文本数据事件检测的研究热点及趋势分析

付 琳,张 媛
(首都师范大学 管理学院,北京 100048)

摘 要:随着科学技术的发展,人们发布信息、表达观点的渠道越来越多,且最常见的信息载体就是文本。文本数据事件检测能够从海量数据中识别和检测当前正在发生的热点或突发事件,可以很好地支持应急管理、舆情监控、信息安全等领域的工作。为了掌握文本数据事件检测的研究状况,揭示该领域的研究热点和发展趋势,该文以中国知网作为文献数据库,2003 年-2021 年的 440 篇期刊论文作为样本,借助科学计量软件 CiteSpace,从发文时间分布、基金支持、主要研究力量、主要刊载平台、高被引文献分布、研究热点以及演化路径等方面,对文本数据事件检测研究进行计量分析。研究表明,事件检测研究发文量趋于稳定,且研究质量正在不断提高。研究热点为突发事件与热点话题的文本事件检测应用研究、基于微博数据的事件检测研究和以聚类为主要方法的事件检测方法研究。研究发展分为概念形成和工具开发两个阶段,并已经出现了领域扩散的现象。现阶段的研究重点集中在事件检测技术,社交媒体事件检测和事件检测在突发事件中的应用等方面。基于深度学习的文本数据检测方法、在社交媒体中的突发事件检测逐渐成为本领域研究的发展趋势。

关键词:事件检测;CiteSpace;研究热点;研究趋势;可视化分析
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)02-0024-08
doi:10.3969/j.issn.1673-629X.2023.02.004

Research Hotspots and Trend Analysis of Text Data Event Detection

FU Lin,ZHANG Yuan
(School of Management,Capital Normal University,Beijing 100048,China)

Abstract:With the development of science and technology,people have more and more channels to publish information and express their opinions. The most common information carrier is text. Text data event detection identifies and detects hot or unexpected events that are currently occurring from massive amounts of data,and it can well support emergency management,public opinion monitoring,information security and other fields. In order to grasp the research status of text data event detection and reveal the research hotspots and development trends in this field,we use CNKI as a literature database,440 journal articles from 2003-2021 as a sample,econometric analysis of text data event detection research in terms of distribution of publication time,funding support,major research power,major publication platforms,distribution of highly cited literature,research hotspots and evolutionary paths. The results show that the number of event detection research publications is stabilizing and the quality of research is improving. The research hotspots are the application research of text event detection for breaking news and hot topics,research on event detection based on twitter data and research on event detection method based on clustering. Research development is divided into two phases:concept formation and tool development,and domain proliferation has already occurred. At this stage,the research focus is divided into three categories,event detection technology,event detection research in social media and event detection research in emergent events. Text data detection methods based on deep learning and emergency detection in social media have gradually become the development trend of this field.

Key words:event detection;CiteSpace;research hotspot;research trends;visual analysis

0 引 言

文本数据事件检测是信息抽取中被广泛研究的一个问题,起源于 1997 年启动的话题检测与追踪(TDT)研究。主要是从海量文本数据中自动提取事件或话题的信息,实现对未知事件或话题的发现。这些文本可以是传统媒体的新闻报道,也可以是社交媒体上的帖子或推文。目前事件检测在突发事件检测、网络舆情检测、热点话题发现等方面有较好的应用。例如,Johnson N F 等人研究了与恐怖组织 ISIS 相关的个人或组织,分析他们在社交网络中的行为与现实世界所发生的极端恐怖事件之间的联系,帮助预测了现实世界中可能出现的恐怖袭击事件[1]。事件检测与话题检

收稿日期:2022-03-24 **修回日期**:2022-07-27
作者简介:付 琳(1997-),女,硕士,研究方向为舆情分析与数据挖掘;通讯作者:张 媛(1978-),女,副教授,CCF 会员(2163M),研究方向为舆情分析、情感分析与数据挖掘。

测最主要的区别在于对事件和话题的定义。一般来说,事件是由特定原因、条件引起,发生在某些特殊时间、地点的重要事情。事件相对于话题来说更具有局限性,同一话题下可能涵盖多个相似或相关事件。

随着大数据和人工智能等技术的发展和突破,事件检测中运用的方法更加丰富,效率和准确率都得到了显著提升。在事件检测领域已有的研究中,大部分是关于计算机科学和情报学领域的内容,尽管已有一些学者对事件检测进行综述研究,但大多只关注和分析了事件检测在社交媒体中的应用,对国内事件检测研究涉及的其他领域和方向分析较少。同时也缺少基于文献计量分析的研究工作,基于文献计量分析的研究能够推进事件检测领域的系统发展,帮助研究人员把握科研工作的方向。因此,该文从文献的角度出发,借助 CiteSpace 软件,采用文献计量分析方法对事件检测研究进行可视化分析,明确事件检测领域的研究内容,梳理研究现状,分析研究热点与重点,探索研究演化趋势和未来研究方向。并且这是首次使用知网数据库对事件检测进行文献计量分析。

