

公共服务热线中基于地域自适应的 突发事件实时检测方法

麦丞程 陈玉婷 仇学明 刘 健 赵 博 袁春风 黄宜华

(南京大学计算机软件新技术国家重点实验室 南京 210023)

(南京大学计算机科学与技术系 南京 210023)

摘 要 随着信息技术的普及应用,城市公共服务热线平台累积了大量亟待分析的民生诉求数据.传统事件检测方法缺少对于地域模式的考虑,同时,其所依赖的GPS地理信息也不易获得.因此,难以直接运用现有的突发事件检测方法挖掘公共服务热线中潜在的民生突发事件.为此,本文提出了一种基于地域自适应的突发事件实时检测方法(RAEDetection).首先,提出一种基于增量式Kleinberg模型的突发词识别算法,克服了现有批处理式Kleinberg模型的局限性,可从流式增量数据中实时识别突发词汇;然后,提出一种基于分层语义分析的候选突发事件识别算法,以突发词为线索,先根据突发词的主题层语义信息确定突发主题事件,再根据诉求记录的事件层语义信息将每个突发主题事件进一步细分为多个候选突发事件;最后,提出一种基于事件地域树的地域模式自适应识别算法,通过构建包含市级、区级、街道级三层结构的事件地域树,并通过基于KL距离的事件地域分布检验与优化,自适应地识别不同事件发生的地域模式,过滤候选突发事件中的噪声数据,得到最终的突发事件.在城市公共服务真实数据集以及Twitter数据集上的实验结果表明,与目前最新的方法对比,本文方法具有更高的检测准确率和更快的计算性能,能够有效地检测出数据流中的突发事件,算法具备良好的数据和系统可扩展性.本文方法已经成功落地应用于江苏省公共服务热线平台,提供高效的自动化和智能化突发事件检测服务.

关键词 事件检测;突发性分析;地域自适应;公共服务热线;数据挖掘

中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2020.02259

A Region-Adaptive Method for Real-Time Bursty Event Detection in Public Service Hotline

MAI Cheng-Cheng CHEN Yu-Ting QIU Xue-Ming LIU Jian ZHAO Bo

YUAN Chun-Feng HUANG Yi-Hua

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023)

(Department of Computer Science and Technology, Nanjing University, Nanjing 210023)

Abstract With the popularization of information technology, the civic public service platform has accumulated a large number of public livelihood complaint data that need to be analyzed. The traditional event detection methods do not take the regional patterns of events into consideration. Meanwhile, the GPS geographic information used by these methods is also not easy to obtain. Therefore, many studies are looking for efficient and accurate methods to deal with the problem of recognizing the region patterns of events. However, it is not efficient to use the existing event

收稿日期:2019-09-10;在线发布日期:2020-02-16. 本课题得到国家自然科学基金项目(61572250, U1811461)、江苏省科技支撑计划项目(BE2017155)、江苏省软件新技术与产业化协同创新中心资助. 麦丞程, 博士研究生, 中国计算机学会(CCF)学生会员, 主要研究方向为大规模文本分析与事件检测等. E-mail: maicc@smail.nju.edu.cn. 陈玉婷, 硕士研究生, 主要研究方向为并行计算与事件检测等. 仇学明, 硕士研究生, 主要研究方向为社会计算与大数据. 刘 健, 硕士研究生, 主要研究方向为机器学习、数据挖掘. 赵 博, 博士研究生, 主要研究方向为主题模型与深度学习等. 袁春风, 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为体系结构与并行计算、操作系统. 黄宜华(通信作者), 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为大数据、云计算、并行计算. E-mail: yhuang@nju.edu.cn.

detection methods to capture the potential events in the civic public service. In this paper, we propose a real-time region-adaptive method for bursty event detection, called RAEDetection. First, the recognition of bursty words from data stream is the basis of discovering the bursty events. The traditional Kleinberg model can only find these bursty words from the static data. Therefore, we propose an improved incremental Kleinberg model to identify the bursty words from the real-time data stream. Then, after obtaining the bursty words, we propose an algorithm based on hierarchical semantic analysis for recognizing the candidate bursty events. With bursty words as clues, this algorithm finds the topic bursty events with semantic information from topics and then divides these events into more fine-grained candidate bursty events with the semantic information from the complaint records. Finally, in order to filter out the noise records in the candidate events, the event region tree is constructed to recognize the regional patterns of events. The event region tree has a three-level structure corresponding to the addresses in the city, district and street level respectively. According to the maximum entropy principle, we assume that the address distribution of one certain event obeys the discrete uniform distribution. We use KL distance to compare the distance between the statistic address distribution and the assumed address distribution. We choose the number of addresses which can minimize the value of the KL distance to indicate the geographical regions of that event, so as to realize the adaptive recognition of regional patterns. The experimental results from two real-world datasets from civic public service and one social media dataset from Twitter show that our method outperforms the state-of-the-art methods for both detection accuracy and computing performance, with good data and system scalability. In the real application scenarios, compared with algorithm TrioVecEvent, GeoBurst, and TopicSketch, the pseudo F1 values of our algorithm RAEDetection are increased by 54.85%, 221.13%, and 84.26% on average, respectively. To further explore the influences of the size of the sliding window and the threshold value of semantic similarity on our method, we carried out the relevant experiments and find that the our RAEDetection achieves the best performance when the sliding window size is set to 40 minutes and the threshold value of semantic similarity is set to 0.5 and 0.6 on the Nanjing and Suzhou dataset respectively, which has an important guiding role for the practical application of our algorithm. Finally, the proposed method has been successfully adopted and validated by the civic public service platform of Jiangsu province.

Keywords event detection; burstiness analysis; region adaptation; public services hotlines; data mining

1 引 言

随着大数据相关治理理念、技术与支撑平台在各个垂直行业的落地与逐步成熟,城市公共服务热线平台作为政府部门及其各级成员单位的重要组成部分,已经累积了大量民生相关,覆盖教育、医疗、就业等多个领域的诉求记录.以江苏省为例,据统计,公共热线服务平台全年累计访问量超过 3.3 亿次,每年诉求工单记录增量超过 1500 万条,是聆听民意、解决民生问题的基础大数据平台^①.因此,从海量诉求记录中实时、准确地检测突发事件是发现民

生突发问题、助力政务服务的必要环节,具有重要的社会价值.

信息传播媒介的演进使得事件检测方法的发展日新月异.早期事件检测方法的数据来源主要是新闻报道.例如,文献[1]将新闻文档映射到向量空间模型中,并将相似度超过给定阈值的文档汇总为一个事件.若某个事件出现的频率超过其历史均值,则认为该事件是一个突发事件.文献[2]利用二项分布对新闻报道中各个词汇出现的频率进行建模,并将其中呈突发性增长趋势的词汇作为突发特征,然后

^① http://www.gov.cn/xinwen/2019-01/27/content_5361507.htm

将使得这些突发词汇出现概率最大化的新闻报道集合作为突发事件. 文献[3]则将新闻事件细分为周期事件与非周期事件,同时,采用傅里叶变换识别文档中词汇的突发周期.

受限于数据来源,这类事件检测方法的局限性在于:(1)新闻报道较长的时延导致事件检测的实时性不足;(2)对于文本长度更短、内容书写更加不规范的社交媒体中的突发事件检测效果有待提升.

2010年以后,社交媒体的兴起对突发事件检测提出了新的挑战.为了提高事件检测的实时性,文献[4]对 Twitter 数据流进行时间窗口的划分,并且假设每个时间窗口内 Tweet 中词汇的产生服从高斯分布.如果某个词汇出现的概率超过其期望值 2 倍标准差,则将该词汇识别为突发词特征,从而快速捕获潜在突发事件.类似地,文献[5]利用 z 分数来衡量 Tweet 记录中词汇的突发性,定位突发事件发生的时间窗口.

由于微博、Twitter 等自媒体的文本长度更短,内容书写也更加随意,基于传统文本聚类的事件检测方法面临挑战.文献[4]利用 Wikipedia 作为外部语料,对 Twitter 文本内容进行规范化与短语重组处理,提升文本质量. Zhang 等人发现微博文本流中情感符号也存在突发现象,提出一种基于情感符号的在线突发事件检测算法,扩展了短文本的语义信息^[6].

此外,移动智能设备提供的 GPS 定位数据对提高突发事件检测的准确率具有重要作用.文献[7]根据 Twitter 的 GPS 地理定位对事件发生的地址空间进行网格划分,将某个空间网格中呈爆发性增长趋势且语义相似的 Tweet 集合检测为一个事件.文献[8]则利用 Epanechnikov 核函数^[9]将地理位置相近且语义相似的 Tweet 记录进行聚合,过滤掉语义上相似却发生在不同地方的噪声数据.

近年来,社会信息化程度的日益提升促使公共服务热线平台亟需从海量诉求记录中发掘民生相关的突发事件.但是现有面向社交媒体的突发事件检测方法与之相比存在较大差异性,难以直接复用,相关的研究也很少.表 1 给出了公共服务热线平台中诉求记录样例.这两类突发事件检测对象的区别总结如下:

(1)在数据来源上,面向社交媒体的突发事件检测方法依赖微博、Twitter 等社交媒体提供的 GPS 坐标信息.文献[10-11]统计发现 Twitter 上大约只有 2%的 Tweet 数据带有 GPS 地理标签,更多

无 GPS 坐标定位的数据还有待开发利用,例如公共服务热线平台上的海量民生诉求数据;

(2)在数据采集上,由于 GPS 坐标信息采集难度大,现有突发事件检测方法难以直接应用在缺少精确 GPS 定位信息的公共热线、政务服务等多个汇集大量民生诉求的垂直行业中;

(3)在数据构成上,社交媒体数据中的地址信息为 GPS 坐标,而公共服务热线平台中地址信息则以文本地址为主.由于 GPS 坐标缺乏语义信息,对于某些发生在同一个地区或行政区划内的事件,其包含的各个数据记录的 GPS 坐标可能相距较远,这将导致发生范围涉及多个地域的广域事件被错误地拆分成多个局部事件,增高事件检测重复率,同时,也难以过滤掉局部事件中的噪声数据.

