长沙 410205；4. 武汉理工大学计算机科学与技术学院，湖北 武汉 430073）

摘 要：针对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，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措施进行处理，使突发事件的负面影响被尽量降低，维护社会的安定。仿真结果表明，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，所提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，而且能准确地检测出事件的相关信息。

关键词：突发事件；单词关系图；多归属谱聚类；检测

中图分类号：TP393

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2020215

Multi-attribute spectral clustering emergency detection based on word correlation feature

JIANG Weijin[1,2,3,4], WANG Yang[1,2], LIU Xiaoliang[2,3], LYU Sijian[2,3]

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
3. College of Computer and Information Engineering, Hunan University of Technology and Business, Changsha 410205, China
4. School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430073, China

Abstract: For current methods for extracting emergencies had problems of low accuracy and low efficiency, an emergency detection method based on the characteristics of word correlation was proposed, which could quickly detect emergency events from the social network data stream, so that relevant decision makers could 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. The simulation results show 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 proposed method can meet the needs of emergency detection. Not only the detailed information of the sub-events can be detected, but also the related information of the events can be accurately detected.

Key words: emergency, word relationship graph, multi-attribute spectral clustering, detection

---

收稿日期：2020−07−13；**修回日期**：2020−10−24

通信作者：王扬，wangyangwy25@163.com

基金项目：国家自然科学基金资助项目（No.61472136, No.61772196）；湖南省自然科学基金资助项目（No.2020JJ4249）；湖南省社会科学基金重点资助项目（No.2016ZDB006）；湖南省社会科学成果评审委员会课题重点基金资助项目（湘社评19ZD1005）；湖南省学位与研究生教育改革研究基金资助项目（No.2020JGYB234）；湖南省研究生科研创新资助项目（No.CX20201074）

Foundation Items: The National Natural Science Foundation of China (No.61472136, No.61772196), The Natural Science Foundation of Hunan Provincial (No.2020JJ4249), The Key Social Science Fund of Hunan Provincial (No.2016ZDB006), The Key Social Science Achievement Review Committee of Hunan Provincial (XSP No.19ZD1005), The Degree and Graduate Education Reform Research Project of Hunan Provincial (No.2020JGYB234), Postgraduate Scientific Research Innovation Project of Hunan Province (No.CX20201074)

1 引言

随着 Web 2.0 的发展，一系列新的社交网络正在迅速兴起。尽管此类网络相对较新，但它们吸引了很多用户来分享其观点和感受，在社交网络上实时讨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，并且他们对事情发表带有主观性、影响力较强的评论，使现实生活中的突发事件在社交虚拟网络上爆发的时间往往比官方发布新闻的时间更早[1]。具有用户发布内容的社交媒体和在线服务已经生成了数量惊人的信息，这些信息在事件分析和应急管理等各个领域都有潜在的应用[2]。突发事件在微博和微信等社交网络上迅速发酵[3-4]，其突发性影响了后续的应急处理，包括舆论以及救援等。通过从紧急灾难等事件检测模型发出大量及时、准确的警报，可以帮助人们迅速采取行动，以减轻损失。因此，在各种突发事件发生后，通过社交网络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意义。

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，可以通过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，因为它们可以带来宝贵的信息。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，以回应用户的投诉并改善决策。与传统的信息传播渠道相比，在社交网络上检测获得的突发事件能使人更快地了解到事件的详细发展情况，以便相关部门迅速采取应对策略，这具有重要的现实意义。本文围绕微博突发词提取及多归属谱聚类检测 2 个核心内容，开展了微博社交网络突发事件检测的研究，主要创新点如下。1) 在突发词提取上，根据微博的时空特点，在综合考虑博文内容及社交关系的基础上，利用词频增长率特征、用户影响力及词权重 3 类指标，提出了新颖的突发词提取模型；2) 在突发事件检测上，针对突发事件检测中参数过多的问题，将文本处理转化为图划分，从特征关系图的角度出发，基于事件突发特征的相似性和共现性构建词关系图，对突发事件进行检测。

2 相关研究

由于本文结合文本和词相关性特征来检测突发事件，因此相关工作集中在文本分析、突发特征分析以及用户特征分析等用于突发事件检测的方法。当前的核心问题和挑战是如何快速、准确地从指数增长的数据中检测到突发事件。现有的突发事件检测方法主要分为 3 类。

1) 以文本为中心。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[6-8]。李莹莹等[9]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。张婧丽等[10]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。陆垚杰等[11]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 2 个角度进行研究，使突发事件的检测模型更具准确性。Zhu 等[12]提出了一种改进的术语频率逆文档频率（TF-IDF, term frequency inverse document frequency）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。但是，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时，引入过多无用信息会对其造成噪声干扰。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，并且阈值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。

2) 以突发特征为中心。这类方法首先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，最后根据聚类算法的结果获取突发事件的相关信息。张鲁民等[13]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著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[14]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，但是该方法检测不到没有地域突发特征的内容。Kalden[15]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，增加了噪声信息。Zou 等[16]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，但在没有任何标签的情况下，这种方法将失败。张仰森等[17]提出了基于最小代价函数的目标检测与跟踪融合算法对突发事件进行检测，以降低检测的错误率。该算法能够自适应地调整跟踪参数的大小，并在丢失目标后重新捕获目标，它可以同时满足多个事件的检测跟踪。Zhang 等[18]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（BTDF, bursty term detection and filtration）方法，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进行过滤。