1　研究设计

1.1　数据来源

该文选择中国知网(CNKI)数据库作为数据收集平台。采用主题检索的方式对 CNKI 中信息科技类文献进行检索,检索时间段为 2003 年 1 月 1 日至 2021 年 11 月 10 日,来源类别为期刊。以事件检测为主要检索词,同时为了改善单一主题或关键词检索中查全率不高的问题,对主题词进行拓展。话题检测和事件检测同属于 TDT 技术,两者的研究相互交叉,有很多重合部分。另外,对于“event detection”的中文翻译不尽相同,存在事件检测、事件发现、事件识别等结果,但其研究方法和方向都可以归到事件检测领域。因此,增加了话题检测、事件探测、事件识别、事件发现、话题识别、话题发现六个同义词语,作为主题和关键词检索的补充。通过检索共获得 4 712 篇期刊文献,去掉非文本数据研究和相关性较小的文献,最终筛选出 440 篇相关文献作为研究对象。

1.2　研究方法

文献计量分析是一种定量分析方法,以文献的各种外部特征作为研究对象,通过数学、统计学等计量方法来描述、评价和预测某个研究领域的现状与发展趋势,总结研究领域知识结构并探索研究前沿动态。

CiteSpace 是由陈超美教授研发的一款信息可视化软件,它主要基于共引分析理论和寻径网络算法,通过对特定领域文献进行计量和可视化图谱的绘制,来形成领域演化潜在动力机制的分析和领域发展前沿的探测[2]。

该文运用文献计量法对发文量、核心期刊占比、基金资助、研究机构和发文期刊等外部特征进行量化分析,借助 CiteSpace 软件对收集的相关文献进行可视化分析。在研究热点与研究重点的分析上,利用关键词共现与关键词聚类分析方法;在研究趋势与研究方向的分析上,利用时间线视图谱和关键词突现分析方法。

2　数据统计与分析

2.1　发文量分析

发文量在一定程度上可以反映一个研究领域在学术界受关注的程度。2003–2021 年,国内事件检测研究领域的发文数量如图 1 所示。

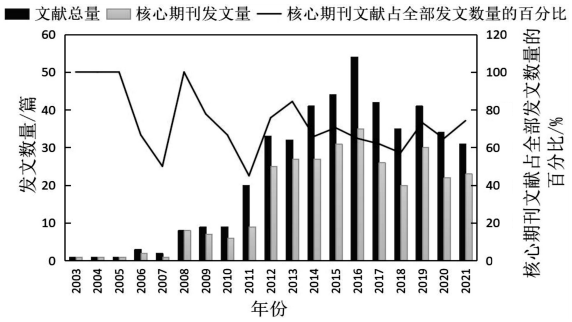

图 1　2003–2021 年发文量统计

2003–2010 年,国内事件检测研究属于起步阶段,发文量较少,但呈上升趋势。2011 年发文量有了显著上升,发文量是 2009 年的 2.2 倍。2011–2016 年,国内事件检测研究发文量逐年增长,2016 年达到顶峰。这与国内互联网普及、社交媒体开始流行有很大关系。社交媒体最大的特点就是能非常迅速即时地将信息传递给每一个用户。随着国内网民的增加,社交媒体中的数据开始爆发式增长。因此,学者们从基于新闻报道的长文本研究转向了对微博等社交媒体中短文本数据的研究。随着自然语言处理、机器学习等技术的不断发展,越来越多的学者将前沿技术运用于事件检测领域。2017–2021 年,该领域发文量基本保持稳定,说明国内事件检测研究正逐步趋于成熟。从核心期刊文献占全部发文数量的百分比可以看出,2012 年发文数量成一定规模以后,只有 2018 年略低于 60%,其余年份核心期刊的发文占比一直在 60% 以上,表明该领域的研究质量整体较高,且研究比较深入。

2.2　基金资助情况

文献的基金资助情况能反映学术研究的科学性和重要性。在 440 篇相关文献中,国家层面的基金支持有 234 篇,占 53%,其中国家自然科学基金委员会资助的论文最多,达到 182 篇;地方层面基金支持和无基金资助的文献有 206 篇,占 47%。说明国家层面对事件检测领域的关注度高,研究价值的认可度高,试图通

过资金支持、项目研发、人才培养等方式推进事件检测领域的研究。

2.3 主要发文机构和刊载平台

在 CNKI 中发表过事件检测相关研究论文的机构共有 281 个,但各发文机构之间合作极少,都是独立进行研究。国内的事件检测领域研究还未形成一个整体,各研究机构应当充分交流、加强合作,共同推进事件检测领域的创新发展。国内发文量最高的是中国科学院,共发表 24 篇,总被引量为 748 次。说明该研究机构较为关注事件检测领域,并对该领域的研究做出了较大贡献。武汉大学和四川大学紧随其后,分别发表了 14 篇和 13 篇文献,总被引量为 134 次和 127 次。除此以外,哈尔滨工业大学、苏州大学、昆明理工大学、南京理工大学、北京信息科技大学也是该领域发文量较多的机构。

在刊载平台方面,《中文信息学报》关于事件检测研究的刊文量最多,达到 24 篇。其次是《计算机应用研究》和《计算机工程》,共 21 篇和 19 篇。《计算机应用》《计算机工程与应用》《计算机科学》《计算机应用与软件》《情报杂志》等期刊也是事件检测研究的重要刊载平台。由此可知,在信息技术分类中,比较关注事件检测研究的是计算机软件与应用领域。

2.4 高被引文章分析

对相关文献进行整理,列出高被引文献及作者,见表 1。其中被引次数最高的两篇都是综述类文献。洪宇的《话题检测与跟踪的评测及研究综述》从 2007 年发表至今总共被引 487 次,平均每年被引 35 次。洪宇在文中对话题检测与追踪(TDT)技术进行了系统阐述,这篇综述文献在国内事件检测领域具有重要意义。

