Journal of Computer Applications
计算机应用,2022,42(9):2667 - 2673
ISSN 1001-9081
CODEN JYIIDU
2022-09-10
http://www.joca.cn

文章编号:1001-9081(2022)09-2667-07
DOI:10.11772/j.issn.1001-9081.2021071330

融合社会影响力和时间分布的微博关键事件抽取方法

赵旭剑*,王崇伟,王俊力
(西南科技大学 计算机科学与技术学院,四川 绵阳 621010)
(*通信作者电子邮箱jasonzhaoxj@gmail.com)

摘　要:针对现有微博事件抽取方法由于基于事件的内容特征,而忽略事件本身的社会属性与时间特征之间的关系,进而无法识别微博热点传播过程中关键事件的问题,提出了一种融合社会影响力和时间分布的微博关键事件抽取方法。首先通过建模社会影响力来刻画微博事件的重要性,然后融合微博事件演化过程中的时间特性以捕获事件在不同时间分布下的差异,最后抽取出不同时间分布下的微博关键事件。在真实数据集上的实验结果表明,所提方法能有效抽取微博热点中的关键事件,较随机选择、词频-逆文本频率(TF-IDF)、最小权重支配集以及度与聚集系数这四种方法在事件集的完整性指标ROUGE-1上在数据集1上分别提升了21%、18%、26%以及30%,在数据集2上分别提升了14%、2%、21%以及23%,抽取效果优于传统方法。

关键词:社会影响力;时间分布;微博;事件抽取;事件演化
中图分类号:TP391.1　　**文献标志码**:A

Key event extraction method from microblog by integrating social influence and temporal distribution

ZHAO Xujian*, WANG Chongwei, WANG Junli
(*School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang Sichuan* 621010, *China*)

Abstract: Aiming at the problem that the existing microblog event extraction methods are based on the content characteristics of events and ignore the relationship between the social attributes and time characteristics of events, so that they cannot identify the key events in the propagation process of microblog hot spots, a key event extraction method from microblog by integrating social influence and temporal distribution was proposed. Firstly, the social influence was modeled to present importance of microblog events. Secondly, the temporal characteristics of microblog events during evolution were considered to capture the differences of events under different temporal distributions. Finally, the key microblog events were extracted under different temporal distributions. Experimental results on real datasets show that the proposed method can effectively extract key events in microblog hot spots. Compared with four methods of random selection, Term Frequency-Inverse Document Frequency (TF-IDF), minimum-weight connected dominating set and degree and clustering coefficient information, the proposed method has the event set integrity index improved by 21%, 18%, 26% and 30% on dataset 1 respectively, and 14%, 2%, 21% and 23% on dataset 2 respectively. The extraction effect of the proposed method is better than those of the traditional methods.

Key words: social influence; temporal distribution; microblog; event extraction; event evolution

0　引言

社交网络平台的开放性与社交性导致其迅速发展,与之对应的则是爆炸式的微博数据增长[1]。微博中突发社会事件以社交网络或新闻网站为传播载体,经过传播发酵产生社会热点事件,随着时间推移与事态发展,热点事件形成动态演化,在各个时间戳上产生不同的关键信息,其中蕴含着事件间错综复杂的发展演化关系。尽管微博拥有丰富的社会热点事件资源,但面对海量数据,用户却难以捕获社会热点事件中的各个演化阶段的关键信息。作为社交网络的代表性平台,新浪微博为事件的传播作出了巨大贡献,但从社会网络信息传播的角度看,微博的“转发”特性带来大量冗余信息也导致了信息泛滥,因此,社会热点事件的过滤筛选和关键事件抽取对用户了解社会热点、追踪热点演化具有重要意义[2-4]。此外,从社会热点事件中提取关键事件,对决策者分

收稿日期:2021-07-26;**修回日期**:2021-09-14;**录用日期**:2021-09-15。　**基金项目**:教育部人文社科基金资助项目(17YJCZH260);四川省科学技术厅重点项目(2020YFS0057);赛尔网络下一代互联网技术创新项目(NGII20180403)。

作者简介:赵旭剑(1984—),男,四川绵阳人,副教授,博士,CCF会员,主要研究方向:文本挖掘、自然语言处理、Web信息处理;　王崇伟(1995—),男,四川泸州人,硕士研究生,主要研究方向:信息抽取、机器学习;　王俊力(1996—),男,四川南充人,硕士研究生,主要研究方向:信息抽取、机器学习。

析舆情态势、引导社会舆论等同样具有研究意义。

目前，社会热点事件的相关研究工作以事件内容特征为基础，传统方法中词频–逆文本频率（Term Frequency-Inverse Document Frequency，TF-IDF）模型是衡量文本重要性的最常用模型，但 TF-IDF 模型仅能在词语级别揭示事件的重要性。最近有研究人员基于贝叶斯网络对网络舆情事件分析[5]，也有一些研究提出将事件建模为图结构，利用支配集算法或计算图节点的度与聚集系数，从图论的角度考虑节点的重要性，从而提取关键事件[6-7]，但构建图需要丰富的事件语义信息，而这是微博数据所缺乏的。总之，以上研究都忽略了微博环境下事件的传播对关键事件的影响。

