

网络舆情事件链演化分析

李仁德, 冯倩, 李瑜, 曹春萍

(上海理工大学图书馆, 上海 200093)

摘要: 为防范舆情风险, 分析网络舆情的细粒度演化, 提出一种去冗余的衍生事件内容关联演化分析框架。通过文本挖掘技术从海量文本流中提取主要的衍生事件, 将舆情内容压缩到人工可判读的数量级; 利用词移距计算相邻两个时间片上的衍生事件相似度, 构建反映演化关系的衍生事件链图。以“上海特斯拉自燃”事件为例进行事件链演化分析, 得到微博网络舆情事件发展不同阶段涉事主体在各个衍生事件中的话题转移关系, 最后鲁棒性分析的结果验证了该分析方法具有降低微博短文本冗余信息的能力, 提高了事件演化关联识别分析的准确性。该研究方法为舆情事件的事后复盘、同类舆情事件的预判和介入、衍生事件的科学研判提供了决策支持。

关键词: 事件链; 舆情演化; 网络舆情; 特斯拉自燃

中图分类号: G 206.3

文献标志码: A

Evolution of online public opinion based on chain of sub-events

LI Rende, FENG Qian, LI Yu, CAO Chunping

(Library, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The evolution analysis of the public opinion in emergency is the foundation for the risk prevention and control. An analysis framework for the evolution of the public opinion was proposed based on the event chain. Firstly, text mining technology was used to extract the main sub-events from the massive text stream, thereby reducing the public opinion content to the order of magnitude that can be manual interpretation and discrimination. Secondly, the word mover's distance was used to calculate the similarity of sub-events on two adjacent time slices, so that the event chain diagram could be constructed. The evolution analysis was carried out with the case of “Tesla spontaneous combustion in Shanghai”. The relationship of topic shifting was built between sub-events at different stages of the evolution. Finally, it was verified by the robustness analysis that the method can reduce the problem of redundant information in short texts of microblogs and improve the accuracy of sub-event correlation. The research results provide decision support for the post-event review, the prediction and intervention of similar public opinion events, and the scientific evaluation of sub-events.

Keywords: chain of sub-events; public opinion evolution; online public opinion; Tesla spontaneously combustion

收稿日期: 2022-11-01

基金项目: 国家自然科学基金青年基金资助项目 (71901144); 中国青少年研究会研究课题 (2023B18); 尚理晨曦社科专项项目 (22SLCX-ZD-005)

第一作者: 李仁德 (1985-), 男, 副研究馆员。研究方向: 网络舆情分析。E-mail: lirende@usst.edu.cn

网络舆情分析与治理关乎国家安全和社会稳定。党的二十大报告提出，需加强全媒体传播体系建设，塑造主流舆论新格局，健全网络综合治理体系，推动形成良好网络生态。如何有效分析海量舆情信息，细粒度研究舆情演化规律，降低舆论负面影响，成为涉事方应对舆论亟需关注的问题。面对各种社会矛盾和利益分化现象的相继涌现，人们为自己发声，表达利益诉求的欲望日益高涨，促使舆情事件的演化表现出多样性和复杂性的特征。传统的网络舆情分析关注舆情在社交媒体中的内容演化^[1-2]，然而，事件的发展会衍生出一系列事件，细粒度的衍生事件关系往往隐含更多的演化规律。本文将衍生事件定义为延续主体事件某一特定方面的事件，对衍生事件进行分析对于辅助网络舆论的引导工作以及预测类似事件的发展方向具有重要的参考价值^[3-4]。

传统舆情演化的文本分析方法经历了两个阶段。第一个阶段以话题检测与追踪 (topic detection and tracking, TDT) 为代表，该阶段的研究能够实现话题信息的聚合^[5-6]，提升人们对海量信息识别的能力。但在获取话题信息后，TDT 技术并没有对话题内容的联系和发展过程进行深入分析，致使用户不能清楚地掌握事件的衍生和演变历程。第二阶段是在第一阶段的基础上，对话题的内容演化进行深入分析，主要采用主题模型和事件链方法^[7-8]。主题模型可以挖掘出事件中隐含的主题信息，通过分析每一时间段的主题分布，结合时间信息得到事件内容的演化脉络。但该方法一般呈现出按时间排列的主题词语集合，且不同时间段的词语集合存在大量的重叠，致使一些不了解该舆情事件的用户很难深入理解话题演化脉络。基于事件链的网络舆情演化研究方法，首先检测话题中包含的相关事件，然后识别事件之间是否存在演化关系，最后通过建立事件链得到舆情内容的演化脉络。事件演化的逻辑源于随时间发展出现的各衍生事件。总体来说，传统舆情演化研究的基本思路是提取事件在不同发展阶段的内容信息，并按时间顺序展示给用户。但这些研究对于提取哪些事件信息、事件之间是否具有演化关系，以及用何种方法能预测类似网络舆情的态势走向等问题并没有作深入的回答。