表 1 高被引文献及作者

标题	作者	年份	被引次数
话题检测与跟踪的评测及研究综述	洪宇	2007	487
话题识别与跟踪研究	李保利	2003	272
一种中文微博新闻话题检测的方法	郑斐然	2012	226
一种基于动态进化模型的事件探测和追踪算法	贾自艳	2004	220
基于隐主题分析和文本聚类的微博客中新闻话题的发现	路荣	2012	164
基于情感分布的微博热点事件发现	杨亮	2012	146
微博文本处理研究综述	张剑峰	2012	144
基于多策略优化的分治多层聚类算法的话题发现研究	骆卫华	2006	135
基于改进向量空间模型的话题识别与跟踪	宋丹	2006	114
基于隐含语义分析的微博话题发现方法	马雯雯	2014	102

文章介绍了 TDT 任务与评测的相关知识,包括相关定义、使用语料、评价体系以及层次结构,并重点论述和分析了国内外在该领域的相关研究及其相互关系[3]。另一篇高引综述文献是李保利的《话题识别与跟踪研究》,这是知网中最早介绍 TDT 的文献。李保利梳理了 TDT 的研究历史,并详细介绍了 TDT 的 5 个子任务:对新闻报道的切片,新事件的识别,报道关系识别,话题识别,话题跟踪[4]。

早期事件检测的研究还是以新闻语料为主,研究者们在信息检索技术的基础上,不断尝试新的方法改进算法模型,以提高新闻事件检测的效率。例如,基于时间距离的相似度计算模型[5]、多策略优化的分治多层聚类算法模型[6]、四向量相似度计算模型[7]。

2012 年,微博的迅猛发展带来了另一种社会化的新闻媒体形式。学术界将视角聚焦于社交媒体中的短文本,对短文本的研究很快成为了主流。所以,另外几篇高引文献均是从短文本数据中进行事件检测。郑斐然等人通过分析微博用户的习惯和数据特征,提出了一套完整的微博数据处理方法和新闻话题的检测算法。在向量空间模型的基础上,从文档主题词的时域分布中,筛选出信息量最大的新闻主题词,并进行聚类[8]。路荣等人通过充分挖掘隐主题来克服短文本数据稀疏性对文本相似度度量的影响,并使用一种两层的 K 均值和层次聚类的混合聚类方法来弥补层次聚类时间慢和 K 均值聚类无法事先指定中心个数的缺点[9]。马雯雯等人对前者的方法进行了优化,在混合聚类的基础上,引入隐含语义分析的方法对中文微博数据建模,解决了传统向量空间模型中高维和同义、多义的问题[10]。

显然并不是所有的微博都是描述新闻事件的,很多微博只是描述用户的心情、状态、工作情况等。有研究表明,当微博中情感词数量增多,并导致相邻时段中情感分布存在差异,这往往意味着热点事件的出现[11]。杨亮等人在此基础上提出了情感分布语言模型 ELM,用于发现微博平台中的热点事件[12]。

3 事件检测研究热点分析

3.1 关键词共现分析

关键词可以揭示文章的主要内容和核心,对事件检测领域相关文献进行关键词共现分析可以更好地了解该领域的研究热点。为了使可视化效果更好,对同义或近似义节点进行合并。最终得到事件检测研究相关文献的关键词共现图谱,如图 2 所示。共包含 348 个关键词,602 条连接,密度为 0.01,其中节点越大表明关键词出现频率越高,连线越多表明两个关键词共现次数越多,连线越粗表明联系程度越强[13]。

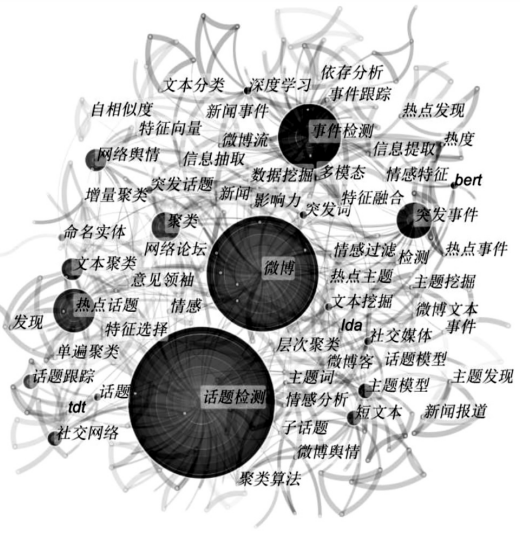

图 2　关键词共现图谱

为了更全面地了解事件检测的研究热点,该文通过统计和排序将共现频次前 6 的关键词及其信息绘制成表格,如表 2 所示。可以发现,最大的三个节点分别是话题检测(141 次)、微博(91 次)和事件检测(52 次)。其中“话题检测”和“事件检测”出现时间较早,是该领域的基础概念。“微博”于 2012 年出现,出现时间较晚,但共现频次很高,说明“微博”一出现就成为研究者们的关注焦点,且很快成为了该领域的研究热点。除此以外,“聚类”“热点话题”“突发事件”的中介中心性较高,因此可以初步判断它们也是事件检测领域的研究热点。

