·先进计算与数据处理·　文章编号：1000-3428(2018)05-0007-07　文献标志码：A　中图分类号：TP391

基于时空分析的突发事件检测方法

梁月仙[1,2,3]，陈自岩[1,2]，王　洋[1,2]，张　跃[1,2,3]，郭　智[1,2]

（1. 中国科学院 空间信息处理与应用系统技术重点实验室，北京 100190；2. 中国科学院电子学研究所，北京 100190；
3. 中国科学院大学，北京 100190）

摘　要： 现有突发事件检测方法多数未考虑事件的重要性，且以孤立的方式看待事件的突发时间域和空间域。为此，提出一种基于时空要素综合分析的突发事件检测方法。引入数据立方体结构存储事件词，通过基于语义相似性的实时事件聚类算法抽取出重要事件。根据 TFIDF 计算事件在时空维度上的出现权重，给出有限状态机-高斯分布模型识别时空突发事件。实验结果表明，该方法能够有效地识别出事件的突发时间段和突发区域，与现有突发事件检测方法相比，检测突发事件的准确率更高。

关键词： 突发事件；时空分析；事件抽取；实时事件聚类；数据立方体

中文引用格式： 梁月仙，陈自岩，王　洋，等. 基于时空分析的突发事件检测方法[J]. 计算机工程，2018，44(5)：7-13.
英文引用格式： LIANG Yuexian，CHEN Ziyan，WANG Yang，et al. Bursty Event Detection Method Based on Spatio-temporal Analysis[J]. Computer Engineering，2018，44(5)：7-13.

Bursty Event Detection Method Based on Spatio-temporal Analysis

LIANG Yuexian[1,2,3]，CHEN Ziyan[1,2]，WANG Yang[1,2]，ZHANG Yue[1,2,3]，GUO Zhi[1,2]

（1. Key Laboratory of Technology in Geo-spatial Information Processing and Application System，Chinese Academy of Sciences，Beijing 100190，China；
2. Institute of Electronics，Chinese Academy of Sciences，Beijing 100190，China；
3. University of Chinese Academy of Sciences，Beijing 100190，China）

【Abstract】 The existing bursty event detection method does not consider the importance of the eveuts，and treats the bursty event time domain and spatial domain of the incident in an isolated manner，and proposes an incident detection method based on comprehensive analysis of spatio-temporal elements. The data cube structure is introduced to store event words，and important events are extracted by a real-time event clustering algorithm based on semantic similarity. TFIDF is used to calculate the occurrence weights of events in the space-time dimension，and the finite state machine-Gaussian distribution model is used to identify spatio-temporal events. Experimental results show that the method can effectively identify bursty time and bursty area of the event，compared with the existing emergency detection method，the accuracy of detecting eveuts is higher.

【Key words】 bursty event；spatio-temporal analysis；event extraction；real-time event clustering；data cube
DOI： 10.19678/j.issn.1000-3428.0046728

0　概述

近年来，世界各地频繁地发生地震、恐怖袭击等突发事件，突发事件的发生严重影响社会秩序的安定和人们生命的安全。互联网上呈现的突发事件信息通常被淹没在众多的普通事件中，人们难以发现潜在的突发性事件，因此，迫切需要一种有效的工具检测出突发性事件。突发事件指在短时间内出现，且其信息量迅速膨胀并随后消亡的事情。突发事件检测旨在从文本中抽取出相关的事件信息并检测其突发性，包括事件抽取和突发性检测两部分。事件抽取指从非结构化的文本中抽取出事件信息并以结构化的形式呈现。

事件抽取主要实现特定事件类型的识别以及事件元素的发现，现有事件抽取方法可分为基于规则匹配的方法、基于监督学习的方法和基于无监督学习的方法。基于规则匹配或监督学习的方法[1-4]依赖于标注语料，存在领域移植性问题，无法有效地运用于开放领域的网络文本。面向开放领域的非监督学习方法采用离线的方式进行事件抽取[5-7]，无法实

基金项目： 国家自然科学基金(41501485)。
作者简介： 梁月仙(1991—)，女，硕士，主研方向为文本数据挖掘；陈自岩、王　洋，助理研究员；张　跃，博士；郭　智，研究员。
收稿日期： 2017-04-10　**修回日期：** 2017-05-25
E-mail： liangyuexian@126.com

时地处理在线的网络数据流。

突发事件检测主要实现事件的突发权重、突发时间段和突发空间区域的识别,已有工作基于事件的词频信息进行突发性检测[8-10],忽略了事件的重要性。另外,事件的突发性不仅与时间序列有关,而且也受地理位置的影响,但是现有大多数工作只考虑事件的突发时间性或突发空间性[11-15]。虽然一些研究[16-17]同时考虑了事件的时空突发性,但是它们以孤立的方式看待事件的突发时间域和突发空间域,未能充分挖掘事件的时空关联性。