表 1 真实诉求工单记录样例

记录编号	99xxxx02571
时间	2018/03/01 19:13:15
诉求内容	“江宁区东山街道金盛路美食烧烤店,此处油烟直排主干道,油烟扰民.”
区域代码	025

为了解决上述问题,扩展突发事件检测研究的对象,弥补现有应用场景的空白,本文围绕突发词实时检测、候选突发事件聚类与突发事件地域模式自适应识别这三个核心内容,提出一种基于地域自适应的突发事件实时检测方法(RAEDetection).首次在公共服务领域,在诉求记录缺失 GPS 坐标的情况下,充分利用诉求记录中的文本地址进行事件地域分布模式的自适应识别,实现对于突发事件实时、准确的识别.

本文的主要创新点与贡献如下:

(1)提出一种基于增量式 Kleinberg 模型的突发词实时检测算法,克服了现有批处理式 Kleinberg 模型的局限性,实现对于流式增量数据的高效突发词实时检测分析;

(2)提出一种基于分层语义分析的候选突发事件识别算法.先根据突发词的主题层语义生成突发主题事件,再根据诉求记录的事件层语义对每个突发主题事件进行更细粒度的划分,形成候选突发事件,提高候选突发事件识别的精度;

(3)首次提出一种基于事件地域树的地域模式自适应识别算法,根据地域范围对候选突发事件进行拆分与噪声数据过滤,减低事件检测的重复率;

(4)应用城市公共服务热线真实数据集的实验结果,验证了本文提出方法的有效性;

(5) 相关研究成果已经成功落地应用于江苏省公共服务热线平台。

本文第 2 节介绍相关工作;第 3 节给出 RAE-Detection 方法的整体框架;第 4 节介绍基于增量式 Kleinberg 模型的突发词实时识别算法;第 5 节阐述候选突发事件识别算法;第 6 节给出突发事件地域模式自适应识别算法;第 7 节进行了实验与分析;第 8 节对本文进行了总结与展望。

2 相关工作

根据突发事件发生的地域范围不同,现有的突发事件检测方法可以分为基于广域模式与基于局部模式的突发事件检测方法。

呈局部模式爆发的事件主要集中在某一个局部地域范围中,例如某社区发生停水、某社区存在违建乱搭现象、某地因道路施工而发生拥堵等事件。文献[8]针对 Twitter 数据流,提出了一种基于 GPS 坐标的局部事件实时检测算法。文献[12]提出了一种基于多模态的嵌入模型 TrioVecEvent,该模型将 Twitter 数据的地理信息、时间信息与语义信息映射到同一个隐式向量空间中,然后采用基于混合贝叶斯模型的聚类算法,从 Twitter 数据流中检测突发事件。文献[13]通过统计某个地域范围内微博热词出现的频率、关联用户数量以及转发、阅读、评论等社交行为发生次数,进行突发词汇检测,并在突发词关联网络中采用层次聚类方法生成突发事件。文献[14]提出了一种可定制化的时空事件检测方法。用户可以根据自身关注的不同领域进行特定事件检测。先从数据中发现该事件,然后再确定其发生的地域范围。文献[15]则提出了一种地理时序模式挖掘算法,先利用聚类算法对微博进行聚类,然后,如果某个地域内某类微博的数量显著增加,则认为该地域发生突发事件。

这类方法的优势在于,能够提高某个局部地域范围内事件检测的准确率,减少噪声数据的引入。但是其局限性在于,会错误地把发生范围涉及多个不同局部地域的广域事件拆分为多个局部事件,增高事件检测的重复率,导致准确率下降。

呈广域模式爆发的事件大多涉及多个局部地域范围,例如台风、地震等灾害事件。文献[16]提出了一种基于 Twitter 数据的地震事件实时检测方法。利用粒子滤波器判断某条 Tweet 是否与地震事件相关,如果与地震相关的 Tweet 数量激增则触发系统报警。文献[17]则通过计算 Tweet 中单词出现频

率加速比的方式,快速识别突发词汇,并提出一种基于单一主题模型的突发词汇聚类算法 TopicSketch。该算法将语义相似的突发词汇检测为同一个事件。由于未对地址信息进行细分处理,该算法存在将语义相似但实则发生在不同局部地域的不同事件误判为同一个广域事件的局限性。文献[18]应用小波变换与自相关系数来度量数据流中词汇的突发性,然后根据互相关系数将突发词汇构建为一种图结构,再通过子图划分确定突发事件。文献[19]以不同国家为地域单位,提出一种基于位置感知的事件检测模型。首先从 Tweet 文本内容中获取地域信息,然后将事件表示为其所涉及的国家范围与参与讨论用户数量的向量,最后基于向量间欧氏距离进行事件识别与检测。

这类方法都是为广域突发事件检测而设计的,如果直接将这些方法应用到公共服务领域,会引入更多的噪声数据,降低事件检测的准确率。

上述工作对突发事件检测做了许多有益的探索,也取得了一些研究成果。然而,目前还缺少一种能够根据事件性质的不同,自适应地确定其地域突发模式的事件检测方法。对此,本文提出了一种基于地域自适应的突发事件实时检测方法,该方法能够:

- (1) 对流式增量数据进行实时突发事件检测;
- (2) 通过基于分层语义分析的候选突发事件识别算法提高候选突发事件检测的准确性;
- (3) 自适应地识别事件发生的地域模式,将分布在各个局部地区的广域突发事件,聚合为一个完整的突发事件,同时,能够过滤局部突发事件中存在的噪声数据,降低事件检测的重复率,提高事件检测准确性。

3 RAEDetection 方法框架

基于地域自适应的突发事件实时检测方法 RAEDetection 的整体框架如图 1 所示。

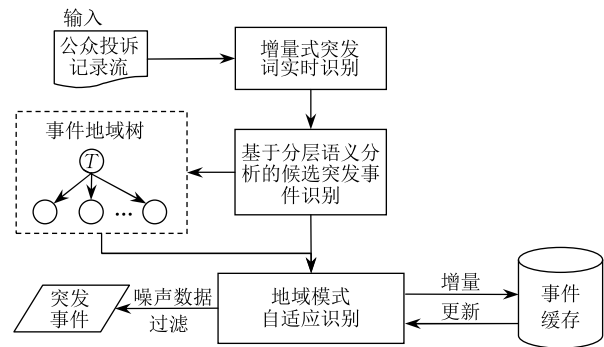


图 1 RAEDetection 方法框架

该方法由 3 个主要部分组成:

(1) 增量式突发词实时识别. 提出一种增量式 Kleinberg 模型, 对诉求记录数据流中词汇的突发状态进行实时识别;

(2) 基于分层语义分析的候选突发事件识别. 采用主题模型实现对于描述相同主题事件的突发词聚类, 再利用文本聚类算法, 如 K-Means 聚类算法^[20], 生成候选突发事件;

(3) 突发事件地域模式自适应识别. 通过构建包含市级、区级、街道级地址的事件地域树, 自适应识别不同事件发生的地域模式. 在此基础上, 对候选突发事件进行噪声数据过滤, 并根据缓存的历史事件信息, 实时获得最终的突发事件.

各部分工作将分别在第 4、5、6 节中给出.

4 增量式突发词实时识别算法

突发词识别对于突发事件检测具有良好的线索作用, 其主要思想是判断某个时间窗口内某个词汇出现的频率是否激增. 如果是, 则有可能存在突发事件^[13].

现有基于 Kleinberg^[21]模型的突发词识别算法只能以批处理的方式对给定时间窗口内的全量数据进行突发词识别. 但是, 其无法对流式增量数据场景下的突发词实时识别.

考虑到公共服务热线中的诉求记录天然包含有序的时序信息, 是一种典型的时序数据流. 因此, 我们对现有的批处理式 Kleinberg 模型进行改进, 提出一种适用于流式数据的增量式 Kleinberg 模型, 克服了现有批处理式 Kleinberg 模型的局限性, 实现突发词汇的实时识别. 该模型结构如图 2 所示.

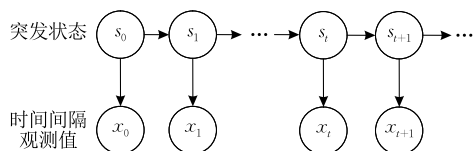


图 2 增量式 Kleinberg 模型

在增量式 Kleinberg 模型中, 单词 w 的突发行为被描述为一段时间窗口内突发状态与非突发状态的相互转换过程. 状态变量 s_t 表示单词 w 在第 t 时刻的突发程度. 若 $s_t = 1$, 则该单词处于突发状态, 即为突发词; 若 $s_t = 0$, 则该单词处于非突发状态. 观测变量 x_t 表示单词 w 在时刻 t 时与上次该单词出现时刻 $t-1$ 之间的时间间隔.

假设单词 w 以一定概率随机出现, 单词 w 前后

两次出现的时间间隔 x 服从的概率分布记为

$B(x_{t+1}, s_{t+1}) = \alpha_{s_{t+1}} e^{-\alpha_{s_{t+1}} x_{t+1}}, \alpha_{s_{t+1}} > 0, s_{t+1} = 0, 1$ (1)

其中, $B(x_{t+1}, s_{t+1})$ 表示输出观测值 x_{t+1} 的概率, $\alpha_{s_{t+1}}$ 表示 $t+1$ 时刻单词 w 在突发状态 s_{t+1} 的条件下的产生速率, 记为

$$\alpha_{s_{t+1}} = (N_w/T) \cdot g^{s_{t+1}}, g > 1 \quad (2)$$

N_w 表示给定时间窗口 $\langle Win_s, Win_e \rangle$ 内单词 w 出现的总次数, T 表示时间窗口 $\langle Win_s, Win_e \rangle$ 的跨度.

从状态 s_t 转移到状态 s_{t+1} 的总体代价函数为

$$A(s_t, x_{t+1}, s_{t+1}) = \tau(s_t, s_{t+1}) - \ln B(x_{t+1}, s_{t+1}) \quad (3)$$

状态 s_{t+1} 取值为 0 或 1. $\tau(s_t, s_{t+1})$ 表示从状态 s_t 转移到 s_{t+1} 的转移代价, 表示为

$$\tau(s_t, s_{t+1}) = \begin{cases} (s_{t+1} - s_t) \gamma \ln N_w, & s_t = 0 \rightarrow s_{t+1} = 1 \\ 0, & s_t = 1 \rightarrow s_{t+1} = 0 \end{cases} \quad (4)$$

规定从突发状态 $s_t = 1$ 转移到非突发状态 $s_{t+1} = 0$ 的代价为 0, 超参数 γ 取值在实验部分给出.