表 2　关键词频次和中介中心性

频次	中介中心性	年份	关键词
141	0.64	2006	话题检测
91	0.39	2012	微博
52	0.44	2004	事件检测
31	0.19	2004	聚类
23	0.12	2009	热点话题
23	0.12	2011	突发事件

3.2　关键词聚类分析

关键词是论文中出现频率最高、同时也是最核心的词汇,对文献进行关键词聚类分析可以从侧面反映出该领域各阶段研究的重点[14]。模块值(Q 值)和平均轮廓值(S 值)两个指标可以作为判断知识图谱绘制效果的依据。一般而言,Q>0.3 就意味着绘制的网络结构是显著的,越接近 1 则可认定该网络图谱所获得的聚类效果就越优秀。当 S 值>0.7 时,认为聚类是令人信服的,若在 0.5 以上,一般认为聚类是合理的。

事件检测关键词聚类图谱如图 3 所示。聚类模块值 Q 为 0.643 1>0.3,聚类平均轮廓值 S 为 0.886 7>0.7,说明聚类效果显著,且令人信服,具有较高的研究价值。共得到 7 个主要聚类,即话题检测(#0)、事件检测(#1)、突发事件(#2)、命名实体(#3)、网络舆情(#4)、主题发现(#5)、社交媒体(#6)。通过对聚类进行比较分析,将 7 个聚类分成 3 组。

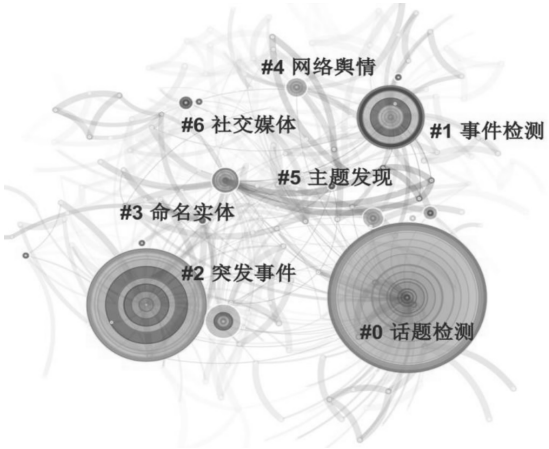

图 3　关键词共现图谱

(1)事件检测技术研究(#0、#1、#3、#5)。

从图 2 和图 3 可以看出,学者们较为关注对事件检测技术的研究。事件检测工作主要分为两部分:文本预处理和事件检测,它们分别对应不同的技术。在文本预处理阶段使用的技术大体可以分为三类:命名实体、特征提取或两者结合。

命名实体是自然语言处理中一项基础性关键任务,其主要任务是识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类。张阔等人利用统计方法优化不同类别新闻对于不同词性词元的权重,再根据已处理的新闻及话题信息动态调整词元权重,实验结果表明,其性能与同类事件检测模型相比有显著提升[15]。

特征提取是将原始数据的维度减少或将原始的特征进行重新组合,从而提高文本分类的准确性和效率。商宪丽等人就对传统文本特征提取进行改进,引入时间因素构建动态共词网络,利用网络统计特征动态提取微博文本特征,在实验中取得了较优的微博话题识别效果[16]。也有一些学者将两者方法结合使用。例如,刘素芹等人将新闻文档表示成基于命名实体及特征词的双特征向量,很好地解决了海量网络数据环境下相似话题难以区分的问题[17]。

在事件检测阶段,研究者们运用的方法主要是统计模型中的聚类分析。自 90 年代以来,统计模型一直是信息抽取的主流方法[18]。有非常多的统计方法被用来抽取文本中的目标信息,其中聚类分析被广泛应用于事件检测领域。聚类技术通常又被称为无监督学习。聚类可以根据给定的标准将数据集分割成不同的类簇,使得同一个类簇内的数据高度相似,从而实现对目标事件的检测。常用的聚类算法有基于划分的聚类

算法、基于层次的聚类算法以及基于模型的聚类算法。随着不断的实践,为了得到更好的聚类结果,学者们对各种聚类算法都进行了改进。

基于划分的聚类算法是聚类算法中最简单的一种。该种聚类要达到的要求是使类簇内部有较高的相似度,而类簇之间的相似度尽可能低。K-means 算法、Single-Pass 增量算法、围绕中心划分(PAM)算法等都得到了广泛的应用。张先飞等人利用触发词来确定 K-means 聚类初始质心,同时结合自相似度策略来确定 K 值,以解决聚类算法中 K 值及初始质心选取的问题[19]。税仪冬等人为解决增量式聚类初始模型不准确的问题,在 Single-Pass 聚类基础上添加了周期分类模块。该模块能够定期对已经聚类的报道分类,有效提高了话题簇的精度[20]。殷风景等人提出了 ICIT 聚类算法,继承 single-pass 算法的原理,通过引入正文和标题双向量的机制提高聚类结果的精确度[21]。

基于层次的聚类算法又称为树聚类算法。与 K-means 算法不同,层次聚类算法不需要预先设定聚类数,只要样本集合通过不断迭代达到聚类条件或者迭代次数即可。龙志祎等人先计算特征词对间基于互信息的相似度,之后采用自底向上的层次聚合聚类算法对特征向量进行聚类[22]。杨长春等人提出了一种改进的 CURE 层次聚类算法。将传统 CURE 算法中的代表点转换为博文种子集,提高了聚类的精确度[23]。