为解决上述问题，本文提出一种融合事件社会影响力和时间分布的微博关键事件提取方法。首先，基于微博中事件特征对事件的社会影响力建模；然后，基于事件演化的时间分布特征，提取不同时间分布下的关键事件；最后，基于真实微博数据集的实验表明，本文方法能有效提取社会热点事件各演化阶段的关键事件。本文主要工作如下：

1）提出了一种建模微博事件重要性的方法。通过建立与事件主题相关的社会影响模型，挖掘微博事件重要元素，构建基于微博社会影响力的事件重要性评价模型。

2）建立融合事件社会影响力和时间分布的微博关键事件检测模型。基于事件社会影响力，融合微博事件演化过程中的时间特性以捕获事件在不同时间分布下的差异，并检测各演化阶段的关键事件。

3）在两个真实微博数据集上对本文提出的抽取方法进行了实验验证并构建了一个微博关键事件抽取系统，实验结果表明，所提方法能有效抽取微博热点中的关键事件，抽取效果优于传统方法。

1 相关研究

面向社交网络的数据挖掘是目前 Web 文本挖掘的重要研究方向。针对微博的数据挖掘分析一般包含话题事件挖掘、情感分析和网络舆情分析等。其中话题事件挖掘包括高质量信息抽取[8]和事件演化挖掘[9-10]等，而目前高质量信息抽取侧重于事件抽取[11]和事件摘要[12]，有别于本文工作所关注的关键事件抽取研究。当前社会热点中关键事件提取方法可分为基于传统内容特征的方法、基于图的方法和基于机器学习的方法，下面将简述不同方法的特点与不足。

1）基于传统内容特征的方法。该类方法利用事件内容特征对事件进行评价排序，通过得分排名提取关键事件。如欧阳逸等[11]计算事件中关键词的 TF-IDF 得分，将关键词得分之和作为事件得分，提取得分排名靠前的事件作为关键事件；彭敏等[8]将事件多特征融合并转换到小波域捕获事件间的细节差异，并引入核主成分分析进行特征变换提取关键事件；夏立新等[13]利用事件热点划分舆情关系，基于 TextRank 算法提取事件关键词和关键事件的文本摘要，最后建立事理图谱并可视化事件摘要。

2）基于图的方法。该类方法基于事件内容特征关系把事件建模为图结构，利用图算法，将提取关键事件转化为提取图中关键节点。如李培等[6]基于相似性将微博建模为多视点图，利用最小权重支配集求解重要节点以提取关键事件，并引入 Top-K 集缓解微博数据量巨大的问题；Yuan 等[7]引入度与聚集系数[14]评价图中节点重要性提取关键事件。

3）基于机器学习的方法。该类方法利用机器学习算法对热点事件建模学习，实现关键事件提取。如田世海等[15]融合网络表示学习与 K 均值聚类算法，将舆情事件用低维向量表示，聚类得到舆情事件；李进华等[16]使用 K 均值聚类算法、K 最近邻分类算法和决策树三类方法建模微博事件的地理特征，检测提取不同地理位置的关键事件。

上述方法从事件内容特征的角度对事件重要性进行评价，并引入了外部模型（例如图算法）进行算法增强，但忽略了事件传播对事件重要性的影响，彭敏等[8]虽然引入微博事件的传播行为特征，但复杂的数学变换将导致巨大的时间开销。基于机器学习的方法虽然对事件特征进行细粒度建模，但忽略了社交网络中的事件特性。本文利用事件的社会影响力弥补基于内容特征方法的不足，此外引入事件时间分布，最大限度保证抽取事件在时间线上分布的合理性，提升关键事件抽取精度。

2 框架概述

2.1 相关定义

定义 1 热点事件。热点事件 E 指社会突发事件在传播介质中经过传播发酵和演化发展形成的具有一定演化阶段的事件聚合体。

定义 2 关键事件。关键事件 e 指热点事件 E 在其演化过程中，各时间戳上最具代表性的事件，是组成热点事件的基本单位。因此，上述热点事件 E 可形式化定义为 $E=\{e_1,e_2,\cdots,e_i,e_{i+1},\cdots,e_n\}$，其中 e_i 表示 E 在第 i 个时间戳上最具代表性的事件。

2.2 框架结构

为有效提取社会热点中的关键事件，反映社会热点事件的发展演化过程，本文提出一种融合微博事件社会影响力和时间分布的关键事件抽取方法，主要包括如下 4 个步骤，如图 1 所示。