为了对实时诉求记录数据流进行增量式突发词识别, 本文给出增量式 Kleinberg 模型中第 $t+1$ 时刻单词 w 状态的计算公式:

$$s_{t+1} = \arg \min_{s_{t+1} = (0, 1)} A(s_t, x_{t+1}, s_{t+1}) \quad (5)$$

求得使代价函数 A 最小化的 s_{t+1} 值, 即为单词 w 在 $t+1$ 时刻的状态. 若 $s_{t+1} = 1$ 使得 A 最小, 则该单词在 $t+1$ 时刻为突发词.

5 基于分层语义分析的候选突发事件识别算法

根据观察, 我们发现突发事件的发生通常伴随着多个相关突发词的出现. 为了根据突发词挖掘出相对应的突发事件, 同时为了避免因直接针对大量短文本诉求记录进行聚类而造成的数据稀疏性问题, 本文提出一种基于分层语义分析的候选突发事件识别算法, 其设计思想如下:

(1) 主题层语义分析. 本文通过引入主题模型捕获突发词的隐层主题语义相似度信息, 将隶属于同一个主题的突发词进行聚类, 指明包含这些突发词的诉求记录与突发主题事件聚类结果间的关系. 文献^[22]曾指出将现有的主题模型直接应用于短文本诉求记录上效果不佳, 为此, 本文给出了一种将短文本诉求记录聚合为长文本数据的方法:

首先, 假设在时间窗口 $\langle Win_s, Win_e \rangle$ 内存在 N 个突发词, 根据第 4 节突发词识别结果, 将所有包含突发词 $w_i (1 \leq i \leq N)$ 的诉求记录聚合为一个长文档, 记为 D_{w_i} . 对所有突发词执行相同操作, 得到长

文档集合 $C = \{D_{w_1}, \dots, D_{w_i}, \dots, D_{w_N}\}$.

然后,运用主题模型在长文档集合 C 上训练得到“文档-主题”分布 θ 与“主题-词”分布 ϕ . 对于诉求记录 r , 根据 Gibbs 采样公式:

$$p(z_{k_i} = k | z_{-k_i}, w), 1 \leq i \leq K \quad (6)$$

计算该记录 r 中的主题分布 ϕ . 其中 K 表示 C 中所有的主题数量, w 表示 r 中的所有词汇向量, z_{k_i} 表示 r 中的第 k_i 个主题.

最后,选择 r 中出现概率最高的主题 $k_{\max}$ 作为该记录的隶属主题,并将隶属于相同主题 k 的记录识别为突发主题事件 TE_k .

(2) 事件层语义分析. 在获得突发主题事件的基础上,我们发现属于同一个突发主题事件的诉求记录集合中时常混杂较多的噪声数据. 例如,诉求记录 1:“广州路社区因道路施工发生停水,请尽快恢复”;诉求记录 2:“广州路社区发生停水,有施工队在附近施工”,诉求记录 3:“上海路水管破裂,道路无法通行”都会被识别为与“停水”相关的突发主题事件. 但实际上,诉求记录 3 显然不是某个特定社区的停水问题,而是由水管破裂引起的交通问题.

为了将突发主题事件进一步细分为多个具体的候选突发事件,过滤其中的噪声数据,本文进一步采用文本聚类算法对每个突发主题事件 TE_k 中的诉求记录集合再次进行事件层级上的语义聚类,最终得到了更细粒度的事件聚类结果,并将每个聚类结果识别为候选突发事件 E_c , 同时将由所有候选事件组成的候选突发事件集合记为 SE_c , $SE_c = \{E_{c_1}, \dots, E_{c_n}\}$.

通过上述基于分层语义分析的候选突发事件识别算法,可以在突发事件检测过程中充分利用突发词及其对应诉求记录的隐层主题语义信息与事件层语义信息,实现对于候选突发事件识别结果的不断细分取精.

6 地域模式自适应识别算法

在获得候选突发事件之后,为了进一步确定事件发生涉及的地域范围,提升突发事件检测的准确性,本文提出了一种基于事件地域树的地域模式自适应识别算法.

首先,判断事件的发生是否与地域相关;若相关,则构建事件地域树,自适应识别该事件发生的地域模式;若不相关,则直接根据事件的语义信息进行事件聚类. 然后再根据识别的地域模式对第 5 节中得到的候选突发事件进行噪声数据过滤,得到最终的突发事件.

6.1 地域相关性判别

给定候选突发事件 E_c , 其包含的每条记录记为 r , $E_c = \{r_1, \dots, r_i, \dots, r_{|E_c|}\}$, 其中, $|E_c|$ 表示候选事件 E_c 中的记录数量. 本文采用 NLTK^① 提供的命名实体抽取工具对每条记录 r 进行地域信息抽取,并保存在其地域集合 $ReginSet(r)$ 中. 设指示函数:

$$I(r) = \begin{cases} 1, & ReginSet(r) \neq \emptyset \\ 0, & ReginSet(r) = \emptyset \end{cases} \quad (7)$$

给出地域相关性阈值 τ , 根据算式(8), 若 E_c 中包含地域信息的记录 r_i 的比例超过 τ , 则认为该候选事件 E_c 的发生与地域相关.

$$\frac{\sum_{i=1}^{|E|} I(r_i)}{|E|} \geq \tau \quad (8)$$

6.2 事件地域树的定义与构建

6.2.1 事件地域树的概念及定义

以南京市诉求记录为例, 候选事件 E_c 中记录 r 的 $ReginSet(r)$ 一般呈现为 {南京市, 鼓楼区, 宁海路街道}, {null, 玄武区, 新街口街道}, {null, null, 夫子庙街道} 等多层级结构. 本文结合国家统计局全国行政区划信息^②, 补全 $ReginSet(r)$ 中缺失的各层级地域信息(标识为 null), 将所有记录 r 的地域信息规范化为 {xx 市, xx 区, xx 街道} 的形式, 记为 $\{L^1, L^2, L^3\}$, 形成一种三层事件地域树的结构. 先给出事件地域树的基本符号解释与定义.

定义 1. 事件地域树 ($ETree$). 事件地域树是一种层数为 3 且满足 3 个条件的多叉树结构:

(1) 第 1 层(市级)中, 存在唯一一个地域节点, 也是该事件地域树的根节点, 记为 L^1 . 其数值 $|L^1|$ 表示候选事件中包含地址 L^1 的记录数量.

(2) 第 2 层(区级)中, 对 L^1 细分为 H 个地域节点, 其中第 i 个节点记为 L_i^2 , $i \in [1, H]$. 其数值 $|L_i^2|$ 表示候选事件中包含地址 L_i^2 的记录数量.

(3) 第 3 层(街道级)中对每个 L_i^2 节点再细分为 m_i 个子节点, L_i^2 节点下第 j 个子节点记为 $L_{i,j}^3$, $j \in [1, m_i]$. 其数值 $|L_{i,j}^3|$ 表示候选事件中包含地址 $L_{i,j}^3$ 的记录数量.

6.2.2 事件地域树的构建

首先, 将 E_c 中所有记录共有的第 1 层(市级)地域信息作为根节点, 记为 L^1 ; 然后, 将各个记录中的第 2 层(区级)地址信息作为 L^1 的子节点, 记为 L_i^2 , 表示地址为 L^1 市的第 i 个市辖区; 最后将记录中第

① <http://www.nltk.org/>

② <http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/>

3 层(街道)地址信息作为 L_i^2 的子节点, 记为 $L_{i,j}^3$, 表示地址为 L_i^2 区的第 j 个街道。构建的事件地域树如图 3 所示。

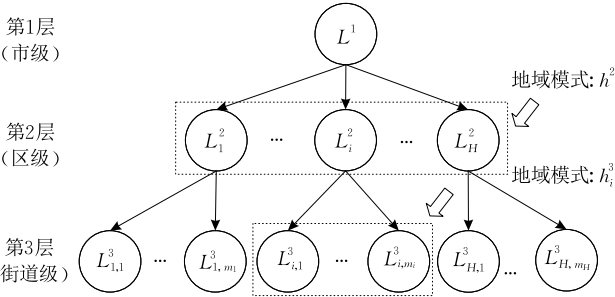


图 3 事件地域树

6.3 地域模式自适应识别

相关研究表明^[7], 局部模式事件主要集中在某一地域范围内; 广域模式事件发生范围则通常涉及多个地域。

基于此, 我们将一个空间划分为 H 个地域, 每个地域表示为 $L_i, i \in [1, H]$ 。对于第 6.1 节中确定的与地址相关的事件 E_c , 其地域模式可能涉及其中的 h 个地域 ($1 \leq h \leq H$)。根据最大熵原理^[23], 假设某个事件 E 的真实地域分布服从参数为 h 的离散型均匀分布 $L_i \sqcup U(h)$, 将其记为事件的地域假设分布 $P(L_i)$, 概率分布函数为

$$P(L_i) = \frac{1}{h}, i = 1, 2, \dots, h \quad (9)$$

其中, $P(L_i)$ 的数值表示事件 E 发生在地域 L_i 中的概率。

通过统计候选事件 E_c 中的地址信息, 可以得到候选突发事件 E_c 实际发生的地域分布概率函数, 将其记为事件的地域频率分布 $Q(L_i)$:

$$Q(L_i) = \frac{N_i}{N}, i = 1, 2, \dots, K \quad (10)$$

其中, N_i 表示候选事件 E_c 发生在地域 L_i 中的记录数量, N 表示该候选事件中的记录总数。

考虑到总有 $h \leq H$, 因此, 取 $Q(L_i)$ 中概率值最大的前 h 个值: $Q(L_1), \dots, Q(L_i), \dots, Q(L_h)$, ($Q(L_i) \geq Q(L_{i+1})$), 并按照概率值降序排序, 得到突发事件 Top- h 地域集合 Top_h 。同时, 对这些地址出现的概率进行归一化处理, 得到:

$$Q'(L_i) = \frac{N_i / N}{\sum_{i=1}^h N_i / N} \quad (11)$$

为了确定 h 的最佳取值, 识别该突发事件的地域模式, 本文采用 KL 距离来衡量统计获得的候选

事件 E_c 的地域频率分布与其对应的地域假设分布之间的差异性, 目标函数如算式(12)所示。

$$J_h = \sum_{i=1}^h \frac{1}{h} \log \frac{P(L_i)}{Q'(L_i)}, Q'(L_i) \geq Q'(L_{i+1}) \quad (12)$$