基于模型的聚类算法是假设每个类簇为一个模型,然后寻找与该模型拟合最好的数据,通常有基于概率和基于神经网络两种方法。前者最常用的方法是基于主题模型的聚类。主题模型假定数据的分布是符合一系列的概率分布,用概率分布模型对数据进行聚类,而不是像层次聚类和划分聚类那样基于距离来进行聚类。主题模型的方法一直备受青睐,学者们通过优化主题模型来改进和完善事件检测的效果和效率。姜晓伟等人提出词项聚合 LDA(term-aggregated LDA,tLDA)策略来解决传统 LDA 无法从短文本中获得足够信息的缺陷[24]。郭蓝天等人引入基于 CBOW(continuous bag-of-word)模型的词向量化方法,通过对 LDA 模型的输入进行相似词的聚类,使话题含义的表达更加明确[25]。为了提高检测的速度,聂文汇等人提出一种基于热度矩阵的主题模型,以词间的共有热度来挖掘各潜在主题间的语义关系。实验显示,在微博数据量达到 60 万条时,该方法依然可以在 1 min 内挖掘出潜在的热点话题[26]。

随着机器学习的发展,深度学习也逐渐成为事件检测的研究热点。相比于传统的主题模型方法,引入深度学习的模型无需人工定义的特征模板,能够自动地学习文本数据中的有效特征。因此,在标注语料充分的情况下,深度学习模型往往能够取得比传统方法更好的性能[27]。侯伟涛等人使用双向 LSTM 神经网络学习文本的隐藏特征,解决了传统方法通用性不强以及无法捕捉前后文隐含信息的缺点[28]。张秀华等人提出卷积神经网络构建中文新闻事件检测模型的方法,通过深度学习抽取文本深层特征[29]。马晨曦等人提出了可以避免误差传播的递归神经网络的事件检测联合模型,该模型不依赖于触发词表的构造和扩展,并且有很好的移植性[30]。

(2)社交媒体事件检测研究(#6)。

随着互联网的普及与高速发展,社交媒体已经成为人们分享观点、抒发情感、交流经验的主要渠道。现阶段的社交媒体包括微博、微信、博客、论坛、播客等。为了从社交媒体数据中获取有效信息,克服数据量大、结构复杂、传播速度快等问题,研究者们不断尝试各种方法来优化事件检测的效果。陈友认为网络论坛下的突发话题发现面临的关键问题是噪音,因此他提出利用词以及用户参与度的突发特性来过滤噪音[31]。赵文清等人针对微博数据稀疏性、实时性、不规范性的特点,提出根据主题词间的共现度构建词共现图的方法[32]。周刚等人注意到微博平台具备一些传统媒体不具有的特性,如关注行为、转发评论行为。他利用这些结构化信息辅助判断,以提高话题检测的性能[33]。申国伟等人针对微博消息流高度动态变化的特点,提出动态窗口选择算法。设置微博窗口调整系数 α 和滑动窗口调整系数 β,在消息流较大时,提高参数 α、β 的值,即增大两个窗口的时间片,能够提高检测粒度,在消息流大小确定时,调整参数 α 能够降低随机噪声对算法的影响。实验表明,在大规模微博消息流中,该算法能够帮助模型更早地检测到突发话题[34]。

还有一些学者关注网络问答平台的研究。黄鲁成等人结合网络问答社区的特点,采用候选关键词与组合词结合进行二次筛选的办法,降低了模糊处理与分词结果不准确带来的误差[35]。

近年来也有很多学者使用深度学习技术来解决社交媒体事件检测中的问题。石磊等人利用循环神经网络来学习词之间的关系,并作为主题模型的先验知识,使主题更加聚焦,解决了短文本稀疏性问题[36]。熊宇等人则提出一种多模态特征深度融合模型来学习事件的多模态特征表达。分别利用深层和浅层的卷积神经网络来提取图片的语义特征和学习短文本的语义信息,从而生成鲁棒性更好的多模态融合特征[37]。

(3)事件检测在突发事件中的应用(#2、#4)。

在海量数据流中检测突发事件是事件检测的研究热点之一。国内《突发事件应对法》中对突发事件做

出了相关定义:“突发事件是指突然发生,造成或者可能造成严重社会危害,需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件。”突发事件的出现会给人们的日常生活、人身安全、财产安全带来巨大影响。因此,对突发事件的检测显得尤为重要。突发事件检测中面临的一个重要问题就是如何准确地识别突发事件。林达真等人通过考虑事件在时间分布特征上的差异来判断该事件在时间特征上是否具有突发性和关联性,从而有效去除虚假突发事件的检测[38]。王勇等人提出一种基于“绝对聚类”的微博突发词文本聚类算法(ACFD 算法)。其思想是如果某一个对象属于既有的一个类,那么它应该和这个类中的每一个对象都相似,即“绝对”属于这个类,否则不属于这个类。并对聚类结果进行热度加权计算,返回各类簇中热度最大的微博作为突发事件的检测结果[39]。

实际情况中,突发事件是经常带有地域属性的,仲兆满等人针对地域性突发事件的检测,提出了地域 Top-k 突发事件检测的系统框架,将地域信息作为突发词提取的指标和热度计算指标之一[40]。李纲等人则关注突发事件的演化规律,结合地理标签和个人信息描述对受灾地区用户和非受灾地区用户进行自动划分,比较两类用户在宏观层面和微观层面的热点话题演化规律。可以帮助灾害管理部门更高效地从社交媒体数据中识别受灾人群及其需求,从而及时采取响应措施[41]。

