DOI：10.11830/ISSN.1000-5013.202108017

# 后疫情时代侨情危机状况识别方法

王华珍，孙雨洁，何霆，陆炫羽，刘晓聪

（华侨大学 计算机科学与技术学院，福建 厦门 361021）

**摘要：** 基于新冠疫情时代海外侨情危机事件频发，提出一种基于自动化信息要素抽取的新闻事件类型识别方法，对后疫情时代侨情新闻事件进行智能危机类别划分. 首先，利用爬虫技术获取特定时间段的相关侨情事件新闻，进而采用信息抽取模型对语料数据进行信息要素抽取；然后，根据要素集的取值判断每条新闻的危机事件类型；最后，对 2020 年 1 月－8 月的侨情新闻数据进行实证研究. 结果表明：该方法不但能提升侨情分析的效率，还能进行多维度的危机状况信息可视化，有助于制定危机事件应对策略.

**关键词：** 后疫情时代；侨情；危机类型；自然语言处理；信息抽取

**中图分类号：** R 181.1；D 634.3；TP 274　　**文献标志码：** A　　**文章编号：** 1000-5013(2022)06-0825-08

# Method for Identifying Crisis Situation of Overseas Chinese in Post-Epidemic Era

WANG Huazhen, SUN Yujie, HE Ting,
LU Xuanyu, LIU Xiaocong

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

**Abstract:** Based on the frequent occurrence of overseas Chinese crisis events in the era of the COVID-19 epidemic, a method for identifying news event types based on automatic information element extraction is proposed, so as to intelligently classify overseas Chinese news events in the post-epidemic era. Firstly, the crawler technology is used to obtain relevant the news of overseas Chinese events in a specific time period, and then the information extraction model is used to extract the information elements from the corpus data. Then, the crisis event type of each news is judged according to the value of the element set. Finally, it makes an empirical analysis on the overseas Chinese news data from January to August 2020. The results show that this method can not only improve the efficiency of overseas Chinese situation analysis, but also visualize multi-dimensional crisis information, which is helpful to formulate crisis response strategies

**Keywords:** post-epidemic era; overseas Chinese; crisis type; natural language processing; information extraction

2020 年 1 月 30 日，世界卫生组织（WHO）宣布，将新型冠状病毒疫情列为国际关注的突发公共卫生事件（PHEIC）[1]. 2020 年 3 月 11 日，WHO 表示，新冠肺炎疫情的爆发已经构成一次全球性的“大流行”. 新冠病毒席卷全球，海外侨胞的生活也因此受到了极大影响. 为贯彻习近平总书记关于侨务工作的

**收稿日期：** 2021-10-21

**通信作者：** 王华珍（1975-），女，副教授，博士，主要从事人工智能、机器学习、增强现实、虚拟现实的研究. E-mail：wanghuazhen@hqu.edu.cn.

**基金项目：** 中央高校基本科研业务费资助项目（TZYB-202005）；华侨大学“华侨华人研究”专项经费资助一般项目（HQHRYB2019-01）；华侨大学“海外华文教育与中华文化传播协同创新中心”项目（HJY201901）

重要论述,需要密切跟踪疫情之下海外侨情的动态,充分借助互联网开展工作,增强底线意识和风险意识,为党和国家工作大局贡献力量[2].涉侨突发事件一般是指在非中国境内突然发生的,会给华人华侨造成或可能造成严重危害或损失,需要采取应急处置措施,以应对自然灾害、事故灾害、公共卫生、社会安全、政治冲突等事件[3],其中包括侨情危机事件.因此,基于网络媒体发布的侨情危机事件新闻来研究海外侨情危机状况,具有重要的现实意义和理论价值.

目前,无论是社会科学领域还是工程技术领域,侨情危机事件的研究已成为一大热点.在社会科学领域,学者们主要关注危机事件概念辨析、危机事件构成要素及分析、危机事件的影响、危机事件政府应对策略等.如骆克任等[3]开展的全球涉侨突发事件的危害等级研究,对涉侨突发事件类型及其信息要素进行定义.在工程技术领域,学者们主要注重获取话题的主要内容、事件关系及变化趋势的分析.如李弼程等[4]构建了一种网络话题智能引导的仿真推演系统,该系统能够在仿真推演的基础上实施网络舆论引导,从而突破传统的机械性引导方式.但侨情危机事件的研究仍处于起步阶段,骆克任等[5]对海外涉侨突发事件的危机类别进行定义,并开展了实证研究,但其在进行新闻要素抽取、危机等级判断时未能实现自动化和智能化,主要仍以人工分析为主,耗费人力成本较高,效率较低.此外,目前尚缺乏针对侨情领域的智能信息处理系统,难以对侨情危机状态进行高效、智能的分析和研究.