通过网格搜索的方式, 确定使得 J_h 最小的 h 值, 即最小化 $P(L_i)$ 与 $Q'(L_i)$ 之间的偏差, 如算式(13)所示。

$$h = \arg \min_{1 \leq h \leq H} \{J_1, \dots, J_h, \dots, J_H\} \quad (13)$$

最终确定该事件 E 的地域分布模式:

(1) 若 $h=1$, 则事件 E 呈现局部模式, 其发生的地域范围只集中在突发事件地域集合 Top_h 中出现概率最高的地址范围内;

(2) 若 $h>1$, 则事件 E 呈现广域模式, 其发生地域范围涉及突发事件地域集合 Top_h 中的前 h 个地址, 其余地域 $L_{h+1}, \dots, L_K$ 则可作为噪声数据进行过滤。

需要指出的是, 对于事件地域树 $ETree$ 而言, 位于第 2 层(区级)地址集合 $\{L_1^2, \dots, L_H^2\}$ 的地域突发模式由 h^2 表示, 其数值含义为该事件发生的地域范围涉及 h^2 个区级地域, 相对应的突发事件地域范围集合表示为 Top_h^2 。

类似地, 位于第 3 层(街道级)的某个 L_i^2 节点下子节点集合 $\{L_{i,1}^3, \dots, L_{i,m_i}^3\}$ 的地域突发模式由 h_i^3 表示, 其数值含义为该事件发生的地域范围涉及 h_i^3 个街道级地域, 相对应的突发事件地域范围集合表示为 Top_h^3 , 如图 3 中黑色虚线框中所示。

6.4 突发事件生成

根据第 6.3 节中地域模式自适应识别的结果, 我们给出从候选事件移除噪声数据的过滤策略:

过滤策略. 如果候选事件 E_c 中记录 r 的地址 L_i^j 不在其对应层级的突发事件地域集合 Top_h^j 中, 则将该记录作为噪声数据移除。

结合第 6.2 节定义的事件地域树, 算法 1 给出基于地域自适应的突发事件检测算法。

算法 1. 基于地域自适应的突发事件检测算法。

输入: 候选突发事件 E_c , 事件地域树 $ETree$, 地域相关性阈值 τ

输出: 突发事件 E

1. IF $\sum_{i=1}^{|E_c|} I(r_i) / |E| < \tau$
2. 将 E_c 标识为 E ; /* 该事件与地域无关 */
3. ELSE
4. 计算 $ETree$ 中 $\{L_1^2, \dots, L_H^2\}$ 的最优 h^2 取值, 获得 Top_h^2 ; /* 根据算式(13)计算区级地域模式 */
5. FOR each $r \in E_c$ do
6. IF $r.L^2 \in Top_h^2$ /* 执行过滤策略 */

```
7.       $k=r.L^2$  在  $\{L_1^2, \dots, L_H^2\}$  中对应次序的下标.
8.      将记录  $r$  放入临时记录集合  $S_k$  中;
9.      END IF
10.    END FOR
11.  FOR  $i=1$  to  $H$ 
12.    IF  $S_i \neq \emptyset$ 
13.      计算  $E_{Tree}$  中  $\{L_{i,1}^3, \dots, L_{i,m_i}^3\}$  的最优  $h_i^3$  取值, 获得  $Top\_h^3$ ; /* 根据算式(13)计算街道级地域模式 */
14.      FOR each  $r \in S_i$  do
15.        IF  $r.L^3 \in Top\_h^3$  /* 执行过滤策略 */
16.          将记录  $r$  放入事件  $E$  中;
17.        END IF
18.      END FOR
19.    END IF
20.  END FOR
21. END ELSE
```

对于候选事件集合 SE_c 中的每个候选事件 E_c 执行算法 1, 得到最终的突发事件集合 $SE = \{E_1, \dots, E_n\}$.

6.5 基于滑动时间窗口的突发事件实时检测

为了实现对于突发事件的实时检测, 本文提出了一种基于滑动时间窗口的突发事件实时检测策略, 如图 4 所示.

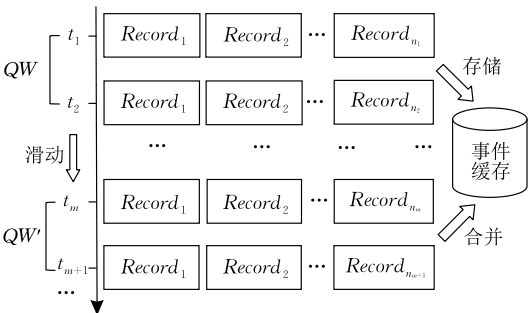


图 4 增量突发事件检测策略

首先, 我们对时间窗口 QW 内所有诉求记录执行 RAEDetection 方法进行突发事件检测, 将识别出的事件记上相应的编号, 并且存储这些历史事件.

随后, 当滑动时间窗口从 QW 滑动到 QW' 时, 我们对时间窗口 QW' 内的所有诉求记录执行相同的突发事件检测方法.

最后, 将在滑动时间窗口 QW' 内新发现的事件与在时间窗口 QW 内发现的历史事件进行相似度比较. 如果新发现的事件与历史事件的语义相似度超过一定阈值 δ_{sim} 且具有相同的地域模式, 则将新检测出的事件与历史事件合并为同一个事件; 否则, 将新检测出的事件作为一个新的事件, 并记上相应

的事件编号. 依次类推, 实现突发事件的实时检测. 滑动时间窗口的大小与相似度阈值 δ_{sim} 的最佳取值将在实验部分第 7.5.3 小节中给出.

7 实验结果与分析

7.1 数据集

本文采用来自公共服务热线中南京与苏州市的真实诉求数据, 以及来自 Twitter 的 New York 市数据^[12]作为实验数据集, 基本信息如表 2 所示.

表 2 数据集描述

城市	诉求记录 总数量/条	时间周期	平均每条记录 长度/字(word)
南京	89795	2018.08.01 to 2018.08.23	42
苏州	417326	2019.02.02 to 2019.02.26	46
New York	32198	2010.06.01 to 2010.06.30	13

7.2 实验设置

7.2.1 数据预处理

对于原始数据集, 进行如下预处理: (1) 对原始记录进行分词; (2) 去除停用词和词性为介词、连词等无实体意义的词; (3) 根据词性, 抽取词性为地址名词的词作为记录的地域信息, 并对地址进行规范化处理; (4) 去除词数少于 3 的记录.

7.2.2 对比方法与参数设置

本文选取 TrioVecEvent^[12]、GeoBurst^[8]与 TopicSketch^[17]这 3 种最新的突发事件检测方法作为对比方法, 各参数按照其各自原始论文中给出的最佳值进行选取, 具体如下:

TrioVecEvent: 将分布维数 D 、最大聚类个数 $MaxClusterNum$ 、Gibbs 采样迭代次数 I 分别设置为 100、500 与 10.

GeoBurst: 依据其原始论文的参数设置, 我们将核函数带宽 h 、重启概率 α 、随机游走相似度阈值 δ 以及平衡时空突发性的排序参数 η 分别设置为 0.01、0.2、0.02 以及 0.5.

TopicSketch: 该方法使用一组关键词表示一个事件. 我们对每个事件选择概率最高的 10 个关键词作为事件关键词, 采用 BM25^[24]检索模型查询最能代表该事件的 20 条诉求记录. 将哈希表规模 B 、时间窗口 ΔT_1 与 ΔT_2 分别设置为 5000、15 与 30.

本文提出的方法 RAEDetection 参数设置如下: 在突发词实时识别算法中, 将时间间隔概率函数参数 g 设置为 2, 转移代价参数 γ 设置为 0.1; 在候选突发事件聚类算法中, 我们选用 LDA^[25]作为主题层语义分析的主题模型, 将主题个数 K 设置为 50; 同

时,选用 K -Means 算法作为事件层语义聚类算法;在滑动时间窗口参数设定部分,滑动时间窗口大小设置为 60 min,语义相似度阈值 δ_{sim} 设置为 0.65.

7.2.3 实验环境与设置

本文采用 Python3.6 实现了各类对比算法,使用的计算服务器配置信息参见表 3.

表 3 计算服务器的配置信息

项目	配置信息
CPU	8 Core Intel Xeon 2.60GHz×2
Memory	192 GB
Disk	8 TB Raid0
OS	CentOS Linux release 7.4

7.3 评价指标

本文借鉴文献[8,12]提出的事件检测方法准确性评价指标,具体如下:(1)查准率 P . $P = N_{true} / N_{detected}$,其中 N_{true} 表示检测出的真实突发事件数量, $N_{detected}$ 表示检测出的所有突发事件数量;(2)伪查全率 R_{pseudo} . 由于难以通过确定数据集上全部突发事件的数量来计算查全率,因此提出伪查全率.将在同一数据集上采用不同方法检测出的真实突发事件进行合并与去重,并令 N_{total} 为合并后事件的数量,伪查全率 $R_{pseudo} = N_{true} / N_{total}$;(3)伪 $F1$ 值 $F1_{pseudo}$. 计算公式为 $F1_{pseudo} = 2 \times P \times R_{pseudo} / (P + R_{pseudo})$.

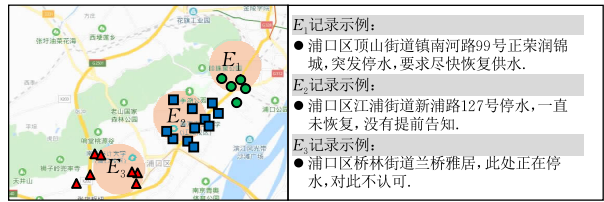
因为数据集中记录数量较大,难以在运行算法前一一标注,所以,对于检测出的突发事件,本文采用人工判断的方式确认检出的事件是否为突发事件.同时,对于南京与苏州市的数据集,本文结合江苏省公共服务热线平台提供的反馈信息进一步确认这些检出的突发事件是局部事件,还是广域事件;对于来自 Twitter 的 New York 市数据,我们同样采用人工判断的方式确定检出的事件是否存在重复现象.通过上述方式,可以发现本文提出的 RAEDetection 方法与基线对比方法是否存在将某一个广域事件重复地识别为多个局部事件的现象.

事件重复检测的概念解释为:如果某一个广域突发事件被错误地拆分为多个局部突发事件,则会产生重复现象.如果某个突发事件检测算法的检测重复率增高,则会导致检出的突发事件中存在多个相同的事件,使得查准率指标 P 下降,具体分析案例将在第 7.4 小节中给出.