针对上述方法存在的问题,本文提出一种联合时空要素综合分析的突发事件检测方法。该方法通过引入数据立方体结构存储事件词,综合分析事件的时空要素并且挖掘事件的时空关联性。同时,给出一种基于语义相似性的实时事件聚类算法,可实时地处理在线的网络数据流,从而摆脱特定领域的限制。在聚类过程中,采用 GloVe 模型挖掘事件词之间的语义关联性,使同一事件类的事件词具有较强的语义相关性,并基于事件类在时空维度上的出现权重,采用有限状态机-高斯分布模型识别时空突发事件。

1 研究方法

本文基于时空要素综合分析的框架,提出一种新颖的突发事件检测方法。该方法首先利用爬虫技术获取大规模的未标注网络文本数据,并通过数据预处理获取时间表达式、地名实体和事件词。其次基于事件词的时空特性,采用数据立方体存储事件词。然后提出一种基于语义相似性的实时事件聚类算法抽取出重要事件。最后基于事件在时空维度上的出现权重,采用有限状态机-高斯分布模型,建模事件的突发特性。突发事件检测的系统框架如图 1 所示。

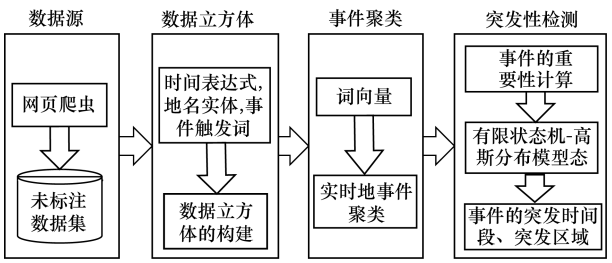

图 1 突发事件检测系统

1.1 数据立方体的构建

通过数据预处理,从网络文本中抽取出事件词、时间表达式和地名实体。

事件触发词是表达事件发生的性质或状态的词[18],例如“由于电池门问题,三星 Galaxy Note7 发生爆炸”,本文将事件触发词作为事件词。为了抽取出事件触发词,将事件触发词的识别视为一个二分类任务。首先随机选取 200 篇新闻文档作为训练语料,这些文档涵盖政治、社会、经济、体育、军事等领域。为了确保训练语料的可靠性,按照 Timebank Corpus[19]标注指导对语料进行人工标注。在众多的分类器中,CRF 模型考虑了文本的语境特征和词性特征,在序列标注任务和分类任务中能够取得较好的效果,因此本文采用 CRF(Conditinal Random Fields)模型[20]抽取出最合适的事件触发词。

一篇文档通常包含多个时间表达式、多个地名实体,新闻媒体或社交网络网站是一个实时报道当天事件的平台,本文将文档的生成时间作为事件词的发生时间,将距离事件词最近的地名实体作为该事件词的发生地点。为了将地名实体转换成空间信息,构建一个完善且全面的地理空间知识库,该知识库包括地名本体子库、规则子库等辅助数据源,并提供相应的查询接口。在地名-空间信息转换过程中,采用了地名消歧和地名经纬度转换等技术。地名消歧通过启发式的规则方法实现[21],通过计算地名和上下文地名之间的地理关联度进行地名的消歧,首先识别出文档中的所有地名,并确定歧义地名对应的所有地理位置,构成候选位置集合,然后设置启发式规则方法,从候选位置集合中确定唯一的地理位置。地名经纬度转换通过启发式的规则匹配方法实现。将事件词的时间信息和空间信息结合,即可获取事件词的时空信息。最后基于事件词的时空信息,将事件词存储于数据立方体中,如图 2 所示。

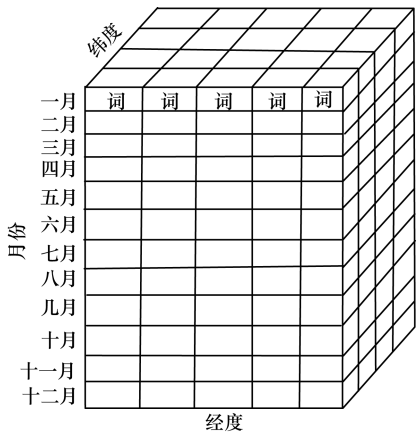

图 2 数据立方体示意图

1.2 基于语义相似性的事件聚类

在 1.1 节的基础上,由于事件词已存储于立方体中,但立方体的事件词是杂乱无章的,需要对这些事件词进行有效的聚类以抽取出重要事件。现有方法研究事件聚类通常采用 K-means 和 Latent Dirichlet Allocatio 等的改进方法[5-7],但它们都是离线的批处理聚类方式,不适用于动态的网络数据流。近年来,随着网络文本数据的兴起,研究者提出了许