3) 以用户行为特征为中心。对用户在社交网络的行为数据进行分析，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[19]对 10 350 条独特的推特信息进行了特征分析，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 等[20]研究用户转发行为，提出了一种基于多层个人信息（MII, multi-layered individual information）和动态时间序列（DTS, dynamic time series）算法的用于谣言事件检测的新型两层门控循环单元（GRU, gated recurrent unit）模型，称为 MII-DTS-GRU。在新浪微博数据集上的实验结果表明，MII-DTS-GRU 模型达到了 96.3%的高精度。赵海林[21]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。介飞等[22]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词功能绑定，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。

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，按时间顺序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。基于突发事件的检测模型流程如图 1 所示。

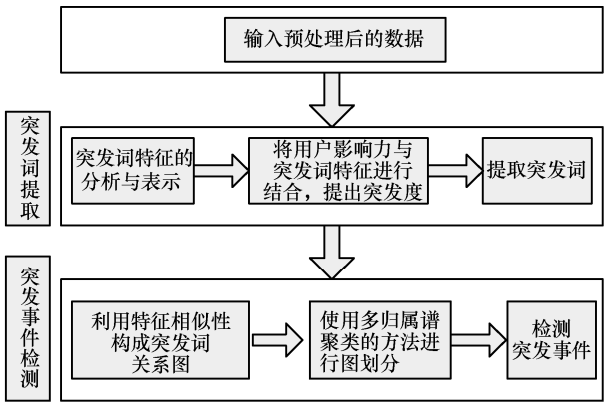


图 1　基于突发事件的检测模型流程

3　基于突发词相关突发特征提取模型

3.1　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。文本预处理首先进行噪声过滤，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[23-24]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、表情符号等。其次使用 BosonNLP 情感词典[25-26]过滤掉含情感的词语，如式(1)所示。最后对文本进

行规范。

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

其中，Se(n)为词语的情感度，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。

3.2 突发词特征的分析与表示

1) 词频增长率特征

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_m 的频率与之前的平均历史频率 $A_{m-1}(\omega)$ 之和。

$$A_m(\omega)=A_{m-1}(\omega)+\frac{f_m(\omega)-A_{m-1}(\omega)}{m} \tag{2}$$

其中，$f_m(\omega)$ 表示词 ω 在时间窗 T_m 下的词频。根据式(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) 用户影响力

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，这会使计算出的突发度不够准确，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 H=(Rep, Com, Fan, Type, Update)，如式(3)所示。

$$B_H=\frac{(\mathrm{Rep}_H+\mathrm{Com}_H)\mathrm{Fan}_H\mathrm{Type}_H}{\mathrm{Update}_H} \tag{3}$$

其中，Rep 和 Com 分别表示用户一个月之内转发和评论微博数量；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，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，普通用户的微博权重为 0.5；Update 表示用户一个月之内的更博数，最小值不能为零。

在社交网络上，用户的粉丝数量越多，影响力越大，如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) 词权重的计算

在突发事件中，与事件有关的微博会呈井喷式爆发，突发词会频繁地出现在同一事件的不同文本中[26]。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[27]中的文档频率−倒文档频率（DF-IDF, document frequency-inverted document frequency）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，能准确地计算词权重。

$$W_{j,t}=\mathrm{dfidf}_j^{t,\tau}=\mathrm{df}_j^t\log\left(1+\frac{1}{\mathrm{df}_j^{t,\tau}}\right) \tag{4}$$

式(4)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−τ~t 天内单词 j 的平均 DF，其 DF=$\frac{|Y_j^t|}{Y^t}$，Y_j^t 表示第 t 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。

3.3 突发度计算方法

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，突发度的计算式为

$$\mathrm{word}_{j,t}=\frac{1}{N}\sum_{k=t-N}^{t-1}\Big(W_{j,t}A_t(\omega)\cdot\sum_{P_n\in aP_{j,t}}\mathrm{lb}(B_{P_n})-W_{j,k}A_k(\omega)\sum_{P_n\in aP_{j,k}}\mathrm{lb}(B_{P_b})\Big) \tag{5}$$

其中，$\mathrm{word}_{j,t}$ 是单词 j 在时间窗 t 内的突发度；B_{P_b} 是包含单词 j 的一条微博的发布者 p_n 的影响力；$P_{j,t}$ 是在时间窗 t 内包含单词 j 的所有微博；N 是时间窗的总数。突发度值高的被提取为突发词。

4 突发事件检测

4.1 词语关系图构建

为迅速获取每日事件的信息，需要选取用于构建关系图的突发词集合，利用突发词集合构建词语关系图。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 是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著变化。边缘节点 g_i 在时间 t_s 边缘权重定义为 $R=(g_i,t_s)$。

给定 2 个词语矩阵 $\boldsymbol{\omega}_i$ 和 $\boldsymbol{\omega}_j$，通过余弦距离定义它们之间的语义相似性为

$$\mathrm{Sim}\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)=\boldsymbol{v}_{\boldsymbol{\omega}_i}\boldsymbol{v}_{\boldsymbol{\omega}_j} \tag{7}$$