基于此,本文采用计算机技术,对骆克任团队海外涉侨突发事件危机类别的识别过程进行复现,提出一种基于自动化信息要素抽取的新闻事件类型识别方法,旨在对后疫情时代侨情新闻事件进行智能危机类别划分和事件信息数据展示.

## 1 研究方法

提出的一种基于自动化信息要素抽取的新闻事件类型识别方法,该方法的研究流程,如图1所示.该方法的核心技术包括网络爬虫和自然语言处理技术(NLP).

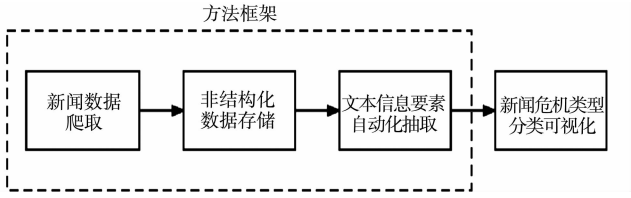


图1 侨情危机事件分类研究流程

Fig. 1 Classification and research process of overseas Chinese crisis events

### 1.1 侨情新闻数据获取

1.1.1 *数据来源* 中国侨网(http://www.chinaqw.com/)是由华声报(电子版)社主办的面对全球华侨华人提供综合性信息服务的专业网站,作为中国内地最大的侨务网络信息平台,推出了同心战"疫"信息服务平台,其内容涵盖全球六大洲际的实时海外侨胞新闻.因此,选择2020年1月—8月的中国侨网同心战"疫"信息服务平台的新闻事件数据作为研究对象,筛选出正文字数不少于200字的侨情事件新闻作为语料数据.

1.1.2 *获取方式* 网络爬虫是一种依据搜索规则自动解析网页并获取网络中符合检索要求的资源的获取程序,可从海量信息中搜寻所需信息.网络爬虫兼具获取数据的精确性与高效性,弥补了传统引擎检索的不足,被应用于自动化新闻分析与管理领域.如朱琪[6]基于网络爬虫技术开发网络舆情分析预警系统;刘娜[7]以主题爬虫和文本分类技术为基础,设计并实现了冬奥会新闻文本采集及分类分析系统.

中国侨网页面中的新闻内容是由js和模板动态加载显示的,而传统爬虫技术擅长获取HTML页面中的静态部分内容,从而无法直接对新闻正文进行爬取.因此,通过python语言环境下的selenium库调用Chrome浏览器驱动,借用Chrome的自动代理框架控制浏览器的操作,从而直接在页面获取动态加载后的新闻内容.由于爬虫获取的数据格式是纯文本,属于非结构化数据,因此,需要先将非结构化文本按新闻标题、新闻链接、发布时间、新闻正文等进行半结构化存储,构成语料数据.

从中国侨网共获取2020年1月—8月的侨情新闻数据3 432篇.为确保新闻叙述完整性,以便在后续新闻要素抽取时获取完整信息,将获取的3 432篇新闻的正文字数进行统计,筛选后获得3 277篇符合字数要求的新闻.

### 1.2 侨情危机事件评估指标体系设计

1.2.1 *事件及其要素的相关理论* 事件是指在某个特定的时间和环境下发生的、由若干角色参与、表现出若干动作特征的一件事情.事件抽取是将文本中描述的事件识别出来并提取事件中的各个信息要素的技术.随着人工智能技术的发展,利用计算机自动抽取文本中的事件要素信息,实现事件自动识别

的方法，在舆情监测、文本摘要、自动问答、事理图谱自动构建等领域有着重要应用. 据此形成了一系列基于开源公共数据集的事件抽取竞赛活动，如 2005 年成立的自动内容抽取（ACE）竞赛（简称 ACE2005）. 而在真实新闻事件中，新闻文本句式繁杂、表述多样，这为事件抽取任务带来了挑战.

根据具体的应用场景和问题焦点，事件的信息要素定义标准和要素集合将呈现不同的形式. 刘宗田等[8]将事件定义为六元组，包含动作、对象、时间、环境、断言和语言表现；而 ACE2005 将事件的信息要素进一步定义为事件类型、事件触发词、事件论元及论元角色. 本研究采用 ACE2005 中的定义标准，其中，事件触发词指的是事件发生的核心词，通常为动词或名词，如遇害、受伤；事件论元指的是事件的参与者，通常由人物、时间、地点、值等组成；论元角色指的是事件论元在事件中充当的角色，如人物实体可划分为攻击者和受害者[9].