多在线的聚类算法[22-24],但是当涉及到相似性计算时,这些方法通常只考虑词之间的空间距离,未挖掘词的语义关联性。

针对现有聚类方法存在的问题,本文提出一种基于语义相似性的实时事件聚类算法,该算法是一种增量式的聚类方式。随着数据流的到来,聚类结果将会动态地改变,该聚类算法如算法 1 所示。

算法 1　事件聚类(E,$\boldsymbol{w}$)

输入　词 $\boldsymbol{w}$,现有事件集 $E=\{e_1,e_2,\cdots,e_K\}$

输出　更新事件集 E

```
If E is null
  e1 = w, c1 = w
Else
  For each event ei in the E do
      Si = Sim(ci, w)
  Return the biggest Sb
  If Sb > threshold T then
    Add w to the existing event eb
    Update the center vector cb of event eb
      For word wi in the eb do
        cb = (1/k) Σ_{i=1}^{1} wi
Else
    add w to E as a new event
```

考虑一个新到达的事件词 $\boldsymbol{w}$,假如 $\boldsymbol{w}$ 是第一个到来的事件词,那么将其作为第一个事件类;否则,将 $\boldsymbol{w}$ 分别与已有的事件类进行相似性计算,然后对所有相似值做降序排序,获得最大的相似值 S_b,假设 S_b 为 $\boldsymbol{w}$ 与事件类 e_b 的相似值,如果 S_b 大于阈值 T,$\boldsymbol{w}$ 被聚到事件类 e_b 中,同时更新事件类 e_b 的质心向量 $\boldsymbol{c}_b$,否则 $\boldsymbol{w}$ 被作为一个新的事件类添加到事件集 E 中,算法 1 中的相似性计算采用余弦相似度公式:

$$\cos\theta=\boldsymbol{w}_i^{\mathrm{T}}\tilde{w}_j/(|\boldsymbol{w}_i||\tilde{w}_j|) \tag{1}$$

上述聚类算法的一个核心环节为事件词间的相似性计算。目前最流行的计算词相似性的方法为词向量的方式。已有的许多表证词的向量空间法,例如文献[25]提出一种全局向量模型(GloVe)训练词向量。GloVe 模型充分利用词的全局共现统计和语境特征来挖掘词之间的语义关联性,在语义相似性任务上,GloVe 模型的实验结果优于 Word2Vec 模型[26],因此,本文采用 GloVe 模型挖掘事件词之间的语义关联性。GloVe 模型的详细推导过程见文献[25]。

1.3　突发事件的检测

在突发性检测中,具有代表性的方法为文献[9]提出的有限状态机模型,该模型基于文档的到达时间间隔,使用有限状态机建模事件的突发性,从而识别出突发的开始时间和结束时间。该模型为一个隐马尔可夫链,模型的隐变量是词所处的状态(突发态或普通态),其假设文档的到达速率服从指数分布,当文档的到达速率加快时,模型会依据状态转换代价判定是否发生状态转换,通过对模型的状态序列进行推理最终获得一条最优的状态序列,序列中2 个时间点的状态改变代表着突发时间段的边界。文献[8]借鉴 Kleinberg 的思想,基于时间序列中话题的出现频率,假设话题的出现频率服从泊松分布,并采用有限状态机-泊松分布模型识别突发性话题。Kleinberg 和 Diao 的方法研究重点在于检测突发事件和突发时间段,未考虑事件的突发区域性,并且它们依据事件的频率信息进行突发性检测,忽略了事件的重要性。本文基于 Kleinberg 和 Diao 识别突发性的方法,提出综合分析事件的时间要素和空间要素,依据事件在时空维度上的出现权重,采用有限状态机-高斯分布模型建模事件的时空突发特性。

1.3.1　事件在时空维度上的重要性计算

现有方法通常依据特征项在时间序列上的出现频率,构建相应的模型判断事件是否为突发性事件。但是特征项的频率信息并不能有效地将某一个特征与其他特征区分开,即频率统计法并不具备很好的区分能力。事件间的重要程度有一定的差异,现有方法考虑事件的出现频率而忽略了事件的重要性,因此,无法有效突显事件的重要程度。词频反文档频率(TFIDF)则可克服该缺点,TFIDF 是一种有效体现特征重要性的值。TFIDF 的思想是:如果词 $\boldsymbol{w}$ 在某一类别中出现的频率高,而在别的类别中出现的频率低,则说明该词能够很好地代表该类别的特征,即可以有效地将某一类别与别的类别区分开。