其中，$\boldsymbol{v}_{\boldsymbol{\omega}}$ 是从 word2vec 模型计算出的单词的单位向量。

归一化将具有表达式的维数转换为无量纲的表达式后，$\boldsymbol{\omega}$ 将成为标量，可将计算量简化。归一化交叉相似度 $\boldsymbol{D}_{\mathrm{cc}}\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)$ 定义如式(8)所示，其中 $\boldsymbol{S}_{\boldsymbol{w}_i}$ 表示单词 $\boldsymbol{\omega}_i$ 的矩阵形式。

$$\boldsymbol{D}_{\mathrm{cc}}\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)=\frac{\boldsymbol{S}_{\boldsymbol{\omega}_i}^{\mathrm{T}}\boldsymbol{S}_{\boldsymbol{\omega}_j}^{\mathrm{T}}}{\sqrt{\boldsymbol{S}_{\boldsymbol{\omega}_i}^{\mathrm{T}}\boldsymbol{S}_{\boldsymbol{\omega}_i}^{\mathrm{T}}}\sqrt{\boldsymbol{S}_{\boldsymbol{\omega}_j}^{\mathrm{T}}\boldsymbol{S}_{\boldsymbol{\omega}_j}^{\mathrm{T}}}} \tag{8}$$

通过式(8)计算，得到词语关系图的相似矩阵，且维度为 n（单词 $\boldsymbol{\omega}_i$ 和 $\boldsymbol{\omega}_j$ 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，遍历词语关系图上的每个节点，如果相似度超过阈值 $\theta_{\mathrm{sim}}^{\mathrm{mc}}$，则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，本文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。

$$C\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)=2^{-\lambda}\left(\frac{f\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)}{f(\boldsymbol{\omega}_i)}+\frac{f\left(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j\right)}{f(\boldsymbol{\omega}_j)}\right) \tag{9}$$

其中，$f(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j)$ 表示单词 $\boldsymbol{\omega}_i$ 和 $\boldsymbol{\omega}_j$ 在某时间段内微博文本中同时出现的次数，$f(\boldsymbol{\omega}_i)$ 表示词语 $\boldsymbol{\omega}_i$ 和 $\boldsymbol{\omega}_j$ 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。

4.2　基于多归属谱聚类的图划分算法（MASCA, multi-attribute spectral clustering algorithm）

谱聚类算法从数据的亲和力矩阵（即相似性矩阵）得出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。多归属谱聚类的图划分流程如图 2 所示。

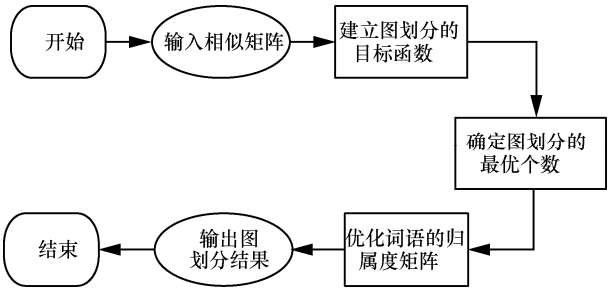


图 2　多归属谱聚类的图划分流程

1) 目标函数建立

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，定义为

$$\boldsymbol{u}_r=\left[\boldsymbol{u}_{1,r},\boldsymbol{u}_{2,r},\boldsymbol{u}_{i,r},\cdots,\boldsymbol{u}_{L,r}\right] \tag{10}$$

其中，$\boldsymbol{u}_{i,r}$ 表示单词 $\boldsymbol{\omega}_i$ 对第 r 个子图的归属程度，$0\leqslant\boldsymbol{u}_{i,r}\leqslant1$，$L$ 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[28]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 如式(11)所示。式(11)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 $\boldsymbol{u}_r$，以免 cut 出一个单独的词语。$\boldsymbol{W}$ 是词语关系图顶点之间的相似度矩阵，$\boldsymbol{D}$ 是相应的度矩阵。

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

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

$$\boldsymbol{U}=\left[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_M\right]$$

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

其中，$\widetilde{\boldsymbol{U}}=\boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U})^{-\frac{1}{2}}$，矩阵 $\boldsymbol{U}_e$ 包含拉普拉斯矩阵 $\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ 的 M 个最大的特征值对应的特征列向量，维度为 $L\times M$。

2) 归属度矩阵近似优化

向量矩阵 $\boldsymbol{U}_e$ 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 $\boldsymbol{U}_{\mathrm{opt}}$，即 $\boldsymbol{U}_{\mathrm{opt}}=\boldsymbol{U}_e\boldsymbol{R}$，其中 $\boldsymbol{R}$ 为旋转矩阵，属于单位正交矩阵。由于在连续域空间中优化 $\boldsymbol{U}_{\mathrm{opt}}$ 无法得到最优结果，属于 NP 难问题，因此本文运用近似方法在离散域中对其优化以期得到最好的结果，近似矩阵 $\boldsymbol{U}^a=\left[\boldsymbol{u}_1^a,\boldsymbol{u}_2^a,\cdots,\boldsymbol{u}_M^a\right]$。

近似方法通过衡量近似矩阵 $\boldsymbol{U}^a$ 与最优归属度矩阵 $\boldsymbol{U}_{\mathrm{opt}}$ 的误差进行优化，即在约束条件下如何使误差最小的问题。$\boldsymbol{U}^a$ 与 $\boldsymbol{U}_{\mathrm{opt}}$ 通过弗罗贝尼乌斯范数（Frobenius norm）进行表示，即

$$\min\widetilde{\boldsymbol{R}}=\arg\min\left\|\boldsymbol{U}^a-\boldsymbol{U}^e\boldsymbol{R}\right\|$$

$$\text{s.t. }\boldsymbol{R}^{\mathrm{T}}=\boldsymbol{R}^{-1} \tag{13}$$

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

其中，$(\boldsymbol{\Pi},\boldsymbol{\Omega},\boldsymbol{\Xi})$ 是矩阵 $\boldsymbol{U}^{a^{\mathrm{T}}}$、$\boldsymbol{U}^e$ 的奇异值分解矩阵，矩阵 $\boldsymbol{\Pi}$ 和 $\boldsymbol{\Xi}$ 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 所示。