1.2.2　指标体系的设计　《中华人民共和国突发事件应对法》第 3 条所述的突发事件是指突然发生的，造成或可能造成严重社会危害，需要采取应急处置措施以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件[10]. 上述定义是按照事件发生领域进行划分的. 骆克任等[3]提出，涉侨突发事件主要是由政治矛盾、社会犯罪、意外事故和民族排斥等引发的危机事件，其划分依据是造成危机事件的原由. 而从危机事件造成的后果来划分，将危机事件划分为生命损失事件、财产损失事件、其他事件 3 种危机类型.

生命损失事件主要是指侨情突发事件中涉及海外侨胞生命损失的事件，包括但不限于人员死亡、受伤. 定义生命损失事件的触发词包含死亡、死去、遇害等词语，论元角色包括受害者、工具、地点. 财产损失事件主要是指侨情突发事件中涉及海外侨胞财产损失的事件，包括但不限于抢劫财产、偷盗. 定义财产损失事件的触发词包含偷窃、遗失、抢劫、诈骗等词语，论元角色包括受害者、方式、地点.

如新闻描述："在约翰内斯堡，南部非洲齐鲁同乡总会会长夫妇在下班途中，遭遇 3 名抢匪持枪械进行武装抢劫，不幸遇害身亡."这条侨情危机新闻同时包含了生命损失事件和财产损失事件，其包含的事件类型及要素汇总，如表 1 所示.

表 1　侨情危机新闻包含的事件类型及要素汇总

Tab. 1　Summary of event types and elements included in overseas Chinese crisis news

| 事件类型 | 触发词 | 论元 | 论元角色 |
|---|---|---|---|
| 生命损失事件 | 遇害、身亡 | 非洲齐鲁同乡总会会长夫妇 | 受害者 |
| | | 枪械 | 工具 |
| | | 约翰内斯堡 | 地点 |
| 财产损失事件 | 抢劫 | 非洲齐鲁同乡总会会长夫妇 | 受害者 |
| | | 抢劫 | 方式 |
| | | 约翰内斯堡 | 地点 |

为了更准确地划分事件类型，制定了生命损失事件和财产损失事件的触发词词典，如表 2 所示. 该词典使用中文突发事件语料库（Chinese emergency corpus，CEC）并结合专家整理得到.

表 2　事件触发词词典

Tab. 2　Event trigger word dictionary

| 事件类型 | 事件触发词集 |
|---|---|
| 生命损失事件 | 死亡、死伤、丧生、遇难、身亡、伤亡、遇害、杀害、去世、离世、谋杀、死、伤、轻伤、重伤、负伤、受伤、受轻伤、受重伤、重创、枪伤、刀伤、捅伤、袭击、自杀、杀死、杀、亡 |
| 财产损失事件 | 抢劫、劫持、偷窃、偷盗、盗窃、诈骗、勒索、打劫、掠夺、洗劫、强抢、抢掠、劫掠、侵夺、骗、损失财产 |

针对文中研究的侨情新闻数据，不属于生命损失事件和财产损失事件的其他类型事件统一视为其他事件. 其他事件的触发词指的是排除表 2 列出的事件触发词，其他事件的论元不需要抽取. 由表 1 展示的事件的触发词、论元、论元角色的定义及表 2 的事件触发词词典，构成了文中设计的侨情危机事件类型指标体系.

## 2　基于 NLP 的侨情危机事件要素抽取技术

从非结构化的新闻正文文本中抽取出侨情危机事件要素，实现事件信息结构化的相关技术，进而根据结构化的事件信息数据识别出事件的类别，实现事件信息的可视化. 由于事件各要素的属性、范畴、性

质都不相同,需要分别研发不同事件要素的抽取技术.

### 2.1 地点要素抽取

侨情事件新闻地点要素抽取研究能够反映各个地区的移民安全状况,为移民安全指数分析提供技术支持,对侨情类型识别具有重要的现实意义.采用深度学习模型和知识推理法实现地点要素抽取任务.首先,对新闻正文和新闻标题的文本运行深度学习模型,抽取出地点实体词汇;然后,采用知识推理法推理出地点的"地区-国家-大洲"三层次地理位置描述集.

2.1.1 *地点抽取模型* 地点要素抽取是自然语言处理常见的信息抽取任务,相对应的抽取理论和模型工具非常多.随着人工智能深度学习的发展,将地点抽取问题转换成序列标注问题,再采用深度神经网络进行端对端学习的方式一般能得到最好性能的模型,因此成为解决地点抽取任务的首选方案.李芳芳等[11]基于图模型和膨胀卷积神经网络,提出交通事件要素抽取算法,针对交通事件文本中的地点要素进行了抽取.