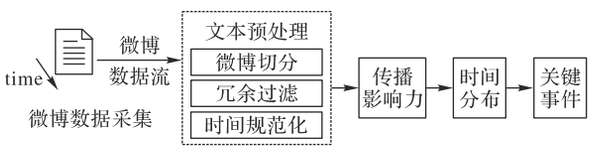


图 1 融合社会影响力和时间分布的微博关键事件提取框架

Fig. 1 Framework of key event extraction integrating social influence and temporal distribution

1）微博数据采集，基于微博的“话题”标签对热点事件的数据进行采集。

2）文本预处理，对微博数据进行切分、冗余过滤、时间表达式规范化等。

3）社会影响力建模，基于微博的社会影响力特征建立事件重要性评价模型。

4）时间分布模型，分析事件演化的时间分布，捕获不同

时间戳上的关键事件。

3　微博数据处理

3.1　数据采集

为采集相关社会事件微博数据，利用微博的“话题”标签，能够有效提升检索事件帖子的相关性。根据微博的搜索工具（https://s.weibo.com）对相关话题进行检索，基于Scrapy爬虫框架捕获相关数据。图2展示了微博数据采集以及储存的设计流程。对于初始化层，首先确定研究的社会热点事件，分析事件核心词和起止时间；在业务层，利用核心词和时间构建爬虫URL地址池，基于Scrapy爬虫框架模拟用户登录并解析页面中的微博帖子数据；最后，将数据规范化并储存。通过爬虫解析并储存到本地的微博数据主要包括微博发布者、微博原始文本、发布时间、转发量、评论量、点赞量和原文链接等核心数据。

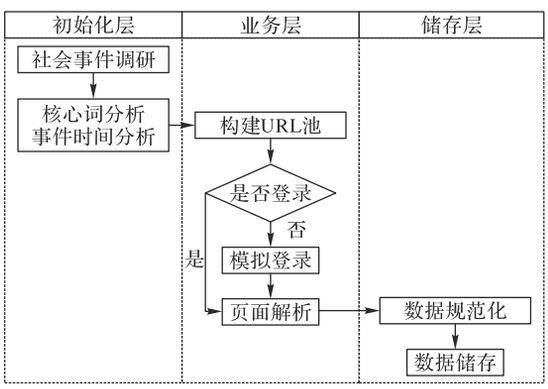


图2　数据采集流程

Fig. 2　Flowchart of data acquisition

3.2　文本预处理

对于微博集合，本文的预处理模块主要包含微博切分、冗余过滤和时间规范化3个步骤。图3展示了文本预处理过程。

微博切分　每一个完整的微博句子都能表示一种完整的事件语义信息，因此基于表达句子结尾意思的标点符号对微博帖子进行切分，得到大致的事件集合，然后考虑将每一个包含时间表达式的微博帖子视为一个事件。事件文本中的链接将被移除，并且少于25个字符的事件将被舍弃，这些事件构成了最初的事件集合。

本文用一个二元组$\langle C, T_e \rangle$表示具体的事件，事件的具体存储格式为“[时间戳]事件文本”，例如“[2018年07月11日]2018年07月11日，由长春长生生物公司内部生产车间的老员工实名举报，一个惊天疫苗造假案震惊寰宇，同时‘疫苗之王’浮出水面。”。此外，本文句子太长或太短都不利于用户理解内容，因此将句子长度量化为可理解性权重（Understandability Weight，UW），由式（1）计算：

$$UW(e_i) = \varepsilon^{-\left(2*Norz(len(e_i))-\frac{1}{2}\right)^2/2} \tag{1}$$

其中：$Norz\left(len(e_i)\right)$为归一化事件$e_i$的内容长度，$\varepsilon$指自然常数。

冗余过滤　由于微博的“社交”特性，转发将导致大量的冗余微博，使事件集合存在大量冗余事件导致信息泛滥。此外，同一社会事件不同发布人员可能包含相似的文本内容，这部分重复数据也应该考虑删除。因此为了有效过滤冗余帖子，基于显式相似度对事件集合进行冗余过滤，提出了使用两层相似度衡量事件相似性，分别是句子层面和事件集合层面：句子层面的相似度由最长公共子串计算，可以有效去除由转发带来的重复微博；事件集合层面的相似度利用TF-IDF算法将事件文本表示为向量，通过计算事件向量的余弦相似性得到，能有效去除由相同事件带来的重复微博。最后总体相似度由式（2）计算：

$$Sim(e_i, e_j) = \alpha \cos(\boldsymbol{v}_{e_i}, \boldsymbol{v}_{e_j}) + (1-\alpha) Sim_{\text{substr}}(d_i, d_j) \tag{2}$$

其中：$\boldsymbol{v}_{e_i}$和$\boldsymbol{v}_{e_j}$分别是事件e_i和e_j的向量表示，d_i和d_j分别是事件e_i和e_j的原始文本，α为权重系数。