算法 1 优化归属矩阵

输入 $n,m,\boldsymbol{U}$

输出 $\boldsymbol{U}_{\mathrm{opt}}$

① $\boldsymbol{R}\leftarrow\mathrm{Ortho}(\boldsymbol{U}[:m,:m])$

② value $\leftarrow$ 1 000.0

③ for $k\leftarrow[1,10]$ do

④ $\quad\boldsymbol{U}^I\leftarrow\boldsymbol{U}[:,:m]\boldsymbol{R}$

⑤ $\quad\boldsymbol{U}^a\leftarrow\mathrm{argmin}(\mathrm{Frobenius}(\boldsymbol{U}^a-\boldsymbol{U}^I))$

⑥ $\quad\boldsymbol{U}_{\mathrm{svd}},s,\boldsymbol{V}_{\mathrm{svd}}\leftarrow\mathrm{SVD}(\boldsymbol{U}^{a^{\mathrm{T}}},\boldsymbol{U}^I)$

⑦ $\quad\boldsymbol{R}\leftarrow\boldsymbol{V}_{\mathrm{svd}}^{\mathrm{T}}\boldsymbol{U}_{\mathrm{svd}}$

⑧ $\quad\boldsymbol{v}\leftarrow\mathrm{Frobenius}(\boldsymbol{U}^a-\boldsymbol{U}^I)$

⑨ $\quad$if $\boldsymbol{v}<$ value then

⑩ $\quad\quad$value $\leftarrow\boldsymbol{v}$

⑪ $\quad\quad\boldsymbol{U}_{\mathrm{opt}}\leftarrow\boldsymbol{U}^a$

⑫ $\quad$end if

⑬ end for

⑭return $\boldsymbol{U}_{\mathrm{opt}}$

3) 聚类个数自适应方法

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，并进一步精确该值。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。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 $\boldsymbol{D}$，$\boldsymbol{W}$

输出 $\boldsymbol{M}_{\mathrm{opt}}$

① $\boldsymbol{M}_{\mathrm{opt}}\leftarrow 0$

② $\boldsymbol{U}\leftarrow\mathrm{Eigenvalue\_Decomposition}(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{\frac{1}{2}})$

③ $\boldsymbol{U}_{\mathrm{diff}}\leftarrow\mathrm{Diff}(\boldsymbol{U})$

④ $\boldsymbol{M}_{\mathrm{opt}}\leftarrow\mathrm{Index}(\boldsymbol{U}_{\mathrm{diff}})$

⑤ return $\boldsymbol{M}_{\mathrm{opt}}$

4) 突发事件识别

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改变代表突发事件产生，伪代码如算法 3 所示。

算法 3　判定突发事件

输入　Graph G, Graph S, μ

输出　true/false

① Get events set μ by searching event index;
② if cos(s Graph G, Graph S) $< \mu$ then
③ return true
④ end if
⑤ 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

① candidate_cluster ←[]
② for word in subgraph do
③ for subcluster in cluster do
④ if word in subcluster then
⑤ candidate _cluster ←subcluster
⑥ end if
⑦ end for
⑧ end for
⑨ new_cluster ← Sort($\text{candidate}_{\text{cluster}}$)
⑩ for subcluster in new_cluster do
⑪ if Similarity(subgraph, subcluster)>sim then
⑫ sim←Similarity(subgraph, subcluster)
⑬ eventcluster ← subcluster
⑭ end if
⑮ end for
⑯return eventcluster

5　实验结果与分析

本文使用的数据集来自新浪微博，通过模拟微博登录来爬取微博数据，采集了 2019 年 11 月 1 日至 11 月 30 日的微博数据，这些数据没有进行事件标注。由于微博不仅包含官方新闻事件，也包含娱乐新闻事件[29-31]，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 3.1 节方法进行文本预处理。所有实验均在具有 8 GB 内存并在 Windows 8 上运行的 4.00 GHz Intel CPU 上进行。本文实现了该算法，以获取准确的突发事件并验证检测是否成功。

5.1　突发词提取

鉴于微博数据中存在的大量噪声，本文对数据进行噪声过滤以及情感过滤，经处理后的微博存储结构如表 1 所示。

表 1　处理后的微博存储结构

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

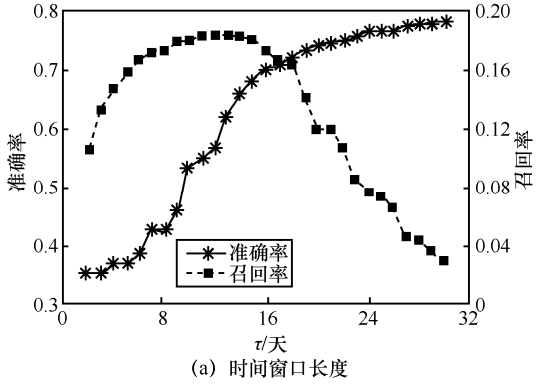
为了测试突发词提取模型的效果，从数据库中抽取 2019 年 11 月 20 日到 2019 年 11 月 30 日的数据。首先分析时间窗口参数对突发事件检测结果的影响，如图 3(a)所示；然后分析提取突发词数量的多少是否会影响实验结果，如图 3(b)所示。