文中选用的深度神经网络模型是结构化预测模型,它是线性条件随机场(Linear-chain CRF)的改良模型,能实现地点要素的抽取.结构化预测模型优化了传统的以字符为单位进行编码的方式,在传统标注的基础上加入 Tri-gram 特征,使高阶预测变量之间的关系同样能够被捕捉[12].该模型由自然语言处理平台 BosonNLP 提供,采用 API 接口方式进行调用.BosonNLP 平台在分词和词性标注中融合了半监督学习的方式,即使用在大规模无标注数据上的统计数据来改善有监督学习中的标注结果,使序列标注的准确率得到提升[13].针对新闻文本的特异性进行参数调优,在最优参数配置的条件下获得新闻事件地点的抽取结果.

2.1.2 *层次地理位置推理* 侨情事件具有全球性,事件发生的地理信息可分别从地区、国家、大洲这三层次进行描述,即地区是侨情事件地点的最小单元地理层面,大洲层面为最大单元地理层面.然而,侨情新闻具有本地化特点,每篇新闻中抽取得到的地点信息要素难以在地理等级上做到统一,且文本中一般不会全部出现地区、国家、大洲的三层次地理信息.因此,需对抽取到的地理信息实体进行知识推理,以同时获取地区、国家和大洲的三层次地理信息.

首先,构建"地区-国家"字典与"国家-大洲"字典;然后,利用 python 编程方式进行知识推理,实现对每一条新闻的事件地点要素实体进行地区、国家、大洲层面的归类.若新闻中抽取到的地点实体为地区,则自动推理出该地点实体所隶属的国家和大洲.由此,最后得到的新闻事件地点要素是"地区-国家-大洲"三层次地理位置描述集 $S=\{\text{area},\text{country},\text{continent}\}$.

### 2.2 事件触发词抽取

采用词典语义匹配法进行事件触发词的抽取.首先,对新闻正文文本进行分词处理,以筛选出候选触发词;然后,采用词典语义匹配法计算出语义匹配的触发词.

2.2.1 *分词处理* 分词处理过程包括分句、分词和词性筛选.由于事件触发词一般为动词、名词、动名词(表 2),因此,需要筛选出动词、名词、动名词并进行词频统计,从而获得触发词候选列表,即

$$L_1=\{(t_1,s_1),(t_2,s_2),\cdots,(t_N,s_N)\}. \tag{1}$$

式(1)中:$t$ 为筛选出的触发词;$s$ 为各词词频;$N$ 为触发词数量.

设置筛选阈值 $s_i>1$,获得高频触发词候选列表 $L_2$.考虑到新闻标题作为新闻正文内容的高度概括,通常包含事件发生地点、事件主要人物、事件核心信息等要素,因此,对新闻标题进行触发词抽取处理,从而获得标题高频触发词候选列表,把它与 $L_2$ 并集得到最终触发词候选列表 $L_3$.

2.2.2 *词典语义匹配* 考虑到词汇表达的多样性和模糊性,如果对最终触发词候选列表 $L_3$ 和事件触发词词典(表 2)进行关键词匹配法,将无法全面准确地抽取到事件所需要素信息,进而影响到事件危机类型的判断效果,因此,引入词向量表示法,把相关各个词汇表达成一阶高维向量,在向量空间中计算词汇之间的距离.研究证明,采用深度自然语言处理技术,如 word2vec[14],BERT[15]等,构建的词向量具有语义一致性,即两个词之间的向量距离越小,其语义相似性越强.对最终触发词候选列表 $L_3$ 中的词汇进行 BERT 向量表达,获得触发词候选矩阵 $\boldsymbol{W}$,同时,对事件触发词词典(表 2)的词汇进行 BERT 向量表达,获得触发词词典矩阵 $\boldsymbol{D}$,其具体表达式为

$$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_k], \tag{2}$$

$$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_m]. \tag{3}$$

式(2)，(3)中：$k$ 为触发词候选矩阵 $\boldsymbol{W}$ 中候选触发词个数；$m$ 为事件触发词词典矩阵 $\boldsymbol{D}$ 中词的个数；$\boldsymbol{d}$ 为事件触发词向量；$\boldsymbol{w}$ 为候选触发词向量.