7.4 定性分析

为了阐述本文所提 RAEDetection 方法的有效性以及与各种基线方法的对比效果,图 5 与图 6 分别给出了本文选出的 3 种不同事件检测方法对具备

不同地域模式突发事件的检测效果例证.图中方框、圆点与三角形符号分别表示检测出的不同事件;叉状符号表示噪声记录.需要说明,因为 TrioVecEvent 算法与 GeoBurst 算法同属局部事件检测算法,而文献[12]的结论指出 TrioVecEvent 算法准确性更佳,因此,我们选取 TrioVecEvent 进行分析对比.



(a) TrioVecEvent



(b) TopicSketch

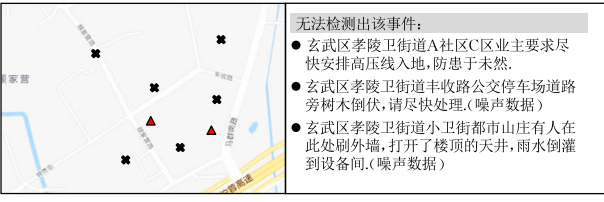


(c) RAEDetection

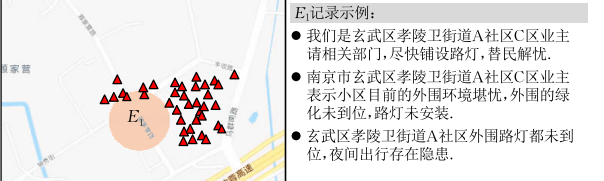
图 5 浦口区多个街道停水事件检测结果(广域模式事件)



(a) TrioVecEvent



(b) TopicSketch



(c) RAEDetection

图 6 某社区缺少照明路灯的事件检测结果(局部模式事件)

图 5 描述了南京市浦口区多个街道突发的停水事件,这是一个发生范围涉及多个地域的广域模式事件.从图 5(a)中可以看出,TrioVecEvent 算法将该事件重复地识别为 3 个事件.这是因为 TrioVecEvent 算法是一种针对局部模式的事件检测算法,更倾向于将地理位置接近的诉求记录集合识别为一个事件.对于发生范围涉及多个地址的广域模式事件,该算法则容易将一个广域事件重复地识别为多个局部事件,增高事件检测的重复率,降低检测的查准率.

从图 5(b)中可以看出,TopicSketch 算法同样将这一个涉及多个地域的停水事件重复地识别为 2 个不同的事件.这是因为该算法只考虑到了诉求记录的语义相似度,忽视了不同事件发生的地域模式存在差异性.此外,同一个事件的发生也存在时延.该算法将两个前后顺接时间窗口中的同一个事件重复地识别为 2 个事件.

从图 5(c)中可以看出,本文提出的事件检测方法能够自适应地识别出该停水事件发生的地域模式为广域模式,其中区级地域突发模式 $h^2=1$,即浦口区,街道级地域突发模式 $h^3=3$,即顶山街道、江浦街道与桥林街道.因此,该方法能够将发生在相隔距离较远但隶属于同一个行政区的不同街道中的停水诉求记录,识别为同一个事件.与 TrioVecEvent、TopicSketch 算法相比,本文方法的检测准确率更高,并且降低了事件检出的重复率.

图 6 则描述了玄武区某居民区因小区路灯等基础设施安装不到位,夜晚没有照明而造成的突发投诉事件.这是发生范围集中在一个地域的局部模式事件.

从图 6(a)中可以看出,TrioVecEvent 算法可以检测出该事件,图中的三角形表示与该事件相关的诉求记录.但是,该事件检测的结果中还存在一些噪声数据,图中用叉状符号标出.这是因为该小区附近也存在一些与路灯相关的投诉记录,而 TrioVecEvent 仅能根据地理距离进行聚类,无法对地址的语义信息进行识别,所以,无法过滤这些发生在该小区周围与路灯相关的噪声数据.

图 6(b)的结果表明 TopicSketch 算法无法检测出该事件.这是因为 TopicSketch 算法是一种针对广域模式的事件检测算法,只利用诉求记录的内容相似性进行事件检测,而对于地址的变化并不敏感.考虑到路灯维修是一种较常见的诉求,因此,TopicSketch 算法并未将该小区路灯未安装的事件识别为一个突发事件.这与文献[14]认为广域事件检测

算法有可能遗漏某些局部事件的结论相符合.

图 6(c)给出了本文提出的 RAEDetection 算法的检测结果.该事件被检测为局部突发事件,其中区级地域突发模式 $h^2=1$,即玄武区,街道级地域突发模式 $h^3=1$,即孝陵卫街道.实验结果再次表明,我们的方法可以正确检测出该事件,并且能够根据该事件发生的局部地域模式,将语义上相似,但在地域分布模式上不同的噪声数据进行过滤.

7.5 定量分析

7.5.1 算法准确性分析

根据 7.3 节中给出的评价指标,我们分别对 TrioVecEvent、GeoBurst、TopicSketch 与 RAEDetection 的准确性进行分析.我们选取南京市与苏州市的真实诉求记录数据作为实验数据集,分别测试了这些方法相应的 P 、 R_{pseudo} 与 $F1_{\text{pseudo}}$ 值,其准确性结果分别如表 4、表 5 所示.

表 4 南京市数据集中 5 种不同方法的准确性对比

Methods	P	R_{pseudo}	$F1_{\text{pseudo}}$
TrioVecEvent	0.529	0.375	0.439
GeoBurst	0.445	0.125	0.195
TopicSketch	0.455	0.417	0.435
RAEDetection _{noReg}	0.545	0.417	0.472
RAEDetection	0.704	0.792	0.745

表 5 苏州市数据集中 5 种不同方法的准确性对比

Methods	P	R_{pseudo}	$F1_{\text{pseudo}}$
TrioVecEvent	0.534	0.600	0.565
GeoBurst	0.636	0.200	0.304
TopicSketch	0.807	0.267	0.401
RAEDetection _{noReg}	0.548	0.667	0.602
RAEDetection	0.782	0.800	0.791

从表 4 与表 5 中可以看出,本文提出的 RAEDetection 算法在查准率 P 、伪查全率 R_{pseudo} 与 $F1_{\text{pseudo}}$ 值这 3 项指标上绝大部分都超过了 TrioVecEvent、GeoBurst 以及 TopicSketch 这 3 种事件检测算法,取得了最佳的算法准确率.算法 TrioVecEvent 的准确性次佳,而算法 GeoBurst 的准确性在这 5 种对比算法中最不理想.

我们对此现象进行进一步的分析,其原因主要在于:

(1) 本文提出的算法 RAEDetection 能够自适应地识别事件发生的地域模式.对于广域事件,能够将发生在不同地点的相同事件合并成一个事件;对于局部事件,可以根据识别出的事件分布地域模式过滤掉其中的噪声数据.因此,我们提出的 RAEDetection 算法各项准确性指标最佳.

(2) TrioVecEvent 算法因为采用基于多模态联

合嵌入的方式对突发事件的地址、语义以及时间信息进行编码与聚类,所以可以在没有精确的二维 GPS 坐标信息的情况下,利用投诉记录中以文本形式存在地址字符串进行事件的编码与按地域划分.表 4 与表 5 的实验结果表明该方法比只能利用精确 GPS 坐标信息进行事件检测的 GeoBurst 算法的适用范围更广,突发事件检测的准确率也更高.但是, TrioVecEvent 算法没有充分考虑不同类型事件的地域分布模式,因此其检测的准确性不如 RAEDetection 理想,对于广域模式的事件存在重复检出的情况.

(3) GeoBurst 算法因为过于依赖 GPS 坐标信息,所以在处理缺乏二维 GPS 地理坐标的数据集时,其各项准确性指标表现地最不理想.

(4) TopicSketch 算法是一种基于主题模型的突发事件检测方法,利用数据流中词频的“加速度”概念进行异常突发事件模式的检测,然后再进行突发主题推断与相关事件生成.因为该算法对每个事件选择概率最高的多个关键词作为表示该事件的代表性关键词,并采用 BM25^[24] 检索模型查询这些关键词对应的诉求记录,得到最终的事件检测结果.这种从“主题”到“关键词”,再从“关键词”到具体“投诉记录”的多次转换过程会引入较多的噪声数据,导致与 RAEDetection 和 TrioVecEvent 相比,其综合检测准确率指标 $F1_{pseudo}$ 降低.

进一步,为了细化分析地域模式自适应识别对于突发事件检测准确性的提升作用,我们将未引入地域模式自适应识别算法的 RAEDetection 算法记为 RAEDetection_{noReg},该简化算法仅将文本地址进行语义聚类,实验结果分别在表 4 与表 5 中给出.可以发现,在南京与苏州市数据集中,具备地域模式自适应识别能力的 RAEDetection 算法的检测准确性综合指标 $F1_{pseudo}$ 比 RAEDetection_{noReg} 分别提高 57.84% 与 31.40%,这表明对于事件发生的地域模式进行识别能够显著提升事件检测的准确性.

此外,我们还选取来自 Twitter 的 New York 市公开数据作为实验数据集,对这 5 种突发事件检测方法的准确性进行评估.因为 Twitter 数据中地址信息主要为 GPS 坐标,所以在运行 RAEDetection 算法前我们采用 xGeocoding^① 软件工具将 Twitter 中的 GPS 坐标解析为“州(state)、市(city)、街区(street)”三级地址结构,从而满足本文提出的突发事件检测算法的数据格式要求,实验结果如表 6 所示.

表 6 New York 数据集中 5 种不同方法的准确性对比			
Methods	P	R_{pseudo}	$F1_{pseudo}$
TrioVecEvent	0.400	0.200	0.267
GeoBurst	0.364	0.200	0.258
TopicSketch	0.500	0.100	0.167
RAEDetection _{noReg}	0.444	0.400	0.421
RAEDetection	0.500	0.650	0.565

根据表 6,在查准率 P 这项指标上,5 种算法的性能表现较接近.其中,算法 RAEDetection 与 TopicSketch 表现最佳,均为 0.5,算法 RAEDetection_{noReg} 表现次之,为 0.444,而 GeoBurst 表现最不理想,为 0.364.但是在伪查全率 R_{pseudo} 指标上,5 种算法的准确性差距较大.其中,本文提出的 RAEDetection 算法性能最佳,RAEDetection_{noReg} 次之,而另外 3 种基线方法的伪查全率指标均不理想.这也导致在综合性评价指标 $F1_{pseudo}$ 上,本文提出的算法准确性最佳.