突发事件总会带来大量的网络舆情,对于网络舆情的识别也是事件检测的研究热点之一。网络舆情具有自由性、交互性、多元性、突发性、群体极化性等特点,能够影响民众的情感和判断,能推动和改变事件的发展和走向,容易被不怀好意的群体利用,已经成为影响社会稳定的重要因素。因此,及时检测、控制并引导舆情的发展具有十分重要的意义。丁杰等人设计了一个网络舆情监控系统 IPSMS,应用了网页清洗及 k-d tree 分类方法,将网络新闻及论坛、BBS 上的帖子依关键词搜索,并依“事件”聚类,让管理者通过阅读事件可以了解正在发生或已经发生的事件[42]。李磊等人关注网络舆情的态势演化,他在对主题词频数进行加权的基础上,计算词对的最大信息系数(MIC)。基于 MIC 计算的主题词集合的密度和中心度充分揭示了话题内容的演化趋势[43]。王曰芬等人以新闻媒体报道来表达社会现实事件、以公众评论来表达舆情事件,通过话题识别与主题关联分析,探究同一事件新闻报道与舆情评论之间的共振与偏离[44]。冯科等人将网络舆情事件发现与分类的复杂问题,分解到三个模型中:基于深度学习的事件句检测模型 ESDM、事件类型判别模型 ETDM 和网络舆情事件专家知识模式库 EKB。三个模型组成的联合模型有效降低了网络舆情重大事件检测的漏判和误判[45]。

4　事件检测研究趋势分析

时间线图谱可以了解聚类之间的关系以及某个聚类中文献的历史演进趋势[46]。因此,根据时间线的变化可以更清晰地了解事件检测领域的发展变化,时间线图谱如图 4 所示。突现关键词表示在一段时期内该研究主题受到了高度关注,近年关键词突现信息如表 3 所示。

表 3　关键词突现信息

关键词	强度	时间跨度
突发事件	4.66	2019-2021
事件检测	4.16	2019-2021
社交媒体	2.90	2019-2021
深度学习	2.33	2018-2021
突发话题	2.24	2020-2021

一个研究领域,一般先经过最初的概念形成阶段,然后随着研究工具的大量出现,研究的能力和范围开始增强,此后进入扩散阶段,研究者将这些方法应用到原本的研究问题之外的领域,最后进入衰减阶段[47]。基于该理论可以看出:2003-2009 年是事件检测领域的概念形成阶段。这一阶段较大的节点是“话题检测”

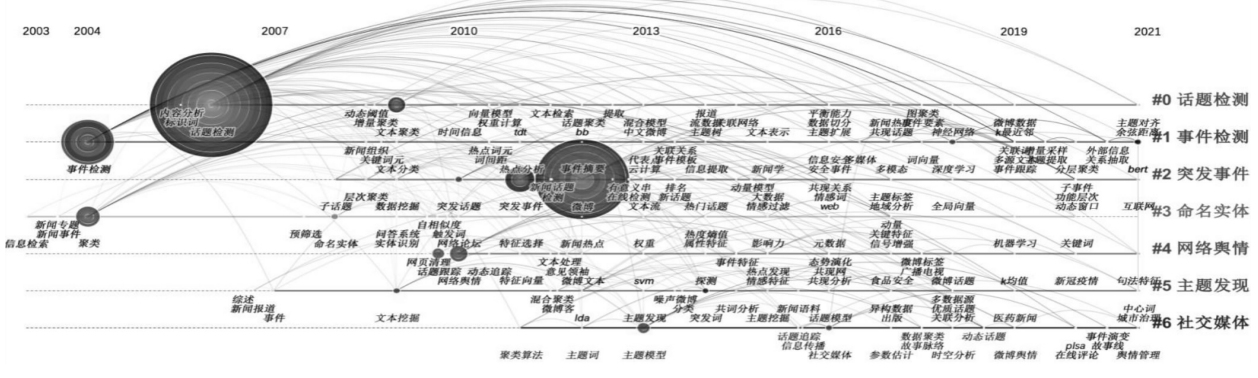

图 4　时间线图谱

“事件检测”和“聚类”。国内的研究刚刚起步,许多方面还不能满足实际应用的需要。所以学者们更多的是对概念的研究,使用的方法也多局限于聚类和信息检索,例如命名实体、增量聚类、层次聚类和文本挖掘。这一阶段的研究热点是从新闻长文本中检测事件,所以主要的研究对象是新闻专题、新闻事件、新闻报道和新闻组织等。在这一时间段内的凸显关键词是“融合特征”和“命名实体”。

2010-2021 年是事件检测领域的工具开发阶段。为克服传统方法的各种缺陷,研究者们不断对检测技术进行改进和完善。这一阶段最大的变化就是微博等社交媒体的流行,彻底改变了事件检测研究的数据类型。研究主题与上一阶段相比成倍增长,“网络舆情”“主题模型”“神经网络”“bert 模型”等内容获得了研究者的大量关注。研究方法更是多种多样,自然语言处理、文本挖掘、主题模型、多模态、深度学习等技术都被应用在该领域。同时,也出现了领域扩散的现象,研究方向不再局限于对技术的探索,已有部分学者将事件检测应用于舆情管理、应急管理、信息安全、食品安全、广播电视、城市治理等领域。