如图 3(a)所示，当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 2~14 时，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，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 3(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。

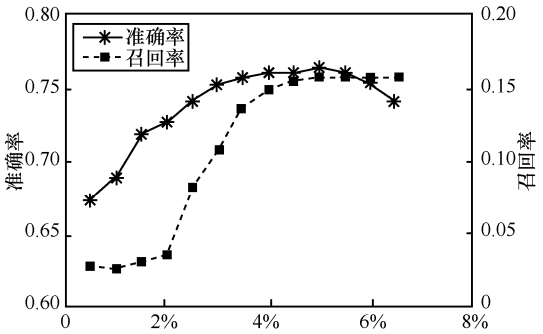
5.2　多归属谱聚类效果测试

1) 单词关系图参数测试

词关系图是进行谱聚类图划分的基础，据此可分析基于图聚类的事件检测效果。图 4 分析了关系图节点近邻数的大小对突发事件检测效果的影响。当节点近邻数较小时，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。



(a) 时间窗口长度



(b) 关键词数量

图 3　不同突发词提取参数对事件检测的影响

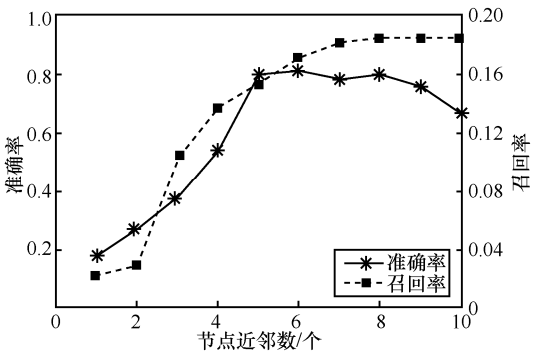


图 4　词关系图节点近邻数对事件检测性能的影响

图 5 显示了突发词相似度阈值的变化对突发事件检测性能的影响。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。

根据上述结果调好参数之后，选取突发度较高的 8 个单词按顺序构建单词关系图，8 个单词的关系网络如图 6 所示。实线表示 2 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。

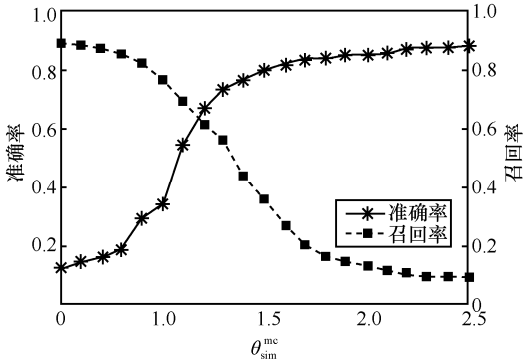


图 5　相似度阈值对事件检测性能的影响

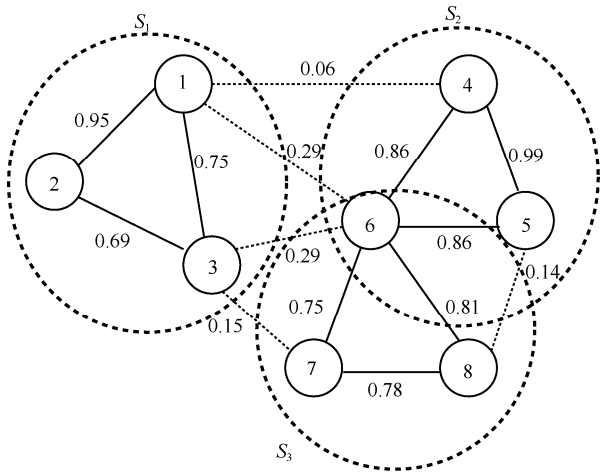


图 6　词关系图效果示意

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式得到了词的突发度，突发关键词的热度频率如图 7 所示，本文对 11 月的突发事件进行分析。在图 7 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词也具有此特征，如图 8 所示。事件 4 与突发词 1、2、3 相关，事件 2 与突发词 1、4 相关。这 2 个图揭示了关于不同事件的关键词彼此之间具有某些语义相关性，并且相互影响。

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（multi-attribute spectral clustering algorithm）对关系图进行划分，并且给出了图划分的最优个数为 7。

5.3　突发事件检测

表 2 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，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，当 μ 为 0.5、0.6、0.7、0.8、0.9 时，计算相对应的指标大小，并对其进行比较。