为了解释该现象,我们进一步对 Twitter 数据集中突发事件识别的结果进行分析,总结为两方面的原因:

(1) Twitter 数据集中地理信息生成的特殊性.由于 Twitter 是一种开放的社交平台,每一时刻都有大量的用户会在某个公园或者“网红”景点进行“打卡”操作.这会导致有大量 Twitter 记录的 GPS 坐标集中在某个景点,但是其记录内容并不反映某个具体的突发事件.

(2) 基线算法对地域模式识别的偏好性. TrioVecEvent 与 GeoBurst 算法在突发事件检测的过程中过于依赖数据集中的地址信息,致使检测出的突发事件大多局限在某一个局部地区,而遗漏了分布在其他区地域中的相关事件.因此,其事件检测的伪查全率 R_{pseudo} 较低.

此外,虽然 TopicSketch 算法本身对于地理位置坐标不敏感,但是由于该算法在聚类 Twitter 记录的过程中采用了基于主题模型的方法,会将多个具体事件混合为同一个主题,引入的噪声数据较多,同样导致最终事件检测伪查全率指标的下降.为了避免该问题,本文提出的 RAEDetection 算法则先将原始记录进行主题划分,再利用 K-Means 算法对每一个突发主题事件进一步细分为多个突发事件,从而提升了算法的整体准确性.

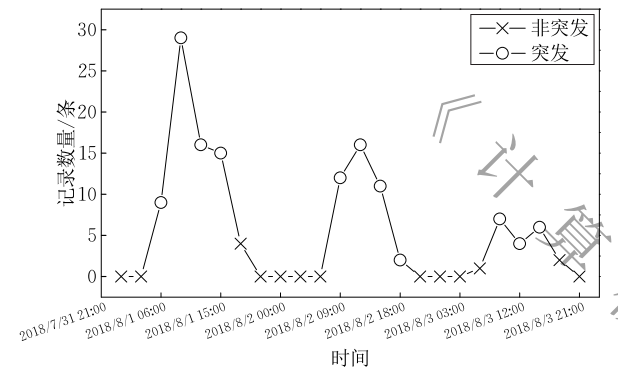
在上述两方面因素的共同作用下,本文提出的 RAEDetection 算法在 Twitter 数据集中也能取得较基线方法更好的准确性.

① <http://www.gpsspg.com/xGeocoding/>

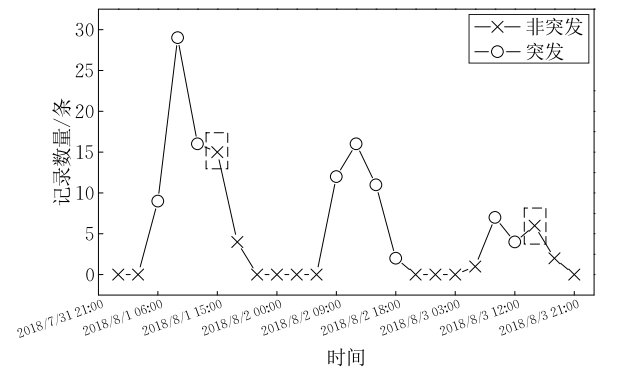
7.5.2 突发词实时识别效果的验证

为了细化验证本文提出的基于增量式 Kleinberg 模型的突发词实时识别算法,同时考虑到真实数据集中的数据规模较大,难以用人工标注的方式一一给出每个词汇在每一时刻的突发状态,因此,我们将现有批处理式 Kleinberg 模型作为对比基准,对突发词实时识别的准确性进行对比分析。

首先,以苏州市某店铺环保问题事件为例,图 7(a)与(b)中分别展示了该事件中突发词“店铺”的突发状态可视化分析结果.图 7 中横轴表示时间,纵轴表示该事件中包含“店铺”的投诉记录数量.图中的圆圈表示将出现在该时刻的词汇“店铺”识别为突发词,叉状符号表示该时刻出现的词汇“店铺”为非突发词。



(a) 基于现有批处理式 Kleinberg 模型的突发词识别效果



(b) 基于增量式 Kleinberg 模型的突发词识别效果

图 7 突发词识别可视化分析示例

从图 7 中可以看出,在 2018/8/1 03:00 至 2018/8/1 15:00,2018/8/2 6:00 至 2018/8/2 16:00,2018/8/3 9:00 至 2018/8/3 15:00 这 3 个时间段内,词汇“店铺”被识别为突发性增长词汇.以此为线索,我们进一步对这些时间段内的诉求记录执行本文提出的突发事件实时检测方法,可以从挖掘出与某店铺环保相关的突发事件。

此外,根据图 7(b),基于增量式 Kleinberg 模型的突发词识别效果与图 7(a)中基于现有 Kleinberg

模型的识别效果整体高度相似,在 24 个时间点上仅有 2 个时间点中词汇“店铺”的突发性识别结果不相同,已在图 7(b)中用黑色虚线方框圈出.与现有批处理式 Kleinberg 模型相比,增量式 Kleinberg 模型的突发词识别准确率达到 91.67%.并且,采用本文提出的增量式 Kleinberg 模型进行突发词识别的优势在于可以对于流式增量投诉记录中关键词的突发性进行实时识别,而现有批处理式 Kleinberg 模型则只能采用批量处理的方式,对划定时间窗口范围内累积投诉记录中的词汇进行突发性识别。

随后,为了进一步定量评估突发词实时识别算法的准确率,本文借鉴机器学习中混淆矩阵(confusion matrix)的概念对突发词识别的准确性进行评估,如表 7 所示.本文将处于突发状态的词汇定义为正元组,非突发状态的词汇定义为负元组. TP 表示将处于突发状态的词正确识别为突发状态; TN 表示将处于非突发状态的词正确识别为非突发状态; FP 表示将处于非突发状态的词错误识别为突发状态; FN 表示将处于突发状态的词错误识别为非突发状态。

表 7 混淆矩阵

真实情况	预测情况	
	突发状态	非突发状态
突发状态	TP	FN
非突发状态	FP	TN

由混淆矩阵计算出相应的单词突发状态识别准确性评价指标为

$$P_{bursty} = \frac{TP}{TP + FP},$$
$$R_{bursty} = TP_{bursty} Rate = \frac{TP}{TP + FN},$$
$$F1_{bursty} = \frac{2 \times P_{bursty} \times R_{bursty}}{P_{bursty} + R_{bursty}}.$$

最后,我们从南京与苏州市的突发事件识别结果中分别随机选取了 10 个代表性突发词进行突发性实时识别准确性评估,时间范围均为 5 天,以 3 小时为划分,共 40 个时间片段.对于这 20 个突发词,我们总计统计了 800 个时间点上的词汇突发性识别结果,具体数据如表 8 所示。

表 8 单词突发性实时识别准确性评估

数据集	P_{bursty}	R_{bursty}	$F1_{bursty}$
南京	0.925	0.829	0.864
苏州	0.909	0.832	0.897

表 8 的结果表明,本文提出的增量式 Kleinberg 模型与现有批处理式模型效果相当,但是在突发词

实时识别上更加具备时效性优势,能够满足流式增量分析场景的需求.

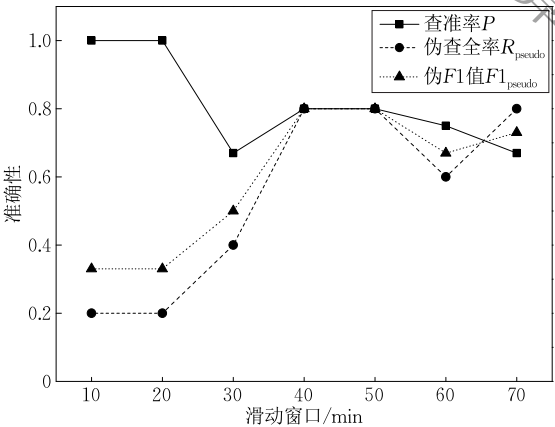
7.5.3 算法参数对准确性的影响

为了确定滑动时间窗口大小 $|QW|$ 以及滑动时间窗口间历史事件语义相似度阈值 δ_{sim} 的最佳取值,我们分别分析了本文提出的 RAEDetection 方法中 $|QW|$ 与 δ_{sim} 对其事件检测的准确性影响.

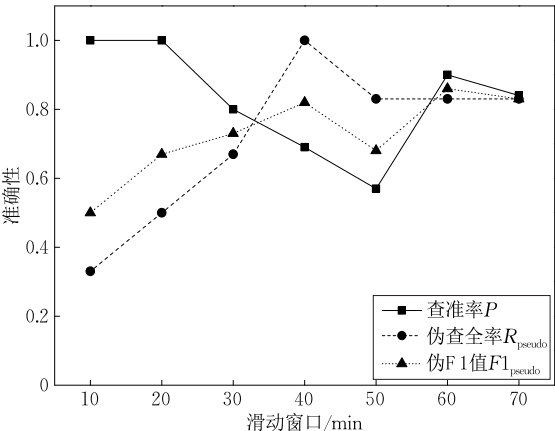
对这些参数进行调优在算法实际落地应用过程中具有重要意义.用户可以根据不同场景下其对于算法响应速度、检测准确率进行综合性权衡与参数选择,进而满足真实应用的需求.

(1) 滑动时间窗口大小对准确性的影响

根据 7.2.2 小节中给出的实验设定,首先将滑动时间窗口间历史事件语义相似度阈值 δ_{sim} 设定为 0.65,然后,滑动时间窗口大小依次设置为 10、20、30、40、50、60 与 70 min,分别在南京市与苏州市数据集上探讨最佳滑动时间窗口大小与事件检测准确性之间的关系.两个数据集集中的实验结果分别如图 8(a)与图 8(b)所示.



(a) 南京市数据集



(b) 苏州市数据集

图 8 滑动时间窗口大小对准确性影响

根据图 8(a),在南京市数据集上,在滑动时间窗口从 10 min 增长到 40 min 的过程中,事件检测的伪查全率 R_{pseudo} 与 $F1_{pseudo}$ 的性能指标逐渐提升,到达 40 min 之后趋于稳定.而查准率 P 的性能指标则从 1.0 逐渐减低到 0.8 附近,同样也在滑动时间窗口大小到达 40 min 的时候趋于稳定.