“突发事件”“社交媒体”和“深度学习”是近三年值得关注的突现词。近年来国内处于突发事件高发阶段,新冠疫情、电动车电池爆炸、城市洪水等突发性灾害事件引起人们的广泛关注。越来越多的研究者和应急管理人员意识到事件检测在应对突发事件中的重要性。而应急管理需要即时访问各种数据源,了解灾难发生期间现场的情况以及各种信息。社交媒体就是当前最重要的信息发布和传播渠道之一。人们能够通过社交媒体主动或被动分享有价值的事件信息,并传递给应急管理人员、决策者或能够提供帮助的人。因此,如何更有效地利用社交媒体中的信息是当前研究者和管理者都在不断探索的问题。深度学习已经成为机器学习的研究热点,它被广泛运用于自然语言处理、图像识别、物体检测等领域,使人工智能等相关技术取得了很大的进步。深度学习不需要人工提取特征,大幅提高了事件检测的效率,同时它能更好地挖掘文本的隐藏特征,使事件检测的结果更加准确。因此,继续探究基于机器学习的事件检测方法将是未来该领域一个重要的研究方向。除此以外,这一阶段的发文量有显著增长,研究角度也更加深入和细化,如子事件检测、事件演化和事件脉络挖掘等方面的研究。

5 结束语

运用文献计量的方法和知识可视化软件 CiteSpace 对事件检测研究成果进行梳理和分析,得出以下结论:

(1)事件检测领域发文量已经趋于稳定,核心期刊占比整体上呈上升趋势,说明对事件检测研究的质量和深度都在提高。中国科学院发文数量最多,但与其他机构的交流合作需要进一步提升,同时其他研究机构之间的合作也较少,从长远来看不利于事件检测领域的发展。各机构之间,尤其是不同学科之间应该加强合作,呈现多样化和交叉性发展态势,有利于事件检测研究的跨学科创新发展。

(2)梳理了研究者在事件检测中应用的方法和技术。虽然方法多种多样,但是很多研究者使用的实验数据是英文语料或 Twitter 等国外平台的数据。面对结构和语义都颇为复杂的中文文本,研究者们还需要继续深化中文数据的处理能力,提出更加高效、精准的检测方法。

(3)在研究热点和研究重点方面,事件检测的研究热点集中在突发事件与热点话题的文本事件检测应用研究、基于微博数据的事件检测案例研究、以聚类为主要方法的事件检测方法研究这三个方面。当前研究重点是事件检测技术,社交媒体事件检测和事件检测在突发事件中的应用。

演化趋势分为两个阶段,2003-2009 年是事件检测领域的概念形成阶段,2010-2021 年是事件检测领域的工具开发阶段。同时出现了领域扩散的现象,研究者将事件检测应用到其他领域,如舆情管理、应急管理、信息安全、食品安全、广播电视、城市治理等。

未来的研究方向包括社交媒体、突发事件、深度学习和突发话题。基于社交媒体的突发事件检测是事件检测领域一个主要的研究方向,如何准确、实时地检测突发事件并对事件的发展进行追踪,是研究者们当前以及未来一段时间关注的焦点。同时可以预见,深度学习将成为未来事件检测的研究重点,将深度学习与自然语言处理结合,可以显著提高事件检测的效率和效果,使事件检测在各个领域的应用具有更好的表现。

参考文献:

[1] JOHNSON N F, ZHENG M, VOROBYEVA Y, et al. New online ecology of adversarial aggregates: ISIS and beyond [J]. Science, 2016, 352(6292): 1459-1463.

[2] 陈 悦,陈超美. 引文空间分析原理与应用:CiteSpace 实用指南[M]. 北京:科学出版社,2014.

[3] 洪 宇,张 宇,刘 挺,等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报,2007,21(6):71-87.

[4] 李保利,俞士汶. 话题识别与跟踪研究[J]. 计算机工程与应用,2003,39(17):7-10.

[5] 贾自艳,何 清,张海俊,等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展,2004,41(7):1273-1278.