基于事件间的总体相似度，利用增量聚类的思想形成事件集合，此时每个事件集合中事件内容高度相似，保留每个事件集合中可理解性权重最大的事件，由此得到了低冗余度的事件集合。

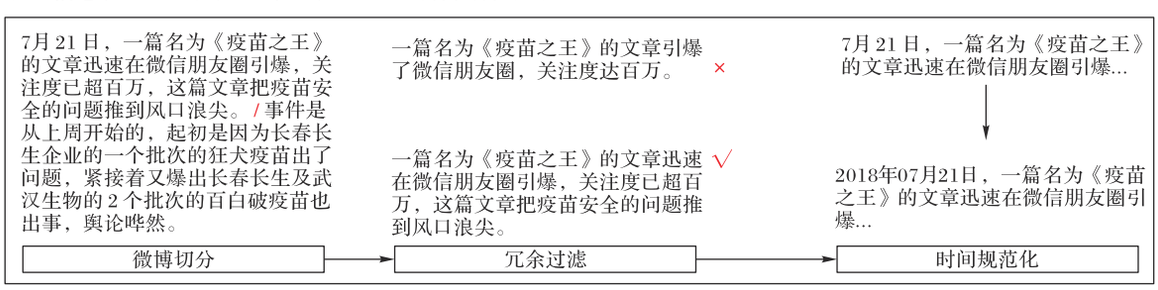


图3　文本预处理流程

Fig. 3　Flowchart of text preprocessing

时间规范化　时间是事件最重要的特征之一，微博事件中时间表达式主要可分为两类[17]，分别为显式时间表达和隐式时间表达，其中显式时间表达指直接的时间戳描述，而隐式事件表达指间接的时间戳描述。本文沿用其中对时间概念的定义，并对时间进行了细粒度划分，把显式时间表达划分为完整显式时间表达和模糊显式时间表达，具体信息如表1所示。时间表达式规范化主要是将模糊显式时间表达和隐式时间表达进行规范，将其统一为完整显式时间表达的格式。

此外，通过数据分析发现微博中社会事件的时间精度往往较低，因此本文仅将事件的时间精度精确到“日”级别。对于显式时间表达，利用正则表达式对事件中的所有显式时间

进行识别，设定了两种模糊粒度，分别为(a)“X 月 X 日”和(b)“X 日”，并提取其时间表达式，采用一种顺序匹配的方法对事件中的模糊时间进行补全。首先，经过时间表达式匹配得到了每个事件中的所有时间表达式集合，并把该事件微博的发表时间作为基准时间；然后，遍历时间集合中的时间表达式，用基准时间补全每次遍历到的时间戳(如果该时间戳为模糊时间的话)；接着，将新补全的时间表达式作为新的基准时间，继续遍历时间集合，直到集合遍历完毕。算法 1 展示了对每个事件包含的模糊时间表达式规范化过程。

表 1 时间表达式分类

Tab. 1 Classification of time expressions

类别	细粒度分类	示例
显式时间表达	完整显式时间表达	2020 年 7 月 3 日
	模糊显式时间表达	7 月 3 日
隐式时间表达	—	昨天、上周五

算法 1 模糊时间表达式规范化。

输入 事件 te_i，时间戳集合 $T=\{t_1, t_2, \cdots, t_i, t_{i+1}, \cdots, t_n\}$。

输出 具有标准化时间表达式的事件 te'_i。

```
//将事件发布时间设为基准时间
baseTime ← sendTime
For t_x in T do
    //完整显式表达情况，更新基准时间
    if '年' in t_x then
        baseTime ← t_x
        //模糊粒度(a)情况，补全“年”，更新基准时间
    elseif '月' in t_x then
        t_x ← strcat(baseTime[0:4], t_x)
        baseTime ← t_x
        //模糊粒度(b)情况，补全“年”月“”，更新基准时间
    elseif '日' in t_x then
        t_x ← strcat(baseTime[0:7], t_x)
        baseTime ← t_x
Return te'_i, T
```

对于隐式时间表达，基于规则编写隐式时间表达式识别的正则表达式，通过建立时间映射规则将隐式时间表达式规范化。例如“今日”“今天”和事件原始微博的发布时间建立映射关系，而“昨日”“昨天”在建立映射关系的前提下，相较基准时间进行时间偏移。表 2 展示了部分高频隐式时间表达式的映射关系。

表 2 隐式时间表达映射

Tab. 2 Implicit time expression mapping

隐式时间表达	时间映射	隐式时间表达	时间映射
今日/今天	微博发布时间	次日/第二天	基准时间($day+1$)
当日/当天	基准时间	一月前	基准时间($month-1$)
昨日/昨天	基准时间($day-1$)	去年今日	基准时间($year-1$)