主要原因在于,当滑动时间窗口取值较小时,其中所包含的真实突发事件的个数与描述信息也较少,RAEDetection 能够检测出其中的突发事件.但是由于无法预知后续时间窗口中的突发事件,所以算法的查准率 P 较高,伪查全率 R_{pseudo} 则较低,进而导致 $F1_{pseudo}$ 的性能指标也较低.当滑动时间窗口的取值逐渐增大到 40 min 时,其中包含的真实突发事件的个数与全局信息也得到增强,因此,RAEDetection 方法的伪查全率 R_{pseudo} 与 $F1_{pseudo}$ 的性能指标也随之提升,同时,查准率 P 的性能指标则会稍稍降低.这是因为当数据量增大时,噪声数据的引入会导致查准率 P 的性能指标降低.当滑动时间窗口继续增大到 40 min 以上时,算法准确性中的各项性能指标最终趋于稳定.

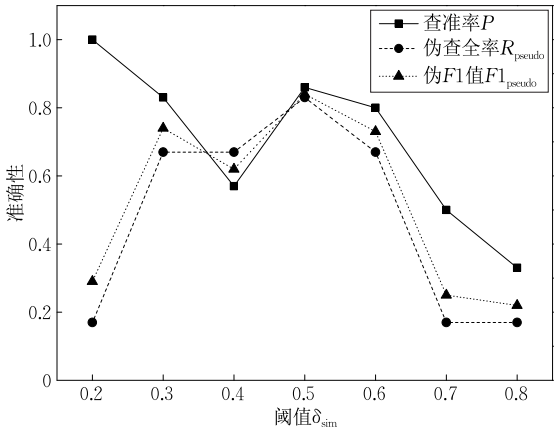
相似的趋势也在图 8(b)中苏州市的数据集上得到验证.当滑动时间窗口从 10 min 增加到 40 min 时,RAEDetection 的 R_{pseudo} 、 $F1_{pseudo}$ 指标逐渐提高并趋于稳定,当滑动时间窗口取值为 60 min 时,RAEDetection 方法准确性各项指标达到最佳.

(2) 滑动时间窗口间相似度阈值对准确性的影响

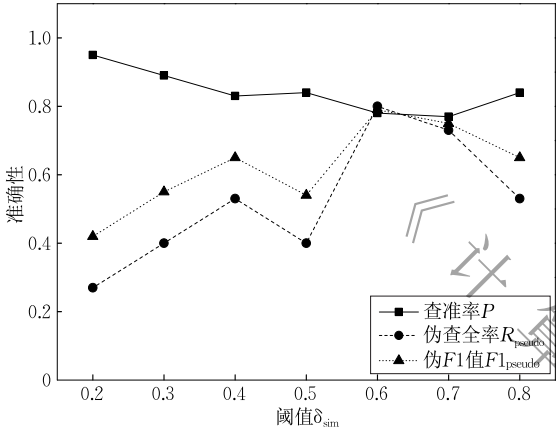
根据图 8 测定的实验结果,我们将滑动时间窗口的最佳取值设置为 40 min.然后,滑动时间窗口间历史事件语义相似度阈值 δ_{sim} 依次设置为 0.2、0.3、0.4、0.5、0.6、0.7、0.8,分别在南京市与苏州市数据集上探讨相似度阈值与事件检测准确性之间的关系.两个数据集集中的实验结果分别如图 9(a)与图 9(b)所示.

由图 9(a)可知,在南京市数据集上,在相似度阈值从 0.2 增长到 0.5 的过程中,事件检测的伪查全率 R_{pseudo} 与 $F1_{pseudo}$ 的性能指标分别从 0.18、0.3 逐渐提升到最高点 0.8 附近,而查准率 P 的性能指标则从 1.0 逐渐减低到 0.82 附近.当相似度阈值继续从 0.5 增加到 0.8 的过程中, P 、 R_{pseudo} 、 $F1_{pseudo}$ 则分别下降到 0.36、0.18 与 0.22 附近.

从图 9(a)的趋势图中可以看出当相似度阈值 δ_{sim} 取到 0.5 时,RAEDetection 准确性指标综合效果最佳.



(a) 南京市数据集



(b) 苏州市数据集

图 9 滑动时间窗口间相似度阈值对准确性影响

主要原因在于,当相似度阈值取值较低时(0.2~0.5 之间),只能检测出 1 至 2 个具有实际意义的突发事件,所以导致查准率虚高.同时,因为相似度阈值设置较低,会引入较多噪声记录,致使伪查全率 R_{pseudo} 与伪 F1 值的性能指标较低.当相似度阈值取为 0.5 时,算法 RAEDetection 准确性的各项指标综合性能达到最佳.当相似度阈值继续从 0.5 增加到 0.8 时,较高的相似度阈值则会导致某些原来能够被正确识别为属于某个突发事件的诉求记录被错误地作为噪声数据过滤掉了,最终使得算法性能各项指标开始下降.

相似的趋势也在图 9(b)给出的苏州市数据集上得到了验证.当相似度阈值取值为 0.6 时,算法 RAEDetection 准确性各项性能综合指标达到最佳.

通过上述两个实验,可以在实际应用过程中为本文提出的 RAEDetection 方法的参数选择提供数据支撑.

7.5.4 算法运行效率分析

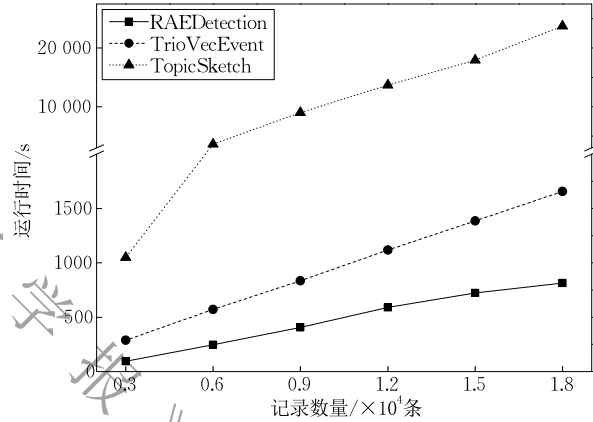
在实际应用中,突发事件检测通常具有很高的

时间效率要求,因此,检测算法不仅要考虑检测精度,同时也要考虑算法的计算性能,这在数据规模较大的应用场景中尤为重要.

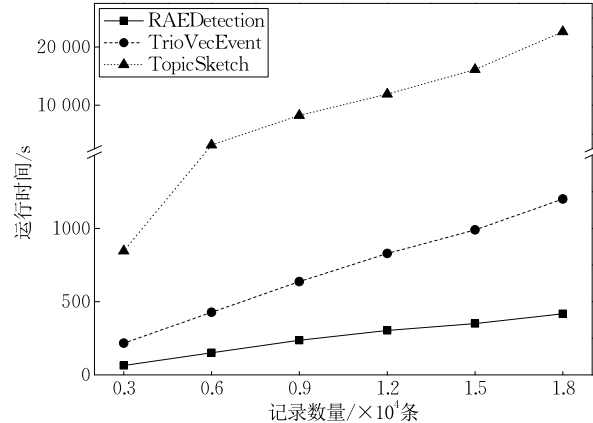
为了验证算法在真实场景中的计算性能,我们将所提出的 RAEDetection 算法与其他两种基线算法进行计算性能和运行效率比较.使用的数据集分别为南京与苏州市两地的真实诉求记录集合.

需要说明, TrioVecEvent 算法与 GeoBurst 算法同属局部事件检测算法,而文献[12]同时指出 TrioVecEvent 算法的运行效率更佳.因此,本文还是选取 TrioVecEvent 算法进行算法运行效率的对比.

从图 10 中看出,RAEDetection 的运行效率最高, TrioVecEvent 算法的运行效率次之, TopicSketch 算法的运行效率最低.



(a) 南京市数据集中算法运行时间与记录条数间关系



(b) 苏州市数据集中算法运行时间与记录条数间关系

图 10 三种事件检测方法的运行时间比较

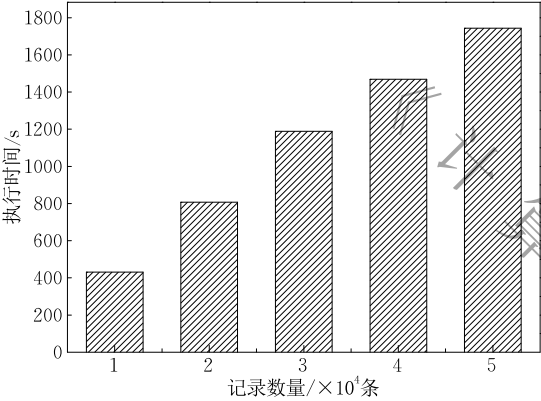
RAEDetection 的运行效率比 TrioVecEvent 高的原因在于,我们通过突发词检测对诉求记录进行过滤,只保留突发词对应的诉求记录,降低了数据规模,而 TrioVecEvent 则需要对所有诉求记录进行词嵌入向量的训练与聚类,因此耗时较多.

TopicSketch 算法运行效率最低的原因在于该算法需要构建二元词对和三元词对矩阵,并且对这些词对共现矩阵进行奇异值分解(Singular Value Decomposition, SVD),这些处理步骤增加了算法复杂度,降低了算法运行效率.

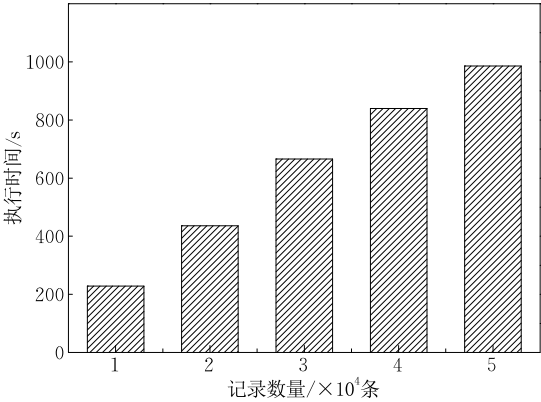
7.5.5 数据可扩展性分析

在该实验中,我们将南京与苏州的真实诉求记录集合作为数据集,验证了 RAEDetection 方法的数据可扩展性. 诉求记录数量从 1×10^4 条增长至 5×10^4 条.

从图 11 中可以看出,在同一数据集中,当数据量线性增长时,RAEDetection 的运行时间也呈现线性增长的趋势. 实验结果表明该方法适用于对大规模诉求记录进行实时突发事件检测.