[6] 骆卫华,于满泉,许洪波,等. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报,2006,20(1):29-36.
[7] 宋　丹,王卫东,陈　英. 基于改进向量空间模型的话题识别与跟踪[J]. 计算机技术与发展,2006,16(9):62-64.
[8] 郑斐然,苗夺谦,张志飞,等. 一种中文微博新闻话题检测的方法[J]. 计算机科学,2012,39(1):138-141.
[9] 路　荣,项　亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能,2012,25(3):382-387.
[10] 马雯雯,魏文晗,邓一贵. 基于隐含语义分析的微博话题发现方法[J]. 计算机工程与应用,2014,50(1):96-100.
[11] AKCORA C G,BAYIR M A,DEMIRBAS M,et al. Identifying breakpoints in public opinion[C]//Proceedings of KDD workshop on social media analytics. Washington:[s. n.],2010.
[12] 杨　亮,林　原,林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报,2012,26(1):84-90.
[13] 李伯华,罗　琴,刘沛林,等. 基于 Citespace 的中国传统村落研究知识图谱分析[J]. 经济地理,2017,37(9):207-214.
[14] 赖　勇,阳富强. 基于 CNKI 数据库的安全疏散文献计量学分析[J]. 安全与环境工程,2018,25(6):114-119.
[15] 张　阔,李涓子,吴　刚,等. 基于词元再评估的新事件检测模型[J]. 软件学报,2008,19(4):817-828.
[16] 商宪丽,王学东. 微博话题识别中基于动态共词网络的文本特征提取方法[J]. 图书情报知识,2016(3):80-88.
[17] 刘素芹,柴　松. 命名实体的网络话题 K-means 动态检测方法[J]. 智能系统学报,2010,5(2):122-126.
[18] 孙　乐,韩先培. 中文信息处理发展报告 第八章 信息抽取研究进展、现状及趋势[R]. 北京:中国中文信息学会,2016.
[19] 张先飞,郭志刚,刘　嵩,等. 基于触发词指导的自相似度聚类事件检测[J]. 计算机科学,2010,37(3):212-214.
[20] 税仪冬,瞿有利,黄厚宽. 周期分类和 Single-Pass 聚类相结合的话题识别与跟踪方法[J]. 北京交通大学学报,2009,33(5):85-89.
[21] 殷风景,肖卫东,葛　斌,等. 一种面向网络话题发现的增量文本聚类算法[J]. 计算机应用研究,2011,28(1):54-57.
[22] 龙志祎,程　葳. 基于词聚类的热点话题检测算法[J]. 计算机工程与设计,2011,32(6):2214-2216.
[23] 杨长春,周　猛,叶施仁,等. 基于改进 CURE 算法的微博热点话题发现[J]. 计算机仿真,2013,30(11):383-387.
[24] 姜晓伟,王建民,丁贵广. 基于主题模型的微博重要话题发现与排序方法[J]. 计算机研究与发展,2013(S1):179-185.
[25] 郭蓝天,李　扬,慕德俊,等. 一种基于 LDA 主题模型的话题发现方法[J]. 西北工业大学学报,2016,34(4):697-701.
[26] 聂文汇,曾　承,贾大文. 基于热度矩阵的微博热点话题发现[J]. 计算机工程,2017,43(2):57-62.
[27] 孙　乐,韩先培. 中文信息处理发展报告 第八章 信息抽取研究进展、现状及趋势[R]. 北京:中国中文信息学会,2016.
[28] 侯伟涛,姬东鸿. 基于 Bi-LSTM 的医疗事件识别研究[J]. 计算机应用研究,2018,35(7):1974-1977.
[29] 张秀华,云红艳,贺　英,等. 基于卷积神经网络和 K-means 的中文新闻事件检测与主题提取[J]. 科学技术与工程,2020,20(3):1139-1144.
[30] 马晨曦,陈兴蜀,王文贤,等. 基于递归神经网络的中文事件检测[J]. 信息网络安全,2018(5):75-81.
[31] 陈　友,程学旗,杨　森. 面向网络论坛的突发话题发现[J]. 中文信息学报,2010,24(3):29-36.
[32] 赵文清,侯小可. 基于词共现图的中文微博新闻话题识别[J]. 智能系统学报,2012,7(5):444-449.
[33] 周　刚,邹鸿程,熊小兵,等. MB-SinglePass:基于组合相似度的微博话题检测[J]. 计算机科学,2012,39(10):198-202.
[34] 申国伟,杨　武,王　巍,等. 面向大规模微博消息流的突发话题检测[J]. 计算机研究与发展,2015,52(2):512-521.
[35] 黄鲁成,蒋林杉,苗　红,等. 基于网络问答社区的话题识别与分析——以知乎"老年人"话题为例[J]. 图书情报工作,2016,60(5):93-100.
[36] 石　磊,杜军平,梁美玉. 基于 RNN 和主题模型的社交网络突发话题发现[J]. 通信学报,2018,39(4):189-198.
[37] 熊　宇,张一飞,冯　时,等. 基于多模态特征深度融合的微博流事件检测与跟踪[J]. 控制与决策,2019,34(7):1409-1416.
[38] 林达真,李绍滋,曹冬林. 基于时间分布特征的博客突发事件检测[J]. 计算机工程与科学,2010,32(10):145-149.
[39] 王　勇,肖诗斌,郭跃秀,等. 中文微博突发事件检测研究[J]. 现代图书情报技术,2013(2):57-62.
[40] 仲兆满,管　燕,李存华,等. 微博网络地域 Top-k 突发事件检测[J]. 计算机学报,2018,41(7):1504-1516.
[41] 李　纲,陈思菁,毛　进,等. 自然灾害事件微博热点话题的时空对比分析[J]. 现代图书情报技术,2019,3(11):1-15.
[42] 丁　杰,徐俊刚. IPSMS:一个网络舆情监控系统的设计与实现[J]. 计算机应用与软件,2010,27(4):187-190.
[43] 李　磊,刘　继,张竑魁. 基于共现分析的网络舆情话题发现及态势演化研究[J]. 情报科学,2016,34(1):44-47.
[44] 王曰芬,许杜娟,杨振怡,等. 舆情评论与新闻报道的话题识别及其主题关联分析[J]. 现代情报,2018,38(6):3-10.
[45] 冯　科,阮树骅,陈兴蜀,等. 基于联合模型的网络舆情事件检测方法[J]. 信息安全研究,2021,7(3):207-213.
[46] CHEN Chaomei. Science mapping:a systematic review of the literature[J]. Journal of Data and Information Science,2017,2(2):1-40.
[47] SHNEIDER A M. Four stages of a scientific discipline;four types of scientist[J]. Trends in Biochemical Sciences,2009,34(5):217-223.