本文采用 TFIDF 计算事件在时间维度、空间维度上的出现权重,用以评估事件在整个事件集中的重要程度。对于事件集 $E=\{e_1,e_2,\cdots,e_i,e_N\}$ 中的事件 e_i,计算其在不同的地理位置 r,不同的单位时间点 t 上的权重 $Weights(e_i,t,r)$。其中,$t\in[1:T]$ 为时间序列中某个单位时间点,$r\in[1:R]$ 为空间区域中某个地理位置。假设一个事件 e_i 由 K 个事件词 $\{w_1,w_2,\cdots,w_j,w_k\}$ 组成,考虑事件元素 w_j,令 $Weights(w_j,t,r)$ 为事件词在单位时间点 t、地理位置 r 上的权重值,则有:

$$Weights(e_i,t,r)=\sum_{j=1}^{K}Weights(w_j,t,r) \tag{2}$$

1.3.2　事件突发性的检测

本文提出采用有限状态机-高斯分布模型对事件的状态进行建模。该模型是一个隐马尔可夫链,模型中的隐变量是词所处的状态,观测数据是事件在时间序列上单位时间点的权重值。该有限状态机模型如图 3 所示,其中,q_i 为自动机的隐状态,“0”代表正常态,“1”代表突发态,模型处在不同的隐状态,就以不同强度的概率来生成观测数据,即状态转移

链的发射概率服从高斯分布。

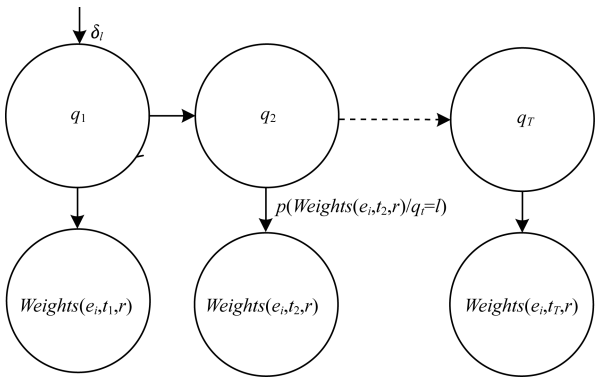

图 3 有限状态机模型

$$p(Weights(e_i,t,r)/q_t=l)=\frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{(Weights(e_i,t,r)-u_l)^2}{2\sigma_l^2}\right) \quad (3)$$

其中，q_t 为事件在单位时间点 t 的状态，$l=0$ 或者 $l=1$，$q_t=0$ 为正常态，$q_t=1$ 为突发态。高斯分布的 4 个参数为 u_0、u_1、σ_0、σ_1。设置 u_0 为事件在时序上的权重均值：

$$u_0=\frac{1}{T}\sum_t Weights(e_i,t,l) \quad (4)$$

其中，设置 $u_1=3u_0$，σ_0 为事件在时序上的权重均方差，$\sigma_1=\sigma_0$。

状态序列 $Q=\{q_1,q_2,\cdots,q_T\}$ 为状态机的状态转移链，其转移规律服从隐马尔科夫假设，由状态转移矩阵 $\boldsymbol{M}$ 和先验概率 θ 控制。在状态 q_1 之前，假设有一个虚拟的正常态 q_0，则状态机的先验概率为 $\theta_l=(p_{00},p_{01})$，状态转移矩阵为：

$$\boldsymbol{M}=\begin{bmatrix}p_{00} & p_{01}\\ p_{10} & p_{11}\end{bmatrix},p_{ij}=p(v_t=j/v_{t-1}=i) \quad (5)$$

其中，设置超参数 $\theta_0=0.7$，$\theta_1=0.6$。

采用维特比算法获取最优的状态转移序列 Q^*。序列中的突发态对应的连续时间段为突发时间段。对于突发时间段 $T=[t_1:t_2]$，其突发权重为：

$$Bursty(e_i,T)=\sum_{t=t_1}^{t=t_2}(p(Weights(e_i,t,r)/q_t=1)-p(Weights(e_i,t,r)/q_t=0)) \quad (6)$$

为了识别出合理的突发时空区域，采用矩形 R 表征事件的突发空间区域，时空窗 W 表征事件的突发时空域。定义事件 e 在突发时间段 T 矩形区域 R 上的突发权重值为事件词落在时间段 T 和矩形 R 上的突发权重值之和，并取多个区间的交叠区段为事件的突发时空域，突发权值为多个区间的权重值之和。事件 e 在时间序列和空间区域上的突发区间如图 4 所示，突发区间在时序上是非交叠的，而在空间区域上存在着交叠。对于突发时间段 $T=[t_1:t_2]$、突发区域 $R=[r_1:r_2]$，获取事件的突发时空窗权重分数为：