4 关键事件抽取

4.1 事件社会影响力模型

微博热点事件往往随时间在各个演化阶段产生事件内容的动态变化，导致传统的事件抽取方法不能准确提取各个演化阶段的事件信息；同时，用于构建热点事件演化集合的事件个体在时间维度上必须能全面地代表热点事件的演化信息。对于社交网络而言，意见领袖对微博社会事件传播具有更强的影响力，因为他们通常比普通用户传递更多的关键信息，因此，意见领袖发表的帖子更有可能成为具有代表性的事件。本文提出使用基于社会影响力的评价模型来衡量事件的代表性，利用微博的转发、评论和点赞来量化事件的代表性程度。如果一篇文章有更多转发、评论和点赞，那么本文认为该帖子包含了大量用户都能识别的基本信息。因此，与那些转发、评论和点赞相对较少的微博相比，这条微博将更有可能讨论具有代表性的事件。具体来说，事件的社会影响力(Social Influence，SI)可以用式(3)来表示：

$$SI(e_i)=UW(e_i)*\ln(\alpha\cdot f_n+\beta\cdot C_n+\gamma\cdot l_n+\varepsilon) \tag{3}$$

其中转发、评论和点赞的数量被定义为 f_n、C_n 和 l_n；α、β 和 γ 表示不同的影响力权重；ε 是自然常数，使计算得到的社会影响力大于 0 并且更加平滑。

4.2 融合社会影响力和时间分布的关键事件抽取

微博热点事件由许多关键事件构成，反映热点事件随时间的演变。对于同时发生的事件，用户的注意力是有限的，这意味着用户通常关注具有更大社会影响力的事件，因此，提取代表性事件需要考虑事件的时间分布。本文根据事件的时间分布，选取具有更大社会影响力的事件来表征关键事件。

关键事件序列在事件演化时间轴上的分布是较为分散的，过于集中在某个时间戳上的关键事件将无法反映事件演化的全部过程。通过考虑事件的社会影响力，对每一个时间戳的重要性程度加权，在每一个时间戳上提取具有更高社会影响力的事件，同时通过时间戳权重确定每个时间戳上提取的事件数量。对事件的时间序列 E_t，每个时间戳的权重 $IW(t_i)$ 定义为：

$$IW(t_i)=\sum_{i=1}^{m}SI(e_i) \tag{4}$$

即 t_i 时刻事件的社会影响力之和。同时，t_i 时刻提取的关键事件个数 $N(t_i)$ 被定义为：

$$N(t_i)=\left\lfloor\frac{IW(t_i)-Min\_IW}{Max\_IW-Min\_IW}*n\right\rfloor \tag{5}$$

式中：Min_IW 和 Max_IW 分别指所有时间戳上社会影响力的最小值和最大值，通过归一化计算得到每个时间戳的重要性程度加权；n 是一个常数，表示每个时间戳提取关键事件的最大值，通过实验本文 n 值取 2，最外层括号表示向下取整。

算法 2 描述了通过融合微博事件演化过程中的时间特性以捕获事件在不同时间分布下的差异，并检测各演化阶段关键事件的具体过程。首先对热点事件 $E=\{e_1, e_2, \cdots, e_i, e_{i+1}, \cdots, e_n\}$，记录每一个事件的时间戳信息，得到事件的时间戳序列 $E_t=\{t_1: [\cdots, e_{(i-1)}, e_i, e_{(i+1)}, \cdots], t_2: [\cdots, e_{(j-1)}, e_j, e_{(j+1)}, \cdots], \cdots\}$；然后遍历每个时间戳 t_i 中的事件，将 t_i 中具有最大社会影响力的事件加入关键事件集合 C 中，并从 t_i 中删除该事件；接着判断 t_i 时刻提取的关键事件是

否为n个,若小于n则重复上述过程直到提取事件个数满足条件;最后得到抽取的关键事件集合C。

算法2　关键事件提取。

输入　热点事件$E=\{e_1,e_2,\cdots,e_i,e_{i+1},\cdots,e_n\}$。

输出　关键事件抽取结果$C=\{e'_1,e'_2,\cdots,e'_j,e'_{j+1},\cdots,e'_m\}$。

1)　$C\leftarrow\{\}$　　//关键事件抽取结果
2)　$E_t\leftarrow\{\}$　　//不同时间戳下的事件字典
3)　For i in $E.size$ do
4)　　$timestamp\leftarrow E.\text{get}(i).\text{get\_time}()$　//记录不同时间戳的事件
5)　　$E_t.\text{get}(timestamp).\text{set}(E.\text{get}(i))$
6)　End For
7)　For t_i in $E_t.\text{size}$ do
8)　　$event\leftarrow E_t.\text{get}(t_i).\text{getMaxWeightEvent}()$
9)　　$C.\text{get}(t_i).\text{add}(event)$
10)　　$E_t.\text{get}(t_i).\text{remove}(event)$
11)　　If $C.\text{get}(t_i).\text{length}<n$
12)　　continue
13)　　Else
14)　　break while
15)　End For
16)　Return C