相似度计算采用余弦相似度算法计算，即通过计算候选触发词向量 $\boldsymbol{w}$ 与事件触发词向量 $\boldsymbol{d}$ 夹角的余弦值来判断对应词向量的相似度. 一般地，夹角越小，余弦值越大，两个词向量语义越相似. 设向量维度为 $n$，$\boldsymbol{w}$ 与 $\boldsymbol{d}$ 的相似度 sim 的计算式为

$$\mathrm{sim}(\boldsymbol{w},\boldsymbol{d})=\frac{\left(\sum_{i=1}^{n}\boldsymbol{w}_i\times\boldsymbol{d}_i\right)}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{w}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{d}_i)^2}}. \tag{4}$$

根据 sim 值的大小，可以判断各个候选触发词的语义重要性值. 分别筛选出与生命损失事件触发词集 $E$、财产损失事件触发词集 $M$(表 2)相对应的最大语义重要性值 $\mathrm{sim}^E_{\max}$，$\mathrm{sim}^M_{\max}$，以及相对应的触发词 $w^E_{\max}$ 和 $w^M_{\max}$. 取语义重要性值 0.7 为阈值，若 $\mathrm{sim}^E_{\max}$，$\mathrm{sim}^M_{\max}$ 均大于等于 0.7，则该新闻同时描述了生命损失事件和财产损失事件；若 $\mathrm{sim}^E_{\max}$，$\mathrm{sim}^M_{\max}$ 均小于 0.7，则该新闻描述的事件类型为其他事件.

## 3　研究结果与数据可视化

运用前述研究技术，对爬虫获取的 3 277 篇侨情事件新闻文本进行地点要素抽取模型和事件类别识别的研究，前者的关键技术是地点要素抽取，后者的关键技术是触发词抽取.

### 3.1　地点要素抽取模型研究结果及可视化

3.1.1　*地点要素抽取结果*　对 3 277 篇侨情事件新闻实行地点要素抽取，其中，有 3 059 篇新闻文本成功实现了地点要素抽取，218 篇新闻未能成功抽取地点要素. 由于新闻描述的简略性，每篇新闻包含的地点实体不一定具有“地区-国家-大洲”三层次地理位置，因此，采用 Linear-chain CRF 抽取 5 211 个地理位置实体词，其中，地区词 3 277 个，国家实体词 23 个，大洲实体词 34 个. 针对这 218 篇新闻进行人工语义分析，并进行地点信息要素抽取，获得新闻事件发生的地理要素实体. 对地点要素抽取模型的性能进行评估，得到模型的准确率为 96.67%，精确率为 100.00%，召回率为 93.35%，精确度和召回率的调和平均值(F1 值)为 96.56%.

3.1.2　*“地区-国家-大洲”三层次地理位置推理结果*　由地点要素抽取结果可知，3 277 篇侨情事件新闻的“地区-国家-大洲”三层次地理位置实体集尚不完整，需要继续采用知识推理获得完整的三层次地理位置实体集. 推理结果汇总为地区实体词有 2 485 个，国家有 2 977 个，大洲有 3 277 个，考虑到向下推理路径无法实现的局限性，地区和国家这两个层次的地点将无法全部获取. 文中方法显示了获取每条新闻三层次地理位置信息要素的有效性，为后续的数据可视化奠定了基础.

3.1.3　*地点要素抽取结果的可视化*　收集的 3 277 篇侨情新闻中，各大洲危机事件的数量分布，如图 2 所示.

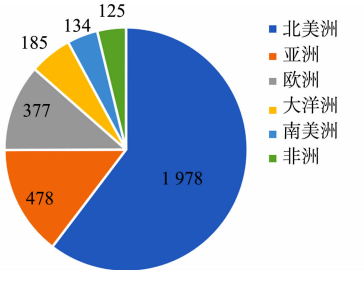


图 2　各大洲危机事件的数量分布

Fig. 2　Number and distribution of crisis events on all continents

由图 2 可知：北美洲危机事件的总数目在 6 大洲中排行第一，是排行第二的亚洲的 4 倍；其次是欧洲、大洋洲；非洲和南美洲的危机事件总数并列最少. 根据中国经济网报道[16]，2020 年 6 月 11 日，美国三大股指出现暴跌，北美经济受疫情影响极大. 新冠疫情对北美洲国家的经济产生了极大冲击，撼动了北美资本主义国家政治及人文的发展，社会的稳定性被打破，严重影响了北美侨胞的日常生活，大量危机事件也随之而来，使北美洲危机事件总数位居洲际第一.

### 3.2　事件类型识别结果及其可视化

为了评估文中提出的词典语义匹配法对事件类型识别的智能化效果，需对算法结果进行人工审核，以统计算法的准确率. 考虑到人工审核需要耗费大量的人力和时间成本，选择对 3 277 篇新闻数据进行洲际等比例抽样，获得精简数据集，数据规模为 369，再对精简数据集进行事件类型识别研究. 精简数据集的洲际分布情况，如表 3 所示.