$$w-score(e_i,T,R)=\sum_{r=r_1}^{r=r_2}Bursty(e_i,T) \quad (7)$$

通过式(7)可获取任意时空窗的权重分数，对权重分数排序，即可获取 Top-rank 突发事件。

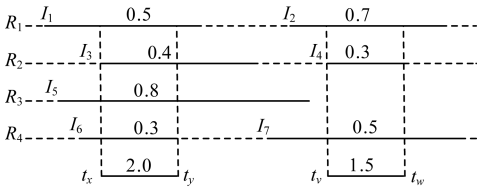

图 4 事件在多个地理位置上的突发时间段示意图

2 实验结果与分析

2.1 数据集与实验设置

采用网络爬虫技术抓取 2015 年 3 月 1 日—2015 年 8 月 30 日的 121 篇、157 篇新闻文档。这些文档涵盖政治、经济、体育等领域。通过数据预处理，获取 184 个事件的发生时间、7 494 个地名实体和 10 022 个事件词，然后基于事件词的时空信息构建立方体。在事件聚类中，基于数据集的相似性统计分析，设置相似度阈值为 0.76。在事件突发性检测中，设置时序上的单位时间为 d。

2.2 事件抽取结果与分析

2.2.1 对比方法

为了证明本文提出的事件抽取方法的有效性，设置基于 StreamCube 方法[27] 和 DTM (Dynamic Topic Models) 模型[28] 的对比实验。StreamCube 方法基于层级时空的 hashtags 聚类实现事件搜索，该方法将 hashtags 作为事件词，考虑了 hashtags 之间的时空关联性，采用在线的聚类算法实现事件搜索。在聚类过程中，StreamCube 采用 one-hot 模型表征词的向量空间，因此未能充分挖掘 hashtags 之间的语义相似性。DTM 是一种离线的主题生成模型，旨在研究基于时间维度的话题演化过程，体现话题随时间变化的特性。DTM 关注了话题随时间变化的演化过程，但是它忽略了话题的空间特性。

2.2.2 评价分析

本文引入 3 个评价聚类质量的指标：*NMI* (Normalized Mutual Information)，*RI* (Rand Index) 和 F_1 值。这 3 个评价指标的含义及计算公式如下所示。

$$NMI(X,Y)=2\times I(X,Y)/(H(X)+H(Y)) \quad (8)$$

其中,$I(\boldsymbol{X},\boldsymbol{Y})$为向量 $\boldsymbol{X}$ 与向量 $\boldsymbol{Y}$ 的互信息,$H(\boldsymbol{X})$为向量 $\boldsymbol{X}$ 的信息熵,同理,$H(\boldsymbol{Y})$为向量 $\boldsymbol{Y}$ 的信息熵。

$$RI=(TP+TN)/(TP+FP+FN+TN) \quad (9)$$

$$F_1=2TP/(2TP+FP+FN) \quad (10)$$

表 1 列举了每种方法的测评结果,StreamCube 方法在聚类过程中,采用 one-hot 模型表征事件词的词向量,即只考虑事件词之间的空间距离,没有挖掘出事件词的语义关联性,因此聚类效果最差。另外,one-hot 模型产生的将是一个高维度的稀疏共现矩阵,容易导致维数灾难的问题。DTM 对随着时间变化的文档集进行主题建模,由文档-词语-主题的生成过程判明出时间片段内文档所包含的主题。从聚类结果可以看出,DTM 可以较为有效地抽取出文档所包含的事件类。但是 DTM 需在整个数据集上迭代计算,是一种离线的抽取方式,因此并不能有效地处理动态的网络数据流。另外,DTM 忽略了话题的空间概念,无法处理事件的空间信息。本文事件抽取方法采用 Glove 模型训练事件词之间的语义相关性,使聚在同一事件类的事件词具有强的语义关联性,因此聚类效果优于 StreamCube 方法和 DTM 方法。另外,本文方法能够用较少的向量维度(200 维、300 维、400 维等)表征事件词的向量空间,因此占用较少的内存空间和聚类时间。

表 1　3 种方法的事件聚类效果

方法	*NMI*	*RI*	F_1
StreamCube	0.550	0.677	0.541
DTM	0.725	0.768	0.687
Our Method	0.803	0.835	0.774

2.3　突发事件检测结果与分析

2.3.1　对比方法

为了证明本文提出的突发事件检测方法的有效性,与 Diao 的方法进行对比,Diao 的方法旨在研究从微博数据流中发现突发性话题,其通过结合用户对话题的关注度以及话题在时序上的出现频率,采用基于有限状态机-泊松分布模型检测出突发性话题。

2.3.2　评价分析

采用本文的突发事件检测方法进行实验,列举了 Top-5 突发事件的实验结果,其中,每个事件列举了 Top-8 个事件词,如表 2 所示。可以看出,所有的突发事件都是有意义的,这些突发事件不仅具有一定的突发时间段,而且还具有一定的突发区域。另外,不同突发事件的突发时间段和突发区域都是不同的,表明了突发时空特性的重要性。