5　实验与分析

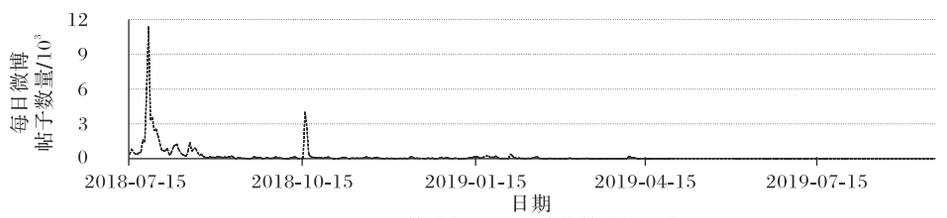
5.1　实验数据集

为了评估系统的性能,在新浪微博上收集了两个真实事件的微博数据集进行实验,通过特定事件相关的查询词,对包含这些查询词的微博帖子进行爬取。数据集详情如表3所示,图4显示了数据集中不同日期的帖子数量。

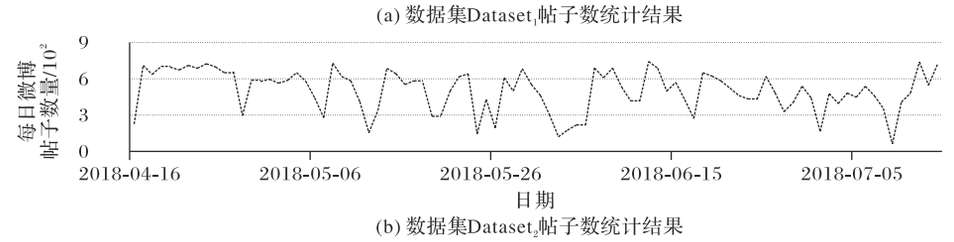
表3　数据集详情

Tab. 3　Details of datasets

数据集	标题	起止日期	数量
Dataset_1	疫苗事件	2018-07-01 — 2019-09-01	73 614
Dataset_2	中兴事件	2018-04-16 — 2018-07-14	45 113



(a) 数据集Dataset_1帖子数统计结果



(b) 数据集Dataset_2帖子数统计结果

图4　不同时间节点的帖子数量

Fig. 4　Number of posts at different time nodes

5.2　实验基准

为了评测事件提取方法的实验性能,将本文方法与4种基准方法进行了比较,如下所述:

1)随机选择(Random),从事件集合中随机选择关键事件,表示一种随机的思想。

2)词频–逆文本频率(TF-IDF)[11],计算TF-IDF分数,选择较高分数的事件作为关键事件。

3)最小权重支配集(Minimum-Weight connected Dominating Set, MWDS)[6],基于事件相似度构建图结构,利用最小权重支配集算法选择关键事件。

4)度与聚集系数(Degree and Clustering Coefficient Information, DCCI)[7],基于事件相似度构建图结构,基于度与聚集系数选择关键事件。

5.3　评价标准

本文的事件提取评价标准基于人工标注,邀请数据挖掘的研究人员从微博帖子中提取标准的关键事件,并使用完整性和冗余度作为事件提取的评价指标。完整性是指集合中的关键事件是否能充分反映热点事件随时间的演变过程。本文基于ROUGE(Recall-Oriented Understudy for Gisting Evaluation)来评估事件集的完整性[18]。具体使用ROUGE-1和ROUGE-L(ROUGE-Longest common subsequence)来评价事件集的完整性;此外,本文还用冗余度来评估事件集的重复信息。本文定义每个事件的冗余度是与集合中最相似事件的相似度,而事件集合的冗余度是每个事件的冗余度之和。这意味着每个事件与其他事件尽可能不同,事件集的冗余度将更小。其中冗余度(Redundancy)由式(6)计算:

$$Redundancy=\sum_{e_i,e_j\in E}\max\left(Sim(e_i,e_j)\right)\tag{6}$$

5.4　实验结果

表4展示了在两个不同数据集上的几种方法的性能比较结果。与其他方法相比,在完整性评估中,本文方法取得了最佳的ROUGE-1和ROUGE-L,这意味着本文提出的基于社会影响力和时间分布的方法能够从微博中准确提取关键事件。此外,本文方法在Dataset_1的冗余度评价中取得了第二名,在Dataset_2中取得了第一名,这意味着本文方法提取的关键事件几乎都是纯净的,即每个事件都能较好地表示热点

事件在不同演化阶段的差异。进一步分析，本文方法在 $Dataset_1$ 上的冗余度性能略差于 TF-IDF 算法，部分原因是 TF-IDF 算法考虑了事件中单词的特殊性，导致 TF-IDF 算法提取的关键事件是尽可能稀有的。而事实上，由于社交网络的转发特性，包含稀有词汇的事件社会影响很小，这恰恰与本文考虑的方法相反。