本文拟解决的关键问题包括：a. 如何从海量微博文本流中提取衍生事件？b. 如何确定舆情事

件间的逻辑关系并构建事件链图？c. 如何利用事件链图分析舆情内容变化？事件链是指在一定的时空条件下，一个事件触发另一个或者多个事件的现象，事件相继触发形成链式结构特征。网络舆情事件链是初始舆情事件在其所处情境的作用下触发了与它关联的其他舆情事件，从而形成的链式效应^[9]。事件链反映了事件之间的因果与时序关系，此类链式结构证明知识在事件之间具有流动性与转化性。本文将“上海特斯拉自燃”事件作为事件链演化分析案例，对事件链形成中的衍生事件提取、衍生事件关联和演化进行系统性建模和分析。

1 相关研究

网络舆情事件的演化分析建模涉及 3 个核心问题：a. 事件提取，包括一类核心事件或活动以及所有与之直接相关的事件或活动，可以由多个衍生事件组成；b. 事件链关联，核心事件发生后所引发的一系列衍生事件，以及形成的事件链式效应；c. 事件演化，刻画事件链上不同类型的衍生事件的信息传播与变化过程。

事件演化分析的首要任务是从海量文本中发现隐含的衍生舆情事件，其文本的非结构化特性以及衍生舆情事件的隐含性是两大瓶颈。在 Automatic Content Extraction^[2] 评测会议中，“事件”被描述为一个动作的发生或状态的变化，但目前学术界对“事件”仍没有统一的定义。Sun 等^[10]将事件表示为谓词+论元结构。事件提取涉及文本挖掘与聚类等方法^[11]，通过事件提取的技术，可以用来生成衍生舆情事件。衍生事件的研究多基于主题模型^[7,12]，主题模型通常假设衍生舆情事件数目固定，然而主题会随着事件的发展以及网络成员的交互行为而发生变化。此外，事件提取研究多以事件的事实发生为依据，对无监督学习情况下衍生舆情事件的多样性、复杂性和适用性的研究需要进一步拓展。对于相似度高、联系紧密的多个衍生事件，有效识别关键衍生舆情事件对于追踪事件演化过程至关重要。针对以上问题，本文基于 Single-Pass 语义聚类提取文本簇，从海量文本流中抽取主要的代表性文本簇并对其进行摘要描述，并归纳衍生事件。

首先，需提取文档中包含的衍生事件，识别

事件之间是否具有演化关系^[13-14], 构建事件链。由于具有演化关系的事件之间内容一般较为相似, 现有研究将该问题转化为事件内容相似度计算问题^[15-16]。Nallapati 等^[15]根据两个事件对应文档之间的相似度平均值来计算两个事件间的相似度。Yang 等^[16]利用事件内容相似度、时间邻近度和网页分布邻近度对事件演化关系进行建模, 从新闻语料库中发现事件演化图。Xu 等^[17]利用关联挖掘规则建立事件间的语义演化关系。多数研究通过比较事件关键词的相似度来识别事件间演化关系, 例如基于向量空间表示事件内容、用余弦相似度计算事件相似性, 或是基于词频概率, 采用 JS 散度或 KL 散度计算事件相似性。这些方法多依赖于词语级的相似度比较, 无法兼顾文档的词汇语义信息。本文基于词移距 (word mover's distance, WMD) 构建事件链。在语义聚类归纳得到的衍生事件基础上, 通过词移距计算相邻两个时间片上的衍生事件相似度来识别事件间的演化关系, 从而得到更好的事件演化关系识别性能, 提高事件链图构建的准确性。

舆情演化分析主要研究舆情事件内容及关系随时间和空间的变化趋势^[18]。Yang 等^[16]通过事件演化识别技术来自动识别事件演化关系, 并通过事件链的形式展示事件内容演化图。Liu 等^[19]提出了一种动态网络知识扩散的方法, 从细粒度层面反映网络知识的演变, 丰富了舆情演变分析的视角。Li 等^[20]通过演化分析的方法, 发现无论信息传播的内容特征如何, 用户的社交结构特征具有一定稳定性。Lian 等^[21]使用 SIR 模型分析了政府、媒体和网民 3 组人群在时间和空间上的相互作用, 从网络空间结构的角度研究了网络舆情传播的规律和效果。这些研究关注事件不同发展阶段的内容信息, 但舆情事件不同侧面的演化过程无法兼顾。特定舆情事件的发展过程往往伴随着相关衍生舆情事件的产生, 每一个衍生舆情事件都在描述事件的不同侧面。