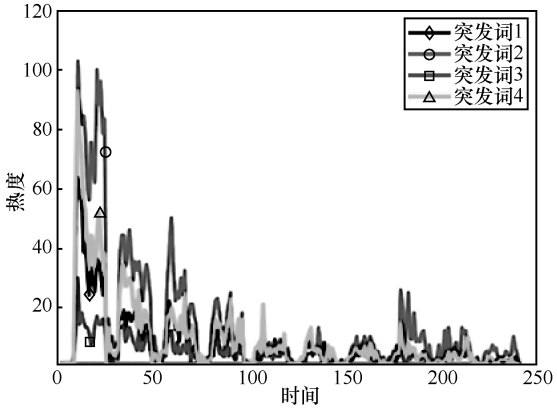


图 7　突发关键词的热度频率

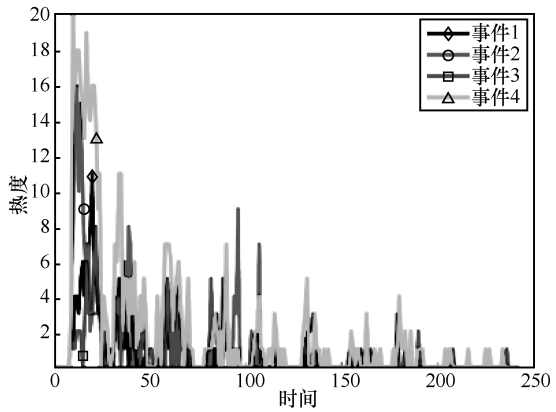


图 8　突发事件的热度频率

表 2　　阈值参数对实验结果的影响

阈值	Precision	Recall	*F*1
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%

Precision、Recall 和 *F*1 在不同相似度阈值参数 μ 下的变化趋势如图 9 所示。Precision 随着 μ 的增加而逐渐下降，0.7~0.9 下降幅度较大；与之相反，μ 越大，Recall 也随着增大，0.8~0.9 基本保持不变；而 *F*1 的变化趋势是先增大然后减小，在 μ=0.7 时，*F*1 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。

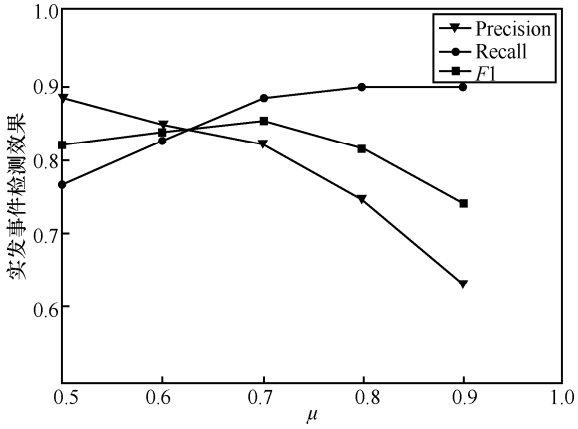


图 9　突发事件检测效果

在国内微博突发事件检测中，尚没有识别手动标记的语料库[32-35]。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 32 个事件。近一个月内社交网络上热议最多的 7 个突发事件在表 3 显示，包含了事件的基本信息，即事件编号、与事件相符的子图词语数量、单词重合率。

表 3　　部分突发事件检测结果

编号	子图单词数量	单词重合率
1	9	1.0
2	9	0.95
3	10	0.9
4	11	0.85
5	10	0.9
6	9	0.8
7	8	0.85

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，子图单词都能描述对应事件的发展经过，同时子图单词重合率平均值为 0.892 9，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地识别突发事件，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。

5.4 评价指标

本节将本文与其他文献的方法进行对比，使用标准指标 Precision、Recall 和 $F1$ 评估量化模型的有效性，计算式为

$$\text{Precision}=\frac{\text{Bcorrect}}{\text{Boutout}}$$

$$\text{Recall}=\frac{\text{Bcorrect}}{\text{Bnumber}}$$

$$F1=\frac{2\text{PrecisionRecall}}{\text{Precision+Recall}}$$

其中，Bcorrect 为系统中识别正确的突发事件个数，Bnumber 为数据集中事件的总数量，Boutout 为数据集手动标注的突发事件个数。

1) 指标对比

文献[29]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，当共现度阈值为 0.6 时 $F1$ 值最高，达到 0.661 5，Precision 是 0.645 4，Recall 是 0.77。文献[20]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为 0.8 时，$F1$ 值最高，达到 0.766 8，Precision 是 0.736 4，Recall 是 0.805 0。将其与本文方法的 Precision、Recall、$F1$ 值相比较，如图 10 所示。

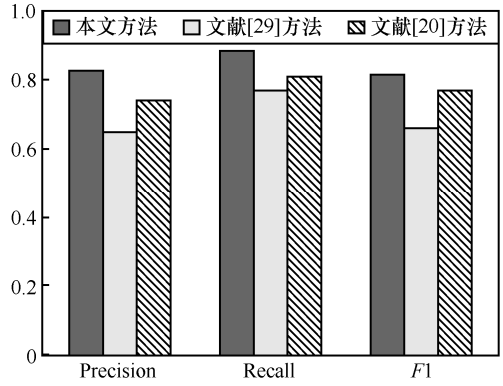


图 10 实验结果对比

由图 10 可知，本文方法在 Precision、Recall 与 $F1$ 值上都要优于其他 2 种方法，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

检测时延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[36-38]。本文选择 30 个通过给定 5 种方法成功检测到的事件。图 11 显示了突发事件检测时延对比。在所有方法中，本文方法花费最少的时间进行事件检测。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，能使相关人员及时采取措施进行控制。

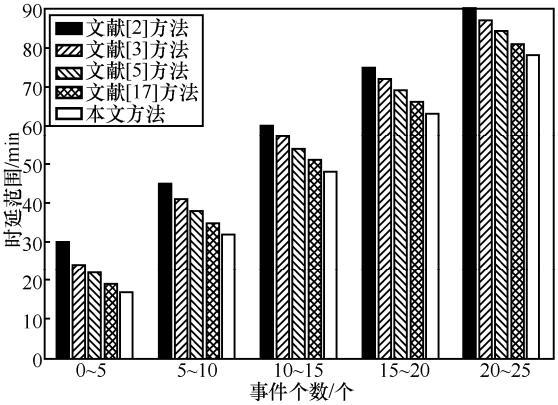


图 11 突发事件检测时延对比

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，检查了真实数据后发现，关系图中最早和最新的事件不一定彼此相似。但是其他方法将它们视为无关事件，因为它没有达到阈值。本文方法获得了由最相似事件之前已经构造的旧关系图，并将本文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。