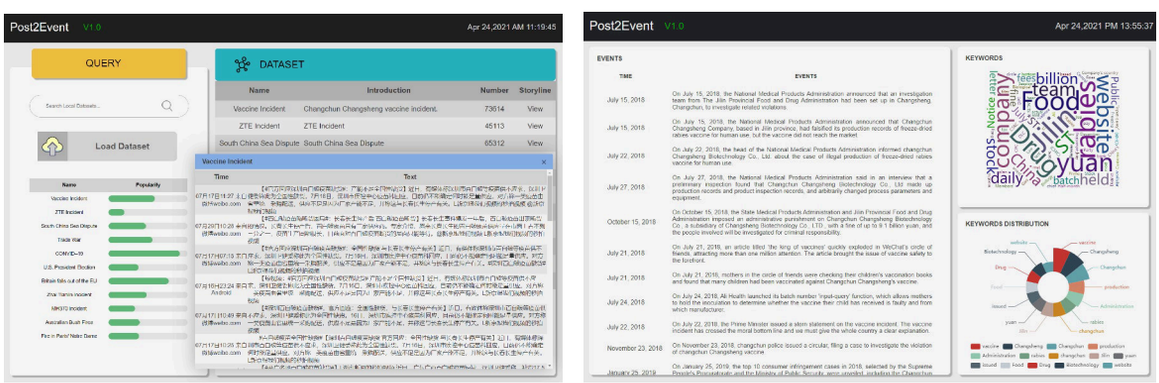
表 4 事件提取性能比较

Tab. 4 Comparison of event extraction performance

数据集	评价标准	方法				
		Random	TF-IDF	MWDS	DCCI	本文方法
$Dataset_1$	ROUGE-1	0.678 1	0.693 1	0.652 4	0.630 6	**0.820 6**
	ROUGE-L	0.218 3	0.210 6	0.320 8	0.345 6	**0.514 2**
	Redundancy	28.152 0	**20.824 1**	36.576 9	47.952 8	23.340 9
$Dataset_2$	ROUGE-1	0.728 9	0.817 7	0.684 7	0.675 1	**0.830 8**
	ROUGE-L	0.438 7	0.467 0	0.443 7	0.431 9	**0.704 1**
	Redundancy	31.806 2	39.201 4	41.484 7	44.356 1	**27.807 7**

5.5 原型系统

为了验证本文方法在微博环境下抽取关键事件的有效性，本文设计并开发了一个面向微博的微博事件抽取系统 Post2Event。如图 5 所示，Post2Event 主要包含两个对象模块：数据模块（图 5(a)）和事件模块（图 5(b)）。数据模块显示本地数据集以及数据集中各热点事件的微博热度。用户可以浏览本地数据集，并在界面中可视化每个数据的详细信息。此外，Post2Event 提供了一个查询接口，用户可以从数据库中检索相关的热点事件。在事件模块中，用户能得到系统自动抽取出的各个时间戳上的关键事件以及整个热点事件的关键词分布，让用户可以准确把握热点事件演化分支的主题方向。



(a) 数据模块 (b) 事件模块

图 5 Post2Event 系统快照

Fig. 5 Snapshots of Post2Event system

6 结语

社会热点事件爆发往往会引起数百万的微博讨论，针对新浪微博信息爆炸式增长的问题，从社会热点事件中提取关键事件具有较高应用价值。本文通过将事件的社会影响力和时间分布进行混合建模，提出一个面向微博热点事件的关键事件提取框架，能够有效抽取微博热点中的关键事件，建模事件的发展演化过程。在两个真实微博数据集上对本文提出的抽取方法进行了实验验证并构建了一个微博关键事件抽取系统，对完整性和冗余度两个指标进行评价，结果表明本文提出的抽取方法能保证提取事件集合的完整性，同时有效减小提取事件集合的冗余度。

参考文献（References）

[1] 中国互联网络信息中心. 第48次中国互联网络发展状况统计报告[R]. 北京：中国互联网络信息中心，2021.（China Internet Network Information Center. The 48th statistical report on China's Internet development[R]. Beijing: CNNIC, 2021.）

[2] MU L, JIN P Q, ZHAO J, et al. Detecting evolutionary stages of events on social media: a graph-kernel-based approach[J]. Future Generation Computer Systems, 2021, 123: 219-232.

[3] MU L, JIN P Q, ZHENG L Z, et al. Lifecycle-based event detection from microblogs[C]// Companion Proceedings of the 2018 Web Conference. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2018: 283-290.