综上, 系统理解舆情事件的演化发展脉络, 亟需从隐含的衍生舆情事件的提取、演化关系的关联构建, 以及演化过程的时空发展趋势上, 进行综合分析和评判。舆情事件分析中, 衍生事件的提取是关联分析的基础, 关联分析是演化分析的前提条件, 演化分析的结果是事件链的形式。本文提出的基于事件链的舆情演化分析方法, 从事件提取、事件链构建和演化图构建 3 方

面研究了网络舆情的演化, 使用户可以直观、清晰地了解舆情事件的细粒度演化过程, 掌握舆情事件的因果关联和来龙去脉。

2 研究方法

2.1 舆情演化框架

事件链的网络舆情演化研究分为事件提取、事件链构建和演化图构建 3 个过程, 对应的技术包括微博短文本聚类 and 衍生事件关联。首先, 利用文本挖掘技术从海量文本流中提取出主要的衍生舆情事件; 其次, 通过词移距计算相邻两个时间片上的衍生事件相似度来识别事件间的演化关系, 从而构建事件链图; 最后, 基于事件链图和事件演化概率构建此次网络舆情的内容演化图。分析框架流程如图 1 所示。

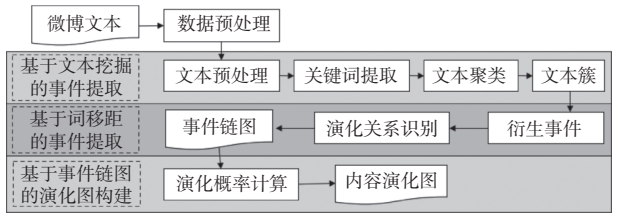


图 1 舆情演化分析框架流程图

Fig.1 Flow chart of public opinion evolution analysis framework

2.2 事件提取

a. 文本预处理。在对数据预处理的基础上进一步对文本进行处理, 删除与事件无关的特殊符号、表情、链接, 如频繁出现的词语“转发”、“微博”、“@用户”, 以及标记性符号等。保留代表话题“#自燃事件#”。使用 Jieba 分词工具对微博文本内容进行分词处理, 同时去除停用词、拟声词、特殊符号。如“的”、“地”、“啊”等词以及无用的标点符号, 得到每条博文词汇集。

b. 关键词提取。通过 TextRank^[22] 算法对分词后的博文词汇集进行关键词提取, 得到博文关键词词汇集 $X = \{x_1, x_2, \dots, x_n\}$, 其中, x_i 为提取的关键词。

c. 文本聚类。为了能够充分研究词汇与词汇之间的关系, 考虑到 word2vec 只是基于词的维度进行语义分析, 并不具有上下文的语义分析能力, 本文在 word2vec 的基础上增加一个段落向量的 doc2vec, 训练文档集中所有词汇, 并进行词向量表征, 然后运用余弦相似度算法通过博文关键词计算任意两两博文之间的语义相似度。相较于

KL 散度或 JS 散度, 余弦距离是从方向上区分差异, 将阈值对相似的文本向量关键词合并, 两文本向量距离越小则相似度越高。

对文档集中所有词汇对应的关键词向量采用 Single-Pass^[23] 增量聚类算法聚成 J 个文本簇, 将事件提取问题转化为对关键词向量的聚类问题。预设一个聚类阈值, 顺序处理输入的文档, 计算新文档与已识别事件之间的相似度, 如果相似度大于该聚类阈值, 则将该文档加入与它相似度最大的文本簇中, 否则将该文档作为新事件创建。聚类得到各个文本簇后按照词频降序排列, 截取频率最高的前 d 个词, 作为聚类后的文本簇关键词集合 $E_i = \{x_1, x_2, \dots, x_d\}$, $i = 1, 2, \dots, s$ 。

d. 衍生事件提取。文本聚类将海量的舆情事件压缩到人工可判读的数量级, 同时有助于获得相互独立的舆情事件。仅仅将舆情事件的表示视为一个简单的词袋不便于之后的演化分析, 因此, 对得到的文本簇关键词集合 E_i 结合原数据对应的文本内容, 概括性描述出舆情事件摘要。

2.3 事件链构建

a. 衍生事件生成。一次网络舆情对应着多个不同阶段以及相关的衍生舆情事件, 记为衍生事件, 根据提取的舆情文本簇概括性描述出其发展阶段对应的衍生事件集 $E = \{E_1, E_2, \dots, E_s\}$ 。

b. 演化关系识别。具有演化关系的两个衍生事件通过语义进行关联, 采用 WMD 计算衍生事件之间的相似度, 即一个衍生事件中所有关键词“流向”另一个衍生事件的最小代价。衍生事件 E_i 和 E_j 之间的相似度通过将 E_i 中所有词“移动”到 E_j 中所有词的最小距离来度量, 即

$$d(E_i, E_j) = \min_{T_{ij} \geq 0} \sum_{i,j=1}^n T_{ij} L(i, j) = \min_{T_{ij} \geq 0} \sum_{i,j=1}^n T_{ij} \|i - j\|_2 \quad (1)$$