(a) 南京市数据集中算法执行时间与记录数量间关系



(b) 苏州市数据集中算法执行时间与记录数量间关系

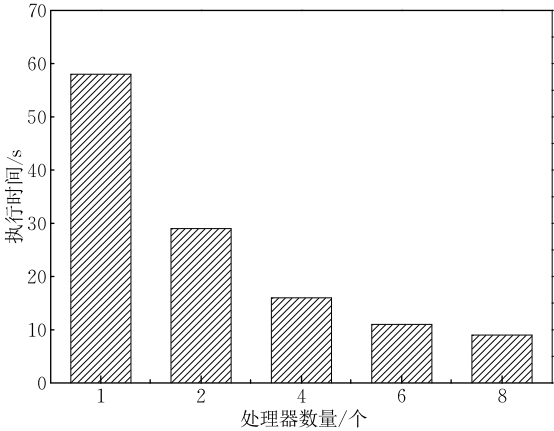
图 11 RAEDetection 算法的数据可扩展性分析

7.5.6 系统可扩展性分析

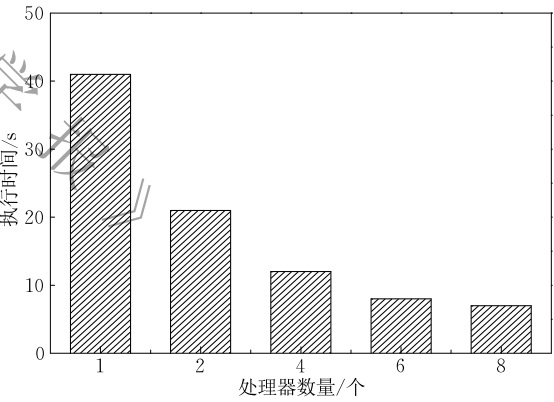
为了分析 RAEDetection 方法运行时间随处理器数量增加时的性能变化情况,本文对 RAEDetection 方法的候选突发事件聚类与地域模式自适应识别部分进行了系统可扩展性实验. 实验数据集同样为南京与苏州市的真实诉求记录集合. 实验中处理器数量从 1 核增加到 8 核.

因为本文提出的突发事件实时检测方法具备对增量数据进行实时分析的能力,所以该方法能够对省公共服务热线平台中已经累积的大量历史数据以及不断涌入的当前数据进行增量式拆分与计算. 因此,虽然本文数据集中给出的数量总量并不是很大,但是能够满足真实应用场景中的需求.

从图 12 中可以看出,在同一数据集中,当处理器数量增长时,RAEDetection 运行的时间也稳步呈现线性下降趋势,验证了方法的系统可扩展性.



(a) 南京市数据集中算法执行时间与处理器数量间关系



(b) 苏州市数据集中算法执行时间与处理器数量间关系

图 12 RAEDetection 算法的系统可扩展性分析

8 总结与展望

现有的突发事件检测方法难以发现海量公共服务热线记录中突发的诉求事件. 对此,本文提出了一种基于地域自适应的突发事件实时检测方法. 通过基于增量式 Kleinberg 模型实现突发词的增量识别与实时处理. 以突发词为线索,采用基于分层语义分析的候选突发事件识别算法获得候选突发事件. 最后,通过构建事件地域树,自适应地识别不同事件发生的地域模式,并据此过滤候选突发事件中的噪声

记录,得到最终的突发事件. 实验结果表明,通过上述不断细分取精的方式,本方法能够有效提高检测的准确率. 与目前最新的方法相比,本文方法具有更好地检测精度和更快的计算性能,相关研究成果已经成功落地应用于江苏省公共服务热线平台上.

下一步工作将尝试更多的突发性检测算法,提高突发事件检测的运行效率与准确率. 同时,考虑对突发事件演进周期中的不同阶段进行趋势追踪,为及时疏解并引导事件发展方向提供指导建议.

参 考 文 献

- [1] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking//Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 37-45
- [2] Fung G P C, Yu J X, Yu P S, et al. Parameter free bursty events detection in text streams//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway, 2005: 181-192
- [3] He Q, Chang K Y, Lim E. Analyzing feature trajectories for event detection//Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, 2007: 207-214
- [4] Li C L, Sun A, Datta A. Twevent: Segment-based event detection from tweets//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Hawaii, USA, 2012: 155-164
- [5] Abdelhaq H, Sengstock C, Gertz M. EvenTweet: Online localized event detection from Twitter. VLDB Journal, 2013, 6(12): 1326-1329
- [6] Zhang Lu-Min, Jia Yan, Zhou Bin, et al. Online bursty events detection based on emoticons. Chinese Journal of Computers, 2013, 36(8): 1659-1667(in Chinese)
(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. 计算机学报, 2013, 36(8): 1659-1667)
- [7] Feng W, Zhang C, Zhang W, et al. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream//Proceedings of the 31st IEEE International Conference on Data Engineering. Seoul, South Korea, 2015: 1561-1572
- [8] Zhang C, Zhou G Y, Yuan Q, et al. GeoBurst: Real-time local event detection in geo-tagged tweet streams//Proceedings of the 39th ACM SIGIR International Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 513-522
- [9] Comaniciu D, Meer P. Mean shift analysis and application//Proceedings of the 7th International Conference on Computer Vision. Kerkyra, Greece, 1999: 1197
- [10] Leetaru K, Wang S W, Cao G F, et al. Mapping the global Twitter heartbeat: The geography of Twitter. First Monday, 2013, 18(5): 1-38
- [11] Unankard S, Li X, Sharaf M A. Emerging event detection in social networks with location sensitivity. World Wide Web, 2015, 18(5): 1393-1417
- [12] Zhang C, Liu L Y, Lei D M, et al. TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 595-604
- [13] Zhong Zhao-Man, Guan Yan, Li Cun-Hua, et al. Localized Top- k bursty event detection in Microblog. Chinese Journal of Computers, 2018, 41(7): 76-88(in Chinese)
(仲兆满, 管燕, 李存华等. 微博网络地域 Top- k 突发事件检测. 计算机学报, 2018, 41(7): 76-88)
- [14] Hua T, Chen F, Zhao L, et al. Automatic targeted-domain spatiotemporal event detection in Twitter. GeoInformatica, 2016, 20(4): 765-795
- [15] Gao X Y, Cao J, He Q, et al. A novel method for geographical social event detection in social media//Proceedings of the 5th International Conference on Internet Multimedia Computing and Service. Huangshan, China, 2013: 305-308
- [16] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 851-860
- [17] Xie W, Zhu F D, Jiang J, et al. TopicSketch: Real-time bursty topic detection from Twitter//Proceedings of the 13th International Conference on Data Mining. Dallas, USA, 2013: 837-846
- [18] Weng J-S, Lee B. Event detection in Twitter//Proceedings of the 5th International Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 401-408
- [19] Quezada M, Pena-Araya V, Poblete B. Location-aware model for news events in social media//Proceedings of the 38th ACM SIGIR International Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 935-938
- [20] Hasan M, Orgun M A, Schwitter R. A survey on real-time event detection from the Twitter data stream. Journal of Information Science, 2018, 44(4): 443-463
- [21] Kleinberg J. Bursty and hierarchical structure in streams//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 91-101
- [22] Jin O, Liu N N, Zhao K, et al. Transferring topical knowledge from auxiliary long texts for short text clustering//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, UK, 2011: 775-784
- [23] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1): 39-71

[24] Jones K S, Walker S, Robertson S E. A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing & Management*, 2000,

36(6): 809-840

[25] Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3: 993-1022



MAI Cheng-Cheng, Ph.D. candidate. His main research interests include event detection and text mining.

CHEN Yu-Ting, M. S. candidate. Her main research interests include parallel computing and event detection.

QIU Xue-Ming, M. S. candidate. His main research interests include social computing and big data.

LIU Jian, M. S. candidate. His main research interests include machine learning and data mining.

ZHAO Bo, Ph. D. candidate. His main research interests include topic model and deep learning.

YUAN Chun-Feng, Ph. D. , professor. Her main research interests include computer architecture, parallel computing and operating system.

HUANG Yi-Huang, Ph. D. , professor. His main research interests include big data, cloud computing and parallel computing.

Background

Event detection is an important technology in data mining and has achieved more and more attention from academia and industry. The data sources of early event detection tasks are mainly news reports. Due to the limitation of data sources, such event detection methods are not suitable for real-time detection. Since 2010, the rise of social media has posed new challenges to bursty event detection. Many studies are looking for efficient and accurate methods to deal with the problem of recognizing the region patterns of the events. According to the different regional scopes of events, the existing detection methods can be divided into local bursty event detection and global bursty event detection.

However, the traditional event detection methods do not take the regional patterns of events into consideration. Meanwhile, the GPS geographic information that these methods depend on is also not easy to obtain. Therefore, it is difficult to use the existing event detection methods to capture the potential hotspots in the public service hotline directly.

In this paper, we propose a real-time bursty event detection method, called RAEDetection, based on the region adaptation. Firstly, an algorithm based on the incremental Kleinberg model is proposed to identify bursty words from the data stream in real-time. Then, based on the topic model, a clustering algorithm for generating candidate bursty events is proposed. With bursty words as clues, this algorithm can divide potential events with similar semantics into more

fine-grained events and form candidate bursty events. Finally, in order to filter out the noise records in the candidate events, the event region tree is constructed to recognize the regional patterns of events. The event region tree has a three-level structure, which corresponds to the addresses in the city, district and street level respectively. According to maximum entropy principle, we assume that the address distribution of one certain event obeys the discrete uniform distribution. We use KL distance to compare the distance between the statistic address distribution and the assumed address distribution. We choose the number of addresses which can minimize the value of the KL distance to indicate the geographical regions of that event, so as to realize the adaptive recognition of regional patterns.

The extensive experimental results show that RAEDetection can effectively detect bursty events in the data stream. Compared with other baseline methods, it can reduce the repetition rate of event detection, and improve the accuracy of event detection. Moreover, our RAEDetection method is more efficient than the compared methods and achieves a good data and system scalability proved by the experimental results on two real-world datasets.

This work is support by the China Science Foundation Research Grant (61572250, U1811461), the Jiangsu Province Science & Technology Research Grant (BE2017155), and the Collaborative Innovation Center of Novel Software Technology and Industrialization of Jiangsu Province.