表 2　突发事件检测结果

突发时间段	突发区域	Top-8 事件词	事件类型
2015-04-20—2015-04-23	上海,北京,广东	减退,上涨,消散,衰退,涨跌,预测,背离,拖累	股票
2015-08-12—2015-08-19	上海,辽宁	泄漏,倒塌,中毒,抢修,燃烧,损坏,爆炸,滑坡	自然灾害
2015-08-27—2015-08-30	中国,日本	抗战,铭记,阅兵,参访,缅怀,抗日,参演,演习	国际关系
2015-03-04—2015-03-05	上海,四川	排放,关停,排污,防治,治污,燃煤,监测,印染	环境保护
2015-04-23—2015-04-25	江西,广东,云南	跳楼,反抗,营救,辱骂,救起,幸存,下跪,杀死	刑事案件

设置基于 Diao 的方法的对比实验。图 5 和图 6 分别为自然灾难事件基于时间序列的事件强度变化过程,其中,图 5 为 Diao 的方法基于事件在单位时间内的出现频率以及,建模有限状态机-泊松分布模型获取的事件强度变化过程。图 6 为 STBEvent 模型中基于事件的 TFIDF 权重以及建模有限状态机-高斯分布模型获取的事件强度变化过程。从图 5、图 6 可以看出,采用 Diao 的方法检测出该自然灾害事件有 4 个异常高频段,模型认为此事件并非一个突发事件,而是一个周期性事件。而采用 STBEvent 模型可正确检测出一个异常高频段,并认为其是一个突发事件。因此,采用 STBEvent 模型检测事件的突发性更为有效。

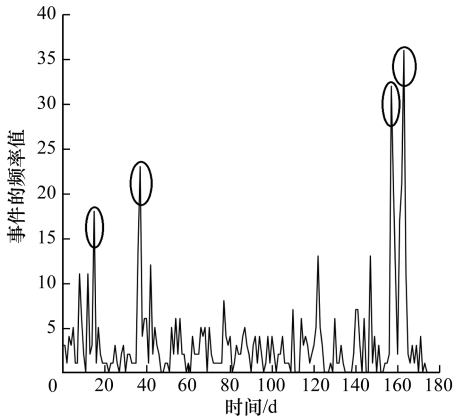

图 5　采用有限状态机-泊松分布模型获取的坍塌事件强度

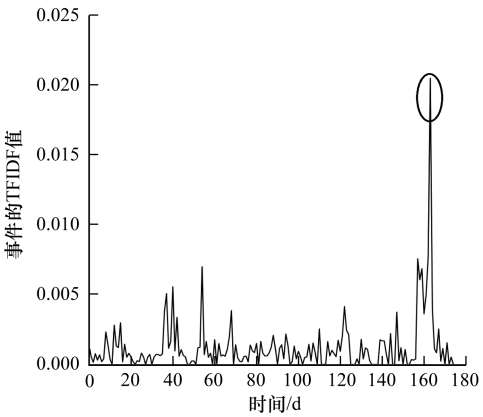

图 6　采用有限状态机-高斯分布模型获取的坍塌事件强度

图 7 为采用有限状态机-泊松分布模型获取的坍塌灾难事件(突发事件)和体育竞技事件(非突发事件)分别基于时间序列的事件强度变化过程,其中,实线为自然灾难事件的事件强度变化过程,虚线为体育竞技事件的事件强度变化过程图。图 8 为采用 STBEvent 基于事件的 TFIDF 权重,以及建模有限状态机-高斯分布模型获取的坍塌灾难事件(突发事件)和体育竞技事件(非突发事件)分别基于时间序列的事件强度变化过程,其中,实线为自然灾难事件的事件强度变化过程,虚线为体育竞技事件的事件强度变化过程。Diao 的方法对于突发事件,其与普通事件的频率分布并不具有很强的区分性。而 STBEvent 模型,对于坍塌灾难事件,在非突发态,其 TFIDF 值是低的;在突发态,其 TFIDF 值骤然增高,并急剧降低,符合突发事件的定义,这表明了 STBEvent 模型检测出的突发性事件与普通事件具有更为明显的区分性。