表3 精简数据集的各大洲分布情况

Tab.3 Continent distribution of reduced data sets

| 项目 | 北美洲 | 亚洲 | 欧洲 | 大洋洲 | 南美洲 | 非洲 |
|---|---|---|---|---|---|---|
| 事件数 | 220 | 54 | 43 | 22 | 15 | 15 |

3.2.1 事件类型识别结果 根据文中方法获取到369篇新闻的事件类型,各大洲事件类型识别结果,如表4所示.

表4 中:以北美洲为例,220条新闻中有35条新闻包含生命损失事件,118条新闻包含财产损失事件,70条新闻属于其他事件,其中,有6篇新闻既包含生命损失事件,又包含财产损失事件.其他洲的事件类型识别结果也按照该规律进行统计.

表4 各大洲事件类型识别结果

Tab.4 Recognition results of event types on all continents

| 大洲 | 生命损失事件数 | 财产损失事件数 | 其他事件数 |
|---|---|---|---|
| 北美洲 | 37 | 119 | 70 |
| 亚洲 | 27 | 17 | 14 |
| 欧洲 | 2 | 3 | 38 |
| 大洋洲 | 6 | 10 | 9 |
| 南美洲 | 1 | 0 | 14 |
| 非洲 | 8 | 3 | 4 |

3.2.2 事件类型识别结果的可视化 可视化的基础维度包括月份、地理、事件类型,由此可以构建事件类型月分布、涉事国家月分布、涉事洲际月分布.多维度统计项的相关指标,如表5所示.

用Excel的数据透视图工具进行数据可视化制作.数据透视表具有表格"透视"的能力,可以挖掘出数据中隐藏的关系,将纷繁的数据有序化,以供研究使用[17].将数据透视技术应用到侨情危机状况研究中,可以实现数据集、透视表、可视化图形的实时联动反应,从而增强数据可视化的交互质量.首先,将研究得到的369篇侨情事件新闻数据作为可视化的语料数据集;其次,根据月份、地理、事件类型等3种可视化的基础维度,结合侨情事件新闻数据分析的需求,设计不同数据透视表的行字段、列字段和求和项,从而得到多个(10个)数据透视表;最后,将语料数据集与数据透视工作表创建关联,通过数据透视图左侧的切片器实现月份、地理、事件类型等3种维度下的数据透视图.

表5 多维度统计项的相关指标

Tab.5 Related indicators of multi-dimensional statistical items

| 统计项 | 指标类别 | 指标数量 |
|---|---|---|
| 事件类型月分布 | 月份、生命损失事件数、财产损失事件数、事件总数、平均生命损失、平均财产损失 | 6 |
| 涉事国家月分布 | 月份、国家、频数 | 3 |
| 涉事洲际月分布 | 月份、大洲、频数 | 3 |

全球危机事件概况可视化结果,如图3所示.图3中:左侧为月份选取栏,使用者可以通过点选不同月份来获知对应月份的危机事件概览结果,实现简单的交互功能;右侧为可视化结果展示栏,通过柱状图的形式直观地展示了对应月份各大洲的危机事件总数,并以折线图的形式分别展示了各大洲生命损失事件和财产损失事件的计算分数,有利于各大洲不同类型危机事件的横向比对.

以3月份的全球概览为例,危机事件数量最多的是北美洲,有35起.针对生命损失而言,亚洲的占比最高;而对于财产损失而言,亚洲的占比也最高.这说明,虽然北美洲危机事件数量最多,但大多数应该属于其他事件类型.

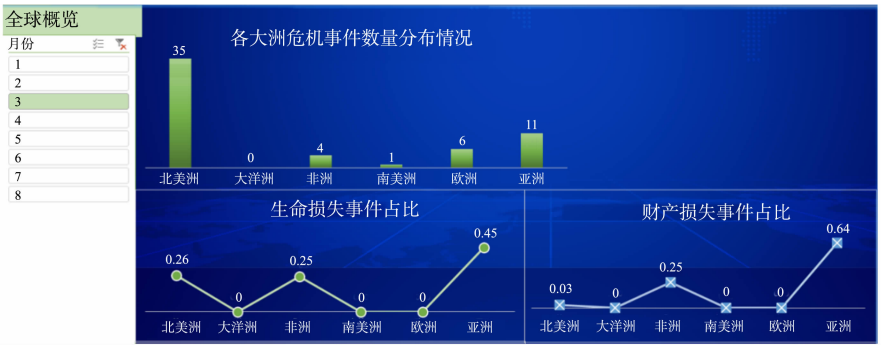


图3 全球危机事件概况可视化结果

Fig.3 Overview on visualized results of global crisis event

新闻来源分析可视化结果，如图 4 所示. 图 4 中：左侧为大洲选取栏，使用者可以通过点选不同大洲来获知对应洲的危机事件来源分析结果，实现简单的交互功能；右侧为可视化结果展示栏，通过饼状图展示了各个国家危机事件的总数，并结合条形图综合展示了各国在不同月份发生危机事件的状况. 该可视化结合了地点要素抽取的结果，针对各大洲中各个国家每月报道的危机事件数量，在地理层面上进行细分，更详细地展示了各大洲危机事件的来源地.