式中: $L(i, j)$ 表示词 i 和词 j 间的欧氏距离; T_{ij} 表示衍生事件 E_i 中的词 i 有多少转化为 E_j 中的词。

为确保能完全转化, 需进行如下约束:

$$\sum_{j=1}^n T_{ij} = E_i', \sum_{i=1}^n T_{ij} = E_j', \forall i, j \in \{1, 2, \dots, n\} \quad (2)$$

其中

$$E_i' = \frac{c_i}{\sum_{i=1}^n c_i} \quad (3)$$

式中, c_i 表示词 i 的词频。

本文只计算相邻两个时间片上衍生事件的相

似度, 如果相邻两个时间片上衍生事件间的相似度值大于或等于该阈值, 则认为两个衍生事件在内容上存在演化关系, 即两者间建立边关系。反之, 如果它们之间的相似度值小于该阈值, 则表明是相互独立的两个衍生事件, 即两个衍生事件在内容上不存在演化关系。根据计算的相似度可得到衍生事件之间的相关关系, 进而构建出事件链图。

2.4 演化图构建

将构建事件链图中的舆情衍生事件以“节点对”的形式两两结合进行计算, 每个节点对代表着网络舆情事件链节点之间的可能演化关系。运用统计概率式 (4) 计算出各衍生事件的先验概率, 然后再利用条件概率式 (5) 计算出节点对间的演化概率。衍生事件之间的演化概率构成事件内容演化图。

$$p(E_i) = \frac{|E_i|}{\sum_{i=1}^n |E_i|} \text{ 且 } \sum_{i=1}^n p(E_i) = 1, i = 1, 2, \dots, n \quad (4)$$

$$p(E_i|E_j) = p(E_i)p(E_j|E_i) \Rightarrow p(E_j|E_i) = p(E_i|E_j)/p(E_i) \quad (5)$$

式中, $|E_i|$ 为频次。

3 实证分析

3.1 实验数据集与预处理

数据集来源于中国新闻史学会计算传播学研究委员会与新浪微热点大数据研究院联合举办的第二届传播数据挖掘竞赛中“上海特斯拉自燃事件”的微博文本数据, 从 2019 年 4 月 21 日到 5 月 5 日共 15 d, 总计 61 688 条博文。该事件发生于 4 月 21 日晚, 上海某小区一地下车库内一辆特斯拉轿车忽然冒烟而引起自燃, 进而引发社会各个群体对新能源电动车安全性问题的一系列热议。

3.2 事件提取

数据预处理去除标签为机器人的数据以及少于 10 个字的微博文本后, 得到 40 119 条博文, 进一步删除与文本话题无关的特殊符号、表情符号等, 并使用 Jieba 分词工具进行博文分词。采用 TextRank 算法对分词后的博文词汇集进行关键词提取, 得到每条博文关键词词汇集。词汇集中存在大量关键词相似的冗余舆情事件, 为了消除冗余, 使用 doc2vec 训练文档集中所有词汇进行词向

量表征, 然后运用余弦相似度算法计算任意两两博文之间的语义相似度, 根据阈值对相似的文本向量关键词进行合并。选取的相似度阈值为 0.83。

根据阈值对相似的文本向量关键词进行合并, 对文档集中所有词汇对应的关键词向量采用 Single-Pass 聚类算法聚成 J 个文本簇, 将事件提取问题转化为对关键词向量的聚类问题。文本簇的聚类, 采用 K 折交叉检验的方法作验证。取 $K=5$, 将原始数据随机分为 5 份, 对每一份数据作相似性计算, 并得到每份样本的衍生事件提取结果。两两对比随机样本的关键词集合, 通过余弦相似性比较不同样本之间衍生事件的语义相似度, 结果如表 1 所示。

表 1 文本簇聚类的 K 折交叉检验
Tab.1 K -fold cross-test for text clustering

相似度	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
$K=1$	1				
$K=2$	0.905±0.002	1			
$K=3$	0.910±0.003	0.968±0.008	1		
$K=4$	0.984±0.007	0.952±0.009	0.902±0.005	1	
$K=5$	0.870±0.002	0.905±0.006	0.966±0.001	0.962±0.002	1