6 结束语

本文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，本文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程度。实验结果表明，本

文方法具有很高的性能和有效性。为了提高性能，本文对检测模型的相关参数进行调整，得到了最优检测性能，当 μ=0.7 时，Precision、Recall 与 $F1$ 值都有良好的效果，本文方法在精度、召回率和时延方面均优于其他对比方法。

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。

参考文献：

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

[2] 闻佳, 王宏君, 邓佳, 等. 基于深度学习的异常事件检测[J]. 电子学报, 2020, 48(2): 308-313.
WEN J, WANG H J, DENG J, et al. Abnormal event detection based on deep learning [J]. Chinese Journal of Electronics, 2020, 48(2): 308-313.

[3] 胡文斌, 王欢, 严丽平, 等. 面向节点演化波动的社会网络事件检测方法[J]. 软件学报, 2017, 28(10): 2693-2703.
HU W B, WANG H, YAN L P, et al. Social network event detection method for node evolution and fluctuation[J]. Journal of Software, 2017, 28(10): 2693-2703.

[4] MU L, JIN P, ZHENG L, et al. Lifecycle-based event detection from microblogs[C]//Companion Proceedings of the The Web Conference 2018. [S.n.:s.l.], 2018: 283-290.

[5] PRADHAN A K, MOHANTY H, LAL R P. Event detection and aspects in twitter: a bow approach[C]//International Conference on Distributed Computing and Internet Technology. Berlin: Springer. 2019: 194-211.

[6] GOTO J, MIYAZAKI T, TAKEI Y, et al. Automatictweet detection based on data specified through news production[C]//Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion. New York: ACM Press, 2018: 1.

[7] 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, 2017, 39(10): 198-202.
ZHOU G, ZOU H C, XION X B, et al. MB-SinglePass: microblog topic detection based on combined similarity[J]. Computer Science, 2017, 39(10): 198-202.

[8] QIU Y F, CHENG L B. Research onsudden topic detection method for microblog[J]. Computer Engineering, 2012, 38 (9): 288-290.

[9] 李莹莹, 马帅, 蒋浩谊, 等. 一种基于社交事件关联的故事脉络生成方法[J]. 计算机研究与发展, 2018, 55(9): 1972-1986.
LI Y Y, MA S, JIANG H Y, et al. A method for generating story context based on social event correlation[J]. Computer Research and Development, 2018, 55(9): 1972-1986.

[10] 张婧丽, 周文瑄, 洪宇, 等. 基于框架语义扩展训练集的有监督事件检测方法[J]. 中文信息学报, 2019, 33(5): 82-92, 131.
ZHANG J L, ZHOU W X, HONG Y, et al. Supervised event detection method based on frame semantic extension training set[J]. Chinese Journal of Information, 2019, 33(5): 82-92, 131.

[11] 陆垚杰, 林鸿宇, 韩先培, 等. 基于语言学扰动的事件检测数据增强方法[J]. 中文信息学报, 2019, 33(7): 110-117.
LU Y J, LIN H Y, HAN X P, et al. Incident detection data enhancement method based on linguistic disturbance [J]. Chinese Journal of Information, 2019, 33(7): 110-117.

[12] ZHU Z, LIANG J, LI D, et al. Hot topic detection based on a refined TF-IDF algorithm[J]. IEEE Access, 2019, 7: 26996-27007.

[13] 张鲁民, 贾焰, 周斌, 等. 一种基于情感符号的在线突发事件检测方法[J]. 计算机学报, 2018, 36(8): 1659-1667.
ZHANG L M, JIA Y, ZHOU B, et al. An online emergency detection method based on emotional symbols[J]. Chinese Journal of Computers, 2018, 36(8): 1659-1667.

[14] 仲兆满, 管燕, 李存华, 等. 微博网络地域 Top-k 突发事件检测[J]. 计算机学报, 2018, 41(7): 1504-1516.
ZHONG Z M, GUAN Y, LI C H, et al. Detection of Top-k emergencies in Weibo networks[J]. Chinese Journal of Computers, 2018, 41(7): 1504-1516.

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

[16] ZOU X M, YANG J, ZHANG J P. Sentiment-based and hashtag-based Chinese online bursty event detection[J]. Multimedia Tools and Applications, 2018, 77(16): 21725-21750.

[17] 张仰森, 段宇翔, 王建, 等. 基于多种词特征的微博突发事件检测方法[J]. 电子学报, 2019, 47(9): 1919-1928.
ZHANG Y S, DUAN Y X, WANG J, et al. Microblog emergency detection method based on multiple word features[J]. Chinese Journal of Electronics, 2019, 47(9): 1919-1928.

[18] ZHANG Q, DU J, KOU F, et al. Bursty topic detection based on bursty term detection and filtration[C]//Chinese Intelligent Systems Conference. Berlin: Springer, 2019: 211-219.

[19] GUPTA A, LAMBA H, KUMARAGURU P, 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. [S.n.:s.l.], 2013: 729-736.

[20] WANG Z H, GUO Y. Empower rumor events detection from Chinese microblogs with multi-type individual information[J]. Knowledge and Information Systems, 2020, 62(2): 3585-3614.