具体地，以亚洲为例，危机事件数量最多的国家为马来西亚，有 29 起. 马来西亚的月分布也是比较密集的：2 月 9 起；3 月 2 起；4 月 5 起；5 月 13 起.

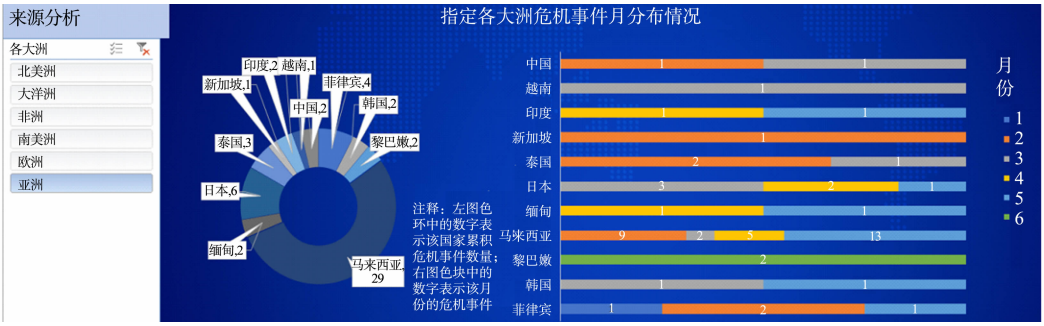


图 4　新闻来源分析可视化结果

Fig. 4　Visualized results of news source analysis

月份事件统计可视化结果，如图 5 所示. 使用者通过点选左侧的月份，在右侧以面积图的形式更直观地展示各国在对应月份中的危机事件数量，同时以饼状图的形式展示对应月份中各大洲的危机事件数量. 以 8 月份为例，全球视角下危机事件数量最多的国家是美国，有 25 起.

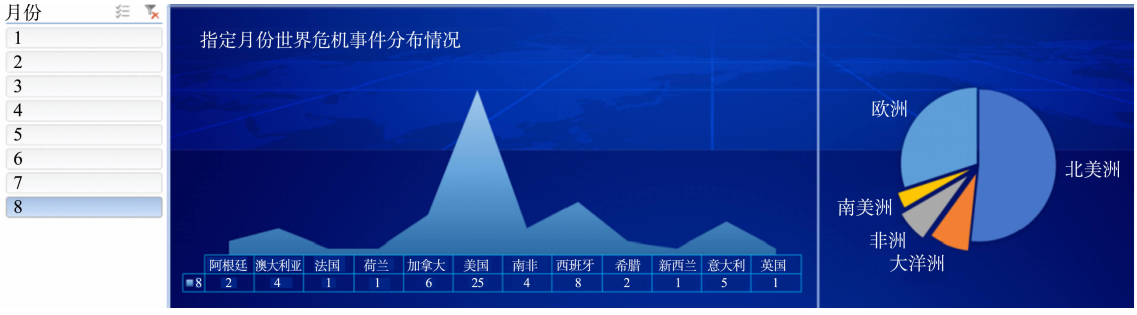


图 5　月份事件统计可视化结果

Fig. 5　Visualized results of monthly event statistics

## 4　结论

提出一种基于自动化信息要素抽取的新闻事件类型识别方法，采用爬虫技术实现新闻数据获取，极大程度地节省了人力查找、获取数据的成本. 在数据的处理部分，文中方法结合自然语言技术和语义词典方法，对新闻文本进行地点要素实体和触发词实体的智能抽取，实现事件类型的智能识别，省去了人工阅览新闻内容、手动划分的过程. 对海外涉侨危机事件危机类别识别的推理方法涉及新闻实时爬取和新闻危机类别判定两个过程，所涉及的时间复杂度也由这两部分组成. 其中，新闻爬取一般采用周期性定时爬取方式，而推理算法的时间复杂性为 $O(N)+O(N^2)$，其中，$O(N)$ 为分词复杂度；$O(N^2)$ 为触发词字典语义匹配过程. 最后，采用数据透视技术实现月份、地理、事件类型等三维度的数据聚合可视化，详细展示了侨情危机的具体情况和变化过程，从而获得侨情危机信息的多角度解析. 综上可知，文中方法能提升侨情分析的效率，且可进行多维度的危机状况信息可视化，有助于制定危机事件的应对策略.