3.3 事件链构建

a. 衍生事件生成。根据文本聚类消除冗余舆情事件后, 最终保留 4 024 个合并文本, 归为 16 个文本簇^[9], 作为基础数据。这些文本簇涵盖了企业、车主、消防、媒体、专家、网友等不同社会群体, 能够较好地反映出消费者维权类网络舆情的主要特征及演化规律。由此, 以文本簇涉及的相似群体为依据, 人工划分并概括性地描述出对应的衍生事件 E_i , 衍生事件可以帮助全面深入地理解事件的各个方面。根据文本聚类统计出事件的出现频次, 为了方便后续的演化概率计算, 此处利用统计概率式 (4) 计算出各衍生事件的初始条件概率。

b. 演化关系识别。通过 WMD 计算相邻两个时间片上衍生事件之间的相似度来识别事件间的演化关系。若相似度大于某一阈值则认为两者之间具有演化关系, 即衍生事件间建立边关系, 构建出网络舆情事件链图, 如图 2 所示。

判别衍生事件关系的阈值取 0.61。在事件发生后, 该网络舆情同时向 4 种潜在的衍生舆情事

件演化, 即消防介入、媒体介入、企业发声和网友参与, 这些衍生事件都是不同参与方对初始舆情事件发生后的回应。例如, 博文“在得知这起发生在上海的事故后, 昨晚我们第一时间派出团队赶往现场。我们正在积极联络相关部门并配合核实情况。根据目前的信息显示, 没有人员伤亡”, 就是企业对事件发生后的回应。而每种潜在衍生事件又可能引发多种不同的演化路径。例如, 企业发声后又出现车主发声、媒体介入和事件处理 3 种潜在衍生事件。该事件链图反映了此次舆情事件发生后产生的一系列衍生舆情事件, 着重揭示了不同事件节点之间的潜在演化路径。

3.4 演化图构建

将 4 024 个聚类样本对应的衍生事件以事件链节点对的形式两两结合, 每个节点对代表着网络舆情事件链节点间的可能演化。衍生事件的先验概率是事件链节点对演化概率计算的基础, 不同舆情阶段的节点演化概率借助条件概率式 (5) 计算获得。基于事件链结构以及各衍生事件节点对的演化概率得到不同演化路径的发生概率, 图 2 即为事件内容演化图。在舆情爆发初始阶段, 该网络舆情同时向 4 个方向演化(消防介入、媒体介入、企业发声、网友参与), 此时向媒体演化的概率最小, 而是主要集中在企业发声上, 反映了初始舆情事件爆发后, 人们更倾向于关注涉事方特斯拉的官方回应。企业发声后, 网络舆情演化出了 3 条发展路径(车主发声、媒体介入、事件处理), 其中车主发声的转移率高达 51.36%, 此时车主急切需要企业给出答复。在媒体介入事件后, 企业和车主演化的概率基本持平, 因为分别作为涉事方和维权方, 媒体对其关注度基本相同。在网友参与阶段, 由于大量的网友参与产生了较多的谣言与曲解事件, 极易引起行业专家进行初步解读, 消除网友的各种猜测。随着时间的推移, 在车主和媒体的共同推动下, 事件进入最终的处理阶段。