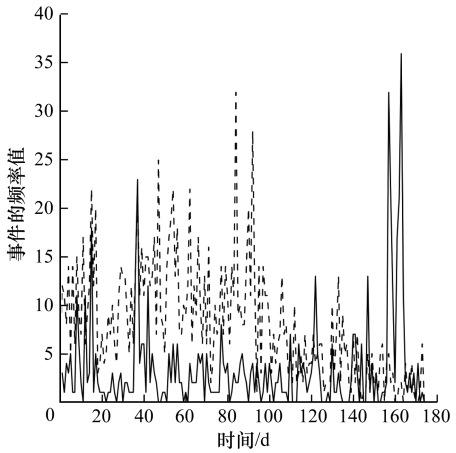

图 7 采用有限状态机-泊松分布模型获取的坍塌事件(突发事件)与体育竞技事件(非突发事件)强度

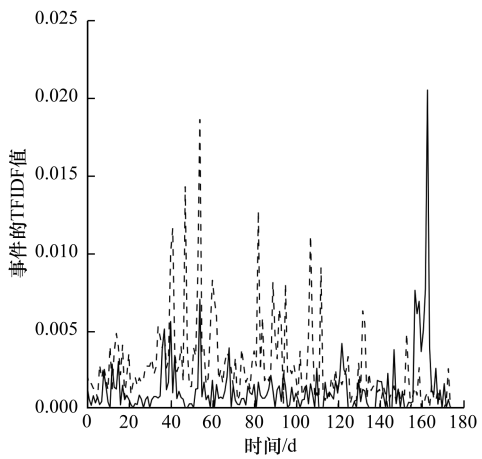

图 8 采用有限状态机-高斯分布模型获取的坍塌事件(突发事件)与体育竞技事件(非突发事件)强度

本文方法不仅能识别出突发时间段,而且可以识别出突发空间区域。图 9 为坍塌事件(突发事件)在不同地理位置序号的 TFIDF 值变化情况,图 10 为体育竞技事件(非突发事件)在不同地理位置序号的 TFIDF 值变化情况。可以看出,坍塌事件的突发区域为 3 个(上海、辽宁、山西),而体育竞技事件无明显的突发区域。

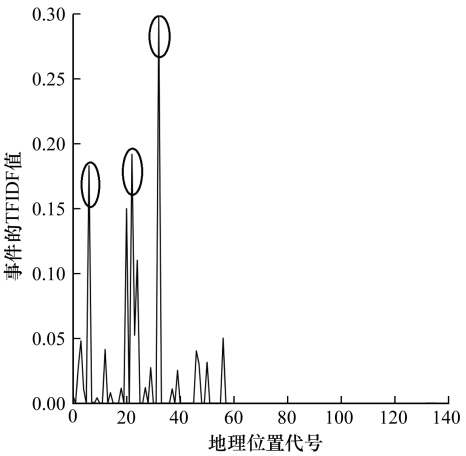

图 9 坍塌事件(突发事件)基于地理区域的权重值变化

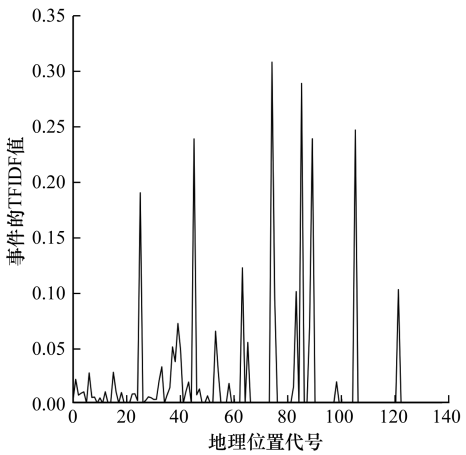

图 10 体育竞技事件(非突发事件)基于地理区域的权重值变化

3 结束语

传统的突发事件检测方法依赖人工标注数据集,以孤立的方式看待事件的时空要素,且忽略事件的重要性等问题。为此,本文提出一种基于时空要素综合分析的突发事件检测方法。该方法首先引入数据立方体结构存储事件词,综合分析事件的时空要素,并且挖掘出事件的时空关联性。然后给出一种基于语义相似性的实时事件聚类算法,实时地处理在线的动态网络数据流,从而摆脱了特定领域的限制。同时,采用 GloVe 模型挖掘出事件词之间的语义关联性,使聚在同一事件类的事件词具有强的语义相关性。其次采用 TFIDF 计算事件的出现权重,评估某一事件在整个事件集中的重要程度。最后采用有限状态机-高斯分布模型识别出时空突发事件。实验结果表明,该方法能够较为准确地抽取出重要的事件,并取得 77.4% 的抽取准确率;在突发性检测时,该方法比现有方法更能准确地检测出突

发事件,且能够有效地识别出事件的突发时间段和突发空间区域。下一步将研究事件抽取和突发性检测的联合学习算法。

参考文献

[1] BETHART S, MARTIN J H. Identification of event mentions and their semantic class[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: [s. n.], 2006: 146-154.