然而，文中方法对文本中暗含的事件地理位置信息无法实现自动抽取，仍需要依靠人工进行手动提取地理信息要素，因此，后续工作可以深入研究语义挖掘相关的技术，以提升隐含地理信息要素的抽取成功率. 同时，由于定义的触发词词典数量有限，不能完全覆盖该危机事件类型中的全部触发词，所以在后续工作中仍然需要不断扩充触发词词典，以提高事件划分的召回率. 另外，关于新闻危机事件研究的

最新发展现状,如舆论引导、传播策略与信息建构等,并未见针对侨情领域的相关研究[18-20],这也是未来可继续探索的方向.

**参考文献:**


[1] 人民日报社. 世卫组织将新型冠状病毒疫情列为国际关注的突发公共卫生事件[EB/OL]. (2020-01-31)[2021-03-28]. https://baijiahao.baidu.com/s?id=1657184212860467637.

[2] 中国共产党新闻网. 2020习近平总书记关于侨务工作重要论述研讨会在南京召开 万立骏出席并讲话[EB/OL]. (2020-09-15)[2021-03-28]. http://cpc.people.com.cn/gb/n1/2020/0915/c432352-31861887.html.

[3] 骆克任,王超,谢婷婷. 全球涉侨突发事件的危害等级研究[M]// 丘进. 华侨华人研究报告(2013). 北京:社会科学文献出版社,2013.

[4] 李弼程,熊尧,黄涛,等. 网络舆论智能引导仿真推演模型与系统构建[J]. 国防科技,2020,41(5):35-40. DOI:10.13943/j.issn1671-4547.2020.05.07.

[5] 骆克任,丘进,王超,等. 海外同胞安全研究:安全预警与风险应对[M]. 北京:社会科学文献出版社,2018.

[6] 朱琪. 基于网络爬虫的舆情分析预警系统设计[J]. 电子设计工程,2020,28(22):56-60. DOI:10.14022/j.issn1674-6236.2020.22.013.

[7] 刘娜. 冬奥会新闻文本采集及分类分析系统的设计与实现[D]. 邯郸:河北工程大学,2020.

[8] 刘宗田,黄美丽,周文,等. 面向事件的本体研究[J]. 计算机科学,2009,36(11):189-192,199. DOI:10.3969/j.issn.1002-137X.2009.11.046.

[9] 秦彦霞,张民,郑德权. 神经网络事件抽取技术综述[J]. 智能计算机与应用,2018,8(3):1-5,10. DOI:10.3969/j.issn.2095-2163.2018.03.002.

[10] 中国人大网. 中华人民共和国突发事件应对法[EB/OL]. (2013-04-16)[2021-03-28]. http://www.npc.gov.cn.

[11] 李芳芳,路毅恒,毛星亮. 基于图模型和膨胀卷积神经网络的交通事件要素抽取算法:110781393A[P]. 2020-02-11.

[12] OSCHINA. BosonNLP分词技术解谜[EB/OL]. (2015-10-22)[2021-03-28]. https://www.oschina.net/question/2448846_2138515.

[13] MIN K,MA Chenggang,ZHAO Tianmei,*et al*. BosonNLP: An ensemble approach for word segmentation and POS tagging[C]// NLPCC 2015:Natural Language Processing and Chinese Computing. Nanchang:Springer International Publishing,2015:520-526.

[14] MIKOLOV T,CHEN Kai,CORRADO G,*et al*. Efficient estimation of word representations in vector space[C]// ICLR 2013 Conference Track. Arizona:arXiv,2013:1-12.

[15] DEVLIN J,CHANG Mingwei,LEE K,*et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minnesota:Association for Computational Linguistics,2019:4171-4186. DOI:10.18653/v1/N19-1423.

[16] 中国经济网. 当重启遭遇疫情抬头美股周四交易日暴跌[EB/OL]. (2020-06-12)[2021-03-28]. https://baijiahao.baidu.com/s?id=1669245895341930874.

[17] 张勇. Excel数据透视表在开放教育学籍数据统计中的应用[J]. 电脑知识与技术,2014,10(13):3050-3052.

[18] 唐净欣. 公共危机事件中电视新闻舆论引导方法研究[J]. 新闻研究导刊,2019,10(19):174,176. DOI:10.3969/j.issn.1674-8883.2019.19.104.

[19] 李涵舒. 论危机事件中新闻媒体的传播策略[J]. 新闻研究导刊,2020,11(3):156,175. DOI:10.3969/j.issn.1674-8883.2020.03.090.

[20] 侯向平. 危机事件传播信息框架建构与类型研究与分析[J]. 公关世界,2016(9):45-47.


(**责任编辑**:黄晓楠 **英文审校**:吴逢铁)