事件发生可利用状态转移矩阵来预测下一步的衍生舆情事件, 直至整个事件趋于稳定终止。经过 7 次迭代, 得到趋于稳定状态下的演化趋势图, 如图 3 所示。

根据传播学中舆情发展的经典四阶段(萌芽、兴起、成熟、衰退)理论, 将萌芽和兴起阶段合并为一个“突发”阶段, 将事件演化分为突发(0—1),

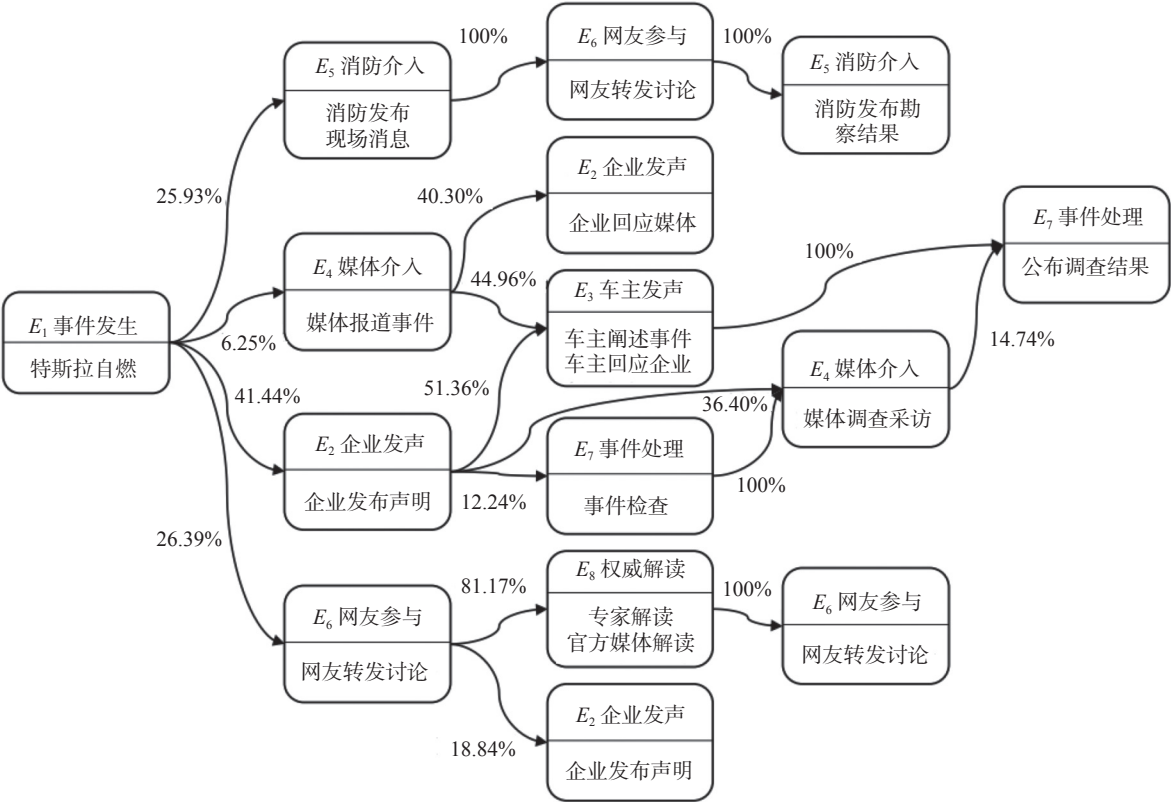


图 2 上海特斯拉自燃舆情事件链图 (内容演化图)

Fig.2 Chain diagram of sub-events for Tesla spontaneous combustion (content evolution diagram)

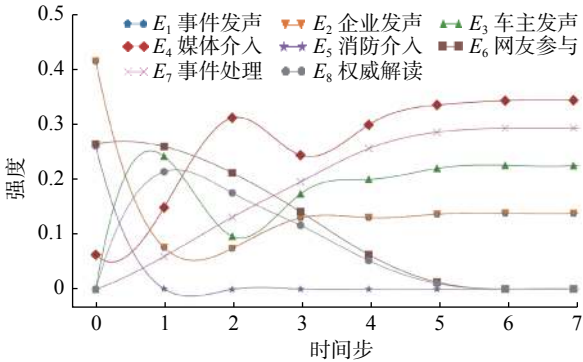


图 3 舆情事件趋势演化图

Fig. 3 Evolution trend of sub-event

成熟 (1—3), 衰退 (3—7) 3 个阶段进行分析。在时间步 0—1 阶段, 衍生事件发生的状态概率较小, 这是由于在上海特斯拉自燃事件发生以前, 网络上就零散地存在着众网友对新能源汽车安全性问题的讨论, 事件发生企业作为涉事方被关注程度最高, 大量网友参与, 直接推动了舆情事件的发生。此时消防受关注程度也较高, 这是因为消防在第一时间介入事故现场进行处理并通报。与此同时, 车主作为维权方也开始发声。在时间步 1—3 阶段, 舆情事件处于成熟期, 可以清晰地

看出媒体受关注程度最高, 说明此时公众更关注媒体对于事件进展情况的报道。在此阶段, 意见领袖在舆情的传播过程中起到了导向作用, 权威解读与媒体的报道也在一定程度上消除了网友因曲解而产生的噪音, 导致网友参与度下降。在时间步 3—7 阶段, 人们对事件处理结果的关注度越来越高, 都期待事件的最终调查结果出现, 对媒体的关注则相对平稳。随着时间的发展, 企业迟迟没有给出最新调查结果, 车主也没有得到及时答复, 因此, 企业和车主作为事件处理的主体方, 受关注度逐渐增长。最后, 舆情随各方主体的关注减少进入衰退期。

3.5 内容演化的评判指标

由于衍生舆情事件的归纳涉及人为标签, 具有一定的主观性, 因此, 将其与 Timeline^[24] 和 Story Telling^[25] 方法进行对比。实验邀请 9 位志愿者从准确性和易理解性两方面对 3 个方法进行评分。其中, 准确性是指该演化过程是否能准确描述舆情事件的发展, 易理解性是指演化过程是否有助于用户理解事件的脉络。评分标准为 1~5 分, 从低到高依次代表很差、差、一般、良好、