[21] 赵海林. 基于用户行为的推特事件检测方法研究[D]. 成都: 电子科技大学, 2018.
ZHAO H L. Research on twitter incident detection method based on user behavior[D]. Chengdu: University of Electronic Science and Technology of China, 2018.

[22] 介飞, 谢飞, 李磊, 等. 社交网络中隐式事件突发性检测[J]. 自动化学报, 2018, 44(4): 730-742.
JIE F, XIE F, LI L, et al. Implicit incident detection in social networks [J]. Acta Automatica Sinica, 2018, 44(4): 730-742.

[23] 姚子瑜, 屠守中, 黄民烈, 等. 一种半监督的中文垃圾微博过滤方法[J]. 中文信息学报, 2016, 30(5): 176-186.
YAO Z Y, TA S Z, HUANG M L, et al. A semi-supervised Chinese junk microblog filtering method[J]. Chinese Information Processing, 2016, 30(5): 176-186.

[24] 王勇, 肖诗斌, 郭跬秀, 等. 中文微博突发事件检测研究[J]. 现代图书情报技术, 2018(2): 57-62.

WANG Y, XIAO S B, GUO J X, et al. Research on Chinese Weibo emergencies detection[J]. Modern Library and Information Technology, 2018(2): 57-62.

[25] 费绍栋, 杨玉珍, 刘培玉, 等. 融合情感过滤的突发事件检测方法[J]. 计算机应用, 2015, 35(5): 1320-1323.
FEI S D, YANG Y Z, LIU P Y, et al. Emergent incident detection method fused with emotion filtering[J]. Journal of Computer Applications, 2015, 35(5): 1320-1323.

[26] 陈国兰. 基于爆发词识别的微博突发事件监测方法研究[J]. 情报杂志, 2014(9): 123-128.
CHEN G L. Research on Weibo emergencies monitoring method based on outbreak word recognition[J]. Journal of Information, 2014(9): 123-128.

[27] 郭跽秀, 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(2): 486-490, 505.
GUO J X, LYU X Q, LI Z. Microblog emergency detection method based on sudden word clustering[J]. Journal of Computer Applications, 2014, 34(2): 486-490, 505.

[28] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000, 22(8): 888-905.

[29] ZHEN F R, MIAO D Q, ZHANG Z F, et al. News topic detection approach on Chinese microblog[J]. Computer science, 2018, 39(1): 138-141.

[30] 熊宇, 张一飞, 冯时, 等. 基于多模态特征深度融合的微博流事件检测与跟踪[J]. 控制与决策, 2019, 34(7): 1409-1416.
XIONG Y, ZHANG Y F, FENG S, et al. Microblog stream event detection and tracking based on deep fusion of multi-modal features[J]. Control and Decision, 2019, 34(7): 1409-1416.

[31] 吴国华, 龚礼春, 袁理锋, 等. 中文文本信息隐藏研究进展[J]. 通信学报, 2019, 40(9): 145-156.
WU G H, GONG L H, YUAN L F, et al. Research progress on information hiding in Chinese text[J]. Journal on Communications, 2019, 40(9): 145-156.

[32] HU L, YU S, WU B, et al. A neural model for joint event detection and prediction[J]. Neurocomputing, 2020, 407: 376-384.

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

[34] WANG Z H, GUO Y. Rumor events detection enhanced by encoding sentimental information into time series division and word representations[J]. Neurocomputing, 2020, 397:224-243.

[35] 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[J]. 中文信息学报, 2019, 33(6): 64-71, 79.
GUO L, LI B C, ZHAO J L. New event detection within topic based on topic word vector clustering[J]. Journal of Chinese Information Processing, 2019, 33(6): 64-71, 79.

[36] 王凯, 洪宇, 邱盈盈, 等. 融合上下文依赖和句子语义的事件线索检测研究[J]. 计算机科学与探索, 2018, 12(3): 423-431.
WANG K, HONG Y, QIU Y Y, et al. Research on event clue detection combining context dependence and sentence semantics[J]. Journal of Computer Science and Exploration, 2018, 12(3): 423-431.

[37] 代翔, 黄细凤, 唐瑞, 等. 基于层次聚类的子话题检测算法[J]. 华南理工大学学报(自然科学版), 2019, 47(8): 84-95.
DAI X, HUANG X F, TANG R, et al. Subtopic detection algorithm based on hierarchical clustering[J]. Journal of South China University of Technology (Natural Science Edition), 2019, 47(8): 84-95.

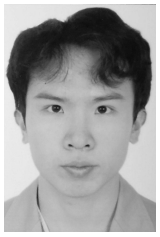
[38] JANANI R, VIJAYARANI S. Text document clustering using spectral clustering algorithm with particle swarm optimization[J]. Expert Systems with Applications, 2019, 134: 192-200.

[作者简介]

蒋伟进（1967− ），男，湖南益阳人，博士，湖南工商大学教授、硕士生导师，主要研究方向为新一代分布式人工智能、软件定义物联网、社会计算、云计算与网络系统安全。

王扬（1996− ），女，湖南衡阳人，湖南工商大学硕士生，主要研究方向为社会计算、应急管理、复杂系统建模和仿真、大数据技术。

刘晓亮（1996− ），男，安徽阜阳人，湖南工商大学硕士生，主要研究方向为群智感知、移动计算、软件定义网络。



吕斯健（1996− ），男，广东深圳人，湖南工商大学硕士生，主要研究方向为移动群智感知、互信计算。