[2] LI P, ZHOU G, ZHU Q. Minimally supervised Chinese event extraction from multiple views [J]. ACM Transactions on Asian and Low-resource Language Information Processing, 2016, 6(2): 13.

[3] NGUYEN M T, NGUYEN T T. Extraction of disease events for a real-time monitoring system [C]// Proceedings of Symposium on Information and Communication Technology. Washington D. C., USA: IEEE Press, 2013: 139-147.

[4] 侯立斌,李培峰,朱巧明. 基于 CRFs 和跨事件的事件识别研究[J]. 计算机工程,2012,38(24):191-195.

[5] TSOLMON B, LEE K S. An event extraction model based on timeline and user analysis in latent dirichlet allocation[M]. New York, USA: ACM Press, 2014.

[6] SILVA J D A, HRUSCHKA E R. A support system for clustering data streams with a variable number of clusters[J]. ACM Transactions on Autonomous & Adaptive Systems, 2016, 11(2): 11.

[7] LIN C X, ZHAO B, MEI Q. PET: a statistical model for popular events tracking in social communities [C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2010: 929-938.

[8] DIAO Q, JIANG J, ZHU F, et al. Finding bursty topics from microblogs[C]//Proceedings of Association for Computational Linguistics. [S. 1.]: Association for Computational Linguistics, 2012: 536-544.

[9] KLEINBERG J. Bursty and hierarchical structure in streams [J]. Data Mining & Knowledge Discovery, 2003, 7(4): 373-397.

[10] LAPPAS T, ARAI B, PLATAKIS M, et al. On burstiness-aware search for document sequences[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2009: 477-486.

[11] ALVES R A D S, ASSUNCAO R M, STANCIOLI V D M P O. Burstiness scale: a parsimonious model for characterizing random series of events [C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2016: 1405-1414.

[12] KALOGERATOS A, ZAGORISIOS P, LIKAS A. Improving text stream clustering using term burstiness and co-burstiness [C]//Proceedings of Hellenic Conference on Artificial Intelligence. Athens, Hellenic: [s. n.], 2016: 1-9.

[13] ZHAO L, CHEN F, LU C T, et al. Online spatial event forecasting in microblogs [J]. ACM Transactions on Spatial Algorithms & Systems, 2016, 2(4): 15.

[14] SCHUBERT E, WEILER M, KRIEGEL H P. SPOTHOT: scalable detection of geo-spatial events in large textual streams[C]//Proceedings of International Conference on Scientific & Statistical Database Management. Washington D. C., USA: IEEE Press, 2016: 1-12.

[15] QUEZADA M, POBLETE B. Location-aware model for news events in social media [C]//Proceedings of International ACM SIGIR Conference. New York, USA: ACM Press, 2015: 935-938.

[16] LAPPAS T, VIEIRA M R, GUNOPULOS D, et al. On the spatiotemporal burstiness of terms[J]. Proceedings of the VLDB Endowment, 2012, 5(9).

[17] TAMURA K, MATSUI T, KITAKAMI H, et al. Identifying local temporal burstiness using MACD histogram [C]//Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. Washington D. C., USA: IEEE Press, 2015: 2666-2671.

[18] DODDINGTON G, MITCHELL A, PRZYBOCKI M, et al. The automatic content extraction program-tasks, data, and evaluation [C]//Proceedings of LREC' 04. Washington D. C., USA: IEEE Press, 2004: 158-165.

[19] PUSTEJOVSKY J, HANKS P, SAURI R, et al. The timebank corpus [C]//Proceedings of Corpus Linguistics Conference. Washington D. C., USA: IEEE Press, 2003: 215-222.

[20] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence Data [J]. Machine Learning 2002, 3(2): 282-289.

[21] 马雷雷,李宏伟,连世伟,等. 地名知识辅助的中文地名消歧方法[J]. 地理与地理信息科学,2016,32(4):5-10.

[22] SILVA J A, FARIA E R, BARROS R C, et al. Data stream clustering: a survey [J]. ACM Computing Surveys, 2014, 46(1): 13.

[23] 蔡偃武. 面向大规模数据的在线新事件检测[D]. 上海:华东理工大学,2014.

[24] YIN J, WANG J. A text clustering algorithm using an online clustering scheme for initialization [C]// Proceedings of ACM SIGKDD International Conference. New York, USA: ACM Press, 2016: 1995-2004.

[25] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Washington D. C., USA: IEEE Press, 2014: 1532-1543.

[26] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2013-01-12]. https://www.mendeley.com.

[27] FENG W, ZHANG C, ZHANG W, et al. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream[C]//Proceedings of IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2015: 1561-1572.

[28] BLER D M, LAFFERTY J D. Dynamic topic models[C]// Proceedings of DBLP' 06. Washington D. C., USA: IEEE Press, 2006: 113-120.

编辑　索书志