[4] 赵旭剑，王崇伟. 基于图卷积网络的微博新闻故事线抽取方法[J]. 计算机应用，2021，41(11)：3139-3144.（ZHAO X J, WANG C W. Storyline extraction method from Weibo news based on graph convolutional network [J]. Journal of Computer Applications, 2021, 41(11): 3139-3144.）

[5] 陈震，王静茹. 基于贝叶斯网络的网络舆情事件分析[J]. 情报科学，2020，38(4)：51-56，69.（CHEN Z, WANG J R. Analysis of network public opinion events based on Bayesian network [J]. Information Science, 2020, 38(4): 51-56, 69.）

[6] 李培，翁伟，林琛. 中文微博故事线生成方法[J]. 中文信息学报，2016，30(3)：143-151.（LI P, WENG W, LIN C. Method for generating microblogs storylines[J]. Journal of Chinese Information Processing, 2016, 30(3): 143-151.）

[7] YUAN R F, ZHOU Q F, ZHOU W B. dTexSL: a dynamic disaster

textual storyline generating framework [J]. World Wide Web, 2019, 22(5):1913-1933.

[8] 彭敏,傅慧,黄济民,等. 基于核主成分分析与小波变换的高质量微博提取[J]. 计算机工程, 2016, 42(1):180-186. (PENG M, FU H, HUANG J M, et al. High quality microblog extraction based on kernel principal component analysis and wavelet transformation [J]. Computer Engineering, 2016, 42(1):180-186.)

[9] 刘国威,成全. 基于网络舆情生命周期的微博热点事件主题演化研究[J]. 情报探索, 2018(4):11-19. (LIU G W, CHENG Q. Research on the topic evolution of Microblog hot events based on the life cycle of network public opinion[J]. Information Research, 2018(4):11-19.)

[10] 王东波,叶文豪,吴毅,等. 基于多特征时间抽取模型的食品安全事件演化序列生成研究[J]. 情报学报, 2017, 36(9):930-939. (WANG D B, YE W H, WU Y, et al. Researches of generating time evolution sequences of food safety events based on multiple time extraction model[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(9):930-939.)

[11] 欧阳逸,郭斌,何萌,等. 微博事件感知与脉络呈现系统[J]. 浙江大学学报(工学版), 2016, 50(6):1176-1182. (OUYANG Y, GUO B, HE M, et al. Event sensing and vein presentation leveraging microblogging [J]. Journal of Zhejiang University (Engineering Science), 2016, 50(6):1176-1182.)

[12] TRAN T A, NIEDERÉE C, KANHABUA N, et al. Balancing novelty and salience: adaptive learning to rank entities for timeline summarization of high-impact events[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, New York: ACM, 2015: 1201-1210.

[13] 夏立新,陈健瑶,余华娟. 基于事理图谱的多维特征网络舆情事件可视化摘要生成研究[J]. 情报理论与实践, 2020, 43(10):157-164. (XIA L X, CHEN J Y, YU H J. Research on the visual summary generation of network public opinion events based on multi-dimensional characteristics of event evolution graph[J]. Information Studies: Theory and Application, 2020, 43(10):157-164.)

[14] 任卓明,邵凤,刘建国,等. 基于度与集聚系数的网络节点重要性度量方法研究[J]. 物理学报, 2013, 62(12): No. 128901. (REN Z M, SHAO F, LIU J G, et al. Node importance measurement based on the degree and clustering coefficient information [J]. Acta Physica Sinica, 2013, 62(12): No. 128901.)

[15] 田世海,董月文,王健. 基于NRL和*k*-means的舆情事件聚类研究[J]. 情报科学, 2021, 39(2):129-136. (TIAN S H, DONG Y W, WANG J. Clustering research on lyrical events based on NRL and *K*-means[J]. Information Science, 2021, 39(2):129-136.)

[16] 李进华, 安仲杰. 基于地理坐标的微博事件检测与分析[J]. 现代图书情报技术, 2016, 32(2):90-101. (LI J H, AN Z J. Analyzing geographical coordinates data for micro-blog trending events[J]. New Technology of Library and Information Service, 2016, 32(2):90-101.)

[17] ZHAO X J, JIN P Q, YUE L H. Discovering topic time from Web news[J]. Information Processing and Management, 2015, 51(6): 869-890.

[18] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]// Proceedings of the ACL-2004 Workshop: Text Summarization Branches Out. Stroudsburg, PA: Association for Computational Linguistics, 2004: 74-81.

This work is partially supported by Humanities and Social Sciences Foundation of Ministry of Education (17YJCZH260), Key Program of Science and Technology Department of Sichuan Province (2020YFS0057), CERNET Innovation Project (NGII20180403).

ZHAO Xujian, born in 1984, Ph. D., associate professor. His research interests include text mining, natural language processing, Web information processing.

WANG Chongwei, born in 1995, M. S. candidate. His research interests include information extraction, machine learning.

WANG Junli, born in 1996, M. S. candidate. His research interests include information extraction, machine learning.