优秀。取平均值作为舆情演化方法的评判结果, 如表 2 所示。

评分结果显示, 采用的事件链方法使得舆情内容演化在准确性和易理解性两方面均表现良好。

此外, 当两个文档没有出现相同词语时, WMD 方法依旧可以识别其语义是否相同, 使得衍生事件的关系配对更加稳定, 并且配合上一步事件提取剔除的冗余信息, 能够得到稳定的演化结果。事件演化图的鲁棒性分析承接上述 K 折交叉检验的方法, 同时对比了词移距与传统的余弦相似性、欧式距离、Jaccard 相似性、海明距离。对每

表 2 基于用户体验的舆情衍生事件内容准确性与易理解性评分

Tab.2 Evaluation of accuracy and accessibility of public opinion content about sub-events based on user experience

评价指标	Timeline	Story Telling	本文方法
准确性	3.33	3.11	3.89
易理解性	3.22	2.89	4.22

一份样本单独进行事件演化图的概率分析, 得到衍生事件转移概率的相似性对比结果, 如表 3 所示。

表 3 舆情衍生事件关联的鲁棒性对比

Tab.3 Robustness comparison of sub-event correlations

衍生事件	词移距	余弦相似性	欧式距离	Jaccard 相似性	海明距离
E_1	0.915±0.002	0.845±0.016	0.801±0.022	0.836±0.018	0.821±0.016
E_2	0.904±0.009	0.884±0.015	0.869±0.039	0.863±0.024	0.803±0.036
E_3	0.967±0.007	0.803±0.045	0.906±0.044	0.904±0.011	0.865±0.034
E_4	0.973±0.001	0.892±0.026	0.814±0.041	0.817±0.019	0.878±0.062
E_5	0.913±0.009	0.832±0.046	0.854±0.034	0.817±0.099	0.815±0.060
E_6	0.955±0.004	0.804±0.022	0.822±0.046	0.866±0.059	0.882±0.038
E_7	0.978±0.008	0.875±0.028	0.823±0.047	0.845±0.036	0.801±0.024
E_8	0.908±0.004	0.810±0.022	0.812±0.011	0.805±0.053	0.819±0.035

4 结束语

舆情事件分析中细粒度的衍生事件演化迫切需要方法论的支撑, 为此, 提出了一种基于事件链的研究方法。利用文本挖掘技术从海量文本流中抽取出主要的衍生舆情, 通过短文本相似性合并, 扩充对衍生事件的富文本信息认知, 解决了微博短文本信息冗余的问题。通过词移距计算相邻两个时间片上的衍生事件相似度, 识别衍生事件间的演化关系, 并构建事件链图, 能够更准确并稳定地获取不同类型事件之间的转移概率, 解决了衍生事件的关联关系识别问题。最后, 对事件链图提出了一种关联事件对演化概率计算的方法, 基于事件链图和事件演化概率构建舆情事件的内容演化图。两种文本相似性的方法组合, 取得了更稳定的事件链构造结果。

本文对微博舆情事件演化分析提供了以下启发: a. 对舆情事件的事后复盘具有参考价值。通过舆情事件的文本分析和计算, 可以了解事件的衍生和演化关系。事后复盘对于舆情案例制作、政策制定及危机公关解决具有关键的参考作用。b. 对同类舆情事件的预判和介入提供管理依据。

用户可通过演化图清楚地掌握舆情事件随时间发展的演变过程, 为类似网络舆情事件的应对工作提供参考, 同时有助于决策主体把握舆情演化的过程与机制, 对于辅助网络舆论的引导工作以及作出合理的应急决策具有现实意义。c. 对细粒度衍生事件引发的舆情异化提供了科学支撑。不同利益主体在舆情事件发展各阶段对于推动衍生话题的作用和关系迥然不同, 通过微博舆情的建模, 梳理主事件与衍生事件关系以及进行事件发展的量化, 有助于形成科学的网络舆情研判机制。

研究的不足之处在于, 衍生事件的提取虽然有利于直观理解事件的不同侧面以及公众舆论的热点迁移, 但衍生事件的含义由人工概括描述。如何引用外部知识提升衍生事件的可理解性与准确性将是下一步的研究工作。同时, 识别演化关系时只考虑了相邻两个时间片上的衍生事件相似度, 未设置一个时间范围阈值以提高演化识别的精准性。在未来的研究中, 如何将人类知识融合机器学习、深度学习、复杂网络等技术提高事件分析的性能, 以及如何充分利用微博的链接关系来提高事件演化关系识别的精准性, 也是值得深入探讨的问题。

参考文献:

- [1] MA H Y. Research of evolution mechanism of network group event based on grounded theory under micro-blog platform[J]. *Journal of Service Science and Management*, 2015, 8(5): 678–684.
- [2] LI S Y, LIU Z X, LI Y L. Temporal and spatial evolution of online public sentiment on emergencies[J]. *Information Processing & Management*, 2020, 57(2): 102177.
- [3] HUO L, MA C Y. The interaction evolution model of mass incidents with delay in a social network[J]. *Physica A: Statistical Mechanics and its Applications*, 2017, 484: 440–452.
- [4] 张磊, 王延章, 陈雪龙, 等. 面向突发事件应急决策的情景建模方法 [J]. *系统工程学报*, 2018, 33(1): 1–12.
- [5] LI R D, GUO Q, ZHANG X K, et al. Reconstruction of unfolding sub-events from social media posts[J]. *Frontiers in Physics*, 2022, 10: 918663.
- [6] LI R D, MA H T, WANG Z Y, et al. Entity perception of two-step-matching framework for public opinions[J]. *Journal of Safety Science and Resilience*, 2020, 1(1): 36–43.
- [7] GUO X, XIANG Y, CHEN Q, et al. LDA-based online topic detection using tensor factorization[J]. *Journal of Information Science*, 2013, 39(4): 459–469.
- [8] HUANG W D, WANG Q, CAO J. Tracing public opinion propagation and emotional evolution based on public emergencies in social networks[J]. *International Journal of Computers Communications & Control*, 2018, 13(1): 129–142.
- [9] 夏立新, 毕崇武, 梅潇, 等. 基于事件链的网络舆情事件演化研究 [J]. *情报理论与实践*, 2020, 43(5): 123–130.
- [10] SUN R, WANG Z C, REN Y F, et al. Query-biased multi-document abstractive summarization via submodular maximization using event guidance[C]//17th International Conference on Web-Age Information Management. Nanchang: Springer, 2016: 310–322.
- [11] CUI W W, LIU S X, TAN L, et al. TextFlow: towards better understanding of evolving topics in text[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2011, 17(12): 2412–2421.
- [12] 关鹏, 王曰芬, 傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析 [J]. *图书情报工作*, 2016, 60(2): 112–121.
- [13] DU Y J, YI Y T, LI X Y, et al. Extracting and tracking hot topics of micro-blogs based on improved latent dirichlet allocation[J]. *Engineering Applications of Artificial Intelligence*, 2020, 87: 103279.
- [14] CURISKIS S A, DRAKE B, OSBORN T R, et al. An evaluation of document clustering and topic modelling in two online social networks: twitter and reddit[J]. *Information Processing & Management*, 2020, 57(2): 102034.
- [15] NALLAPATI R, FENG A, PENG F C, et al. Event threading within news topics[C]//Proceedings of the 13th ACM International Conference on Information and Knowledge Management. Washington: ACM, 2004: 446–453.
- [16] YANG C C, SHI X D, WEI C P. Discovering event evolution graphs from news corpora[J]. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2009, 39(4): 850–863.
- [17] XU Z, WEI X, LUO X F, et al. Knowle: a semantic link network based system for organizing large scale online news events[J]. *Future Generation Computer Systems*, 2015, 43–44: 40–50.
- [18] FANG S W, ZHAO N, CHEN N, et al. Analyzing and predicting network public opinion evolution based on group persuasion force of populism[J]. *Physica A: Statistical Mechanics and its Applications*, 2019, 525: 809–824.
- [19] LIU J G, ZHOU Q, GUO Q, et al. Knowledge diffusion of dynamical network in terms of interaction frequency[J]. *Scientific Reports*, 2017, 7(1): 10755.
- [20] LI R D, LIU J G, GUO Q, et al. Social signature identification of dynamical social networks[J]. *Physica A: Statistical Mechanics and its Applications*, 2018, 508: 213–222.
- [21] LIAN Y, DONG X F, LIU Y J. Topological evolution of the internet public opinion[J]. *Physica A: Statistical Mechanics and its Applications*, 2017, 486: 567–578.
- [22] ZHANG Y, CHEN F, ZHANG W F, et al. Keywords extraction based on word2Vec and TextRank[C]//The 3rd International Conference on Big Data and Education. London: ACM, 2020: 37–42.
- [23] 赵爱华, 刘培玉, 郑燕. 基于 LDA 的新闻话题子话题划分方法 [J]. *小型微型计算机系统*, 2013, 34(4): 732–737.
- [24] WANG Z H, SHOU L D, CHEN K, et al. On summarization and timeline generation for evolutionary tweet streams[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(5): 1301–1315.
- [25] SMITH D, SCHLAEPFER P, MAJOR K, et al. Cooperation and the evolution of hunter-gatherer storytelling[J]. *Nature Communications*, 2017, 8(1): 1853.

(编辑: 丁红艺)