引用格式：元方，卢伟，沈浩．基于无监督技术的中文新闻事件数据构建与分析[J]．中国传媒大学学报（自然科学版），2023，30(05)：01-09.
文章编号：1673-4793(2023)05-0001-09

基于无监督技术的中文新闻事件数据构建与分析

元方，卢伟，沈浩*
（中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：本文针对面向媒介和传播学研究的中文新闻事件数据构建任务进行探索，利用自然语言处理、深度学习和无监督聚类等技术，构建了一套开放性的新闻事件提取框架。构建中文新闻事件数据库的过程可以概括为将原始的新闻文本进行处理，然后进行句法分析和语义角色识别，从中提取三元组，再提取动词并转换为向量表示，之后通过降维和聚类结合人工标注形成结构化数据，最后提出了事件重要性得分以评估新闻中事件的分布情况。利用《人民日报》的新闻数据进行了实验，验证了本文研究的理论与实践价值。

关键词：新闻事件；事件数据；无监督学习
中图分类号：TP391　**文献标识码**：A

Construction and analysis of Chinese news event data based on unsupervised techniques

YUAN Fang, LU Wei, SHEN Hao*
(State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)

Abstract: In this paper the task of constructing Chinese news event data for media and communication research was explored, technologies such as natural language processing, deep learning, and unsupervised clustering were utilized to construct an open-ended news event extraction framework. The process of constructing the Chinese news event database could be summarized as processing the original news text, performing syntactic analysis and semantic role recognition, extracting triplets from it, then extracting verbs and converting them into vector representations, followed by dimension reduction and clustering combined with manual annotation to form structured data. Finally, an event importance score was proposed to assess the distribution of events in the news. The framework was tested using news data from the People's Daily, validating the practical value of the research.

Keywords: news event;event data; unsupervised learning

基金项目：中国传媒大学中央高校基本科研业务费专项资金资助（CUC23GY004）
作者简介（*为通讯作者）：元方（1987-），女，博士研究生，主要从事面向舆情分析的自然语言处理和信息抽取研究。Email：yuanfang@cuc.edu.cn；卢伟（1999-），男，硕士研究生，主要研究方向为数据挖掘、数据可视化和系统工程。Email：luwei@cuc.edu.cn；沈浩（1963-），男，博士，教授，主要从事媒体融合、传播效果、大数据与人工智能研究。Email：shenhao@cuc.edu.cn

1 引言

作为社会信息传播的主要渠道,新闻在互联网信息中占有重要地位,其由全球媒体组织采集、整理、发布,构成了公众了解世界信息的主要途径。随着互联网技术的演进,媒体的覆盖范围和内容日益丰富,使得全面阅读新闻变得不切实际。因此,信息的精简合并,以提高阅读效率,已成为媒体发展的重要趋势。对于传播和社会研究者,丰富的历史新闻信息构成了宝贵的研究资源。媒体传播的观点、形象和信息是文化的核心部分,塑造了人类对社会现实和社会规范的理解,服务于公共社会生活。新闻,作为媒体内容的重要组成部分,既是现实的反映,也是历史的记录。然而,新闻的来源分散,缺乏统一的组织规范,存在大量冗余,且其非结构化特性使得资源利用和分析困难。传统社会科学研究方法如内容分析法,通过抽样和人工编码进行分类和统计分析,其覆盖范围有限,不适合处理广泛性问题。因此,利用计算机技术进行大规模新闻信息分析已成为研究者的新方向。对新闻文本的研究是传播学内容研究的一部分,而传统的内容分析受研究者主观态度影响,可能导致不同研究者得出不同结论。利用机器技术处理新闻文本,将非结构化的新闻文本结构化,以建立新闻事件库,可以实现更好的客观性和一致性,扩大传播学的研究视野,扩展新的传播学研究手段。本研究立足于此,设计了一套结合词向量表示、降维和聚类在内的基于无监督技术的中文新闻事件数据构建框架,提出了事件重要性得分评估新闻中事件的分布情况,并在此基础上以《人民日报》的新闻数据进行了分析验证,具有较强的研究意义与实践价值。

2 主流新闻事件库

目前主流的新闻事件数据库从构建方法上看,可以分为人工编码事件数据库和自动编码事件数据库。从应用上来看,自动编码数据库因为数量更大、覆盖面更广而得到研究者们的青睐[1-3]。自动编码事件数据库的工作流程都比较类似,即在人工收集词典的基础上检测事件,这些词典与行动者及事件领域本体相关联。文本中的每个句子被视为一个或多个事件,其中的谓语动词表示触发事件的动作,左右的名词表示行动者的实体,分别视为源行动者和目标行动者。核心算法的主要技术包含自然语言处理和机器学习,例如对文本进行预处理、浅层解析以及对非英语文本的机器翻译等。

全球事件、位置和音调数据库项目(Global Data on Events, Location, and Tone,GDELT)被称为“有史以来最大、最全面、最高分辨率的人类社会开放数据库”[4]。GDELT项目包含四个数据库:全球事件数据库、全球知识图谱、可视化全球知识图谱和数字化书籍中抽取的知识图谱特殊集合。GDELT数据库使用冲突和调解事件观察(Conflict and Mediation Event Observations,CAMEO)编码框架提供编码事件数据,事件的范围主要是政治、军事、外交、灾害等。编码内容主要是事件类型、事件行动者、日期、地点、主题和情绪等数十个属性变量。此外从2016年起,GDELT还在可视化全球知识图谱数据库中提供了通过Google的Vision API进行处理的图片信息,包括图片的地理位置、徽标、文本的识别及推断的图片情感等。

综合冲突预警系统(Integrated Conflict Early Warning System,ICEWS)提供类似于GDELT的编码事件数据库[5]。数据记录表示源和目标参与者的特征,包括名称、类型和国家等。事件动作的特征包括日期、来源、简短的文本描述和位置描述、强度得分等。编码体系同样基于CAMEO框架。

SPEED项目基于《纽约时报》的新闻档案、FBI外国广播信息服务和BBC的世界广播摘要等新闻提供商抽取事件[6],在文本分类阶段,使用基于朴素贝叶斯的分类器对数据集进行分类,选出与政治相关的文章,然后通用NLP流水线进行处理,例如标记化、句子分割、词性标注、实体抽取、分块、依从句法分析、共指消解和情感检测等。

EventRegistry项目没有一个预设的编码体系,而是通过对新闻文章的聚类实现[7]。它的数据库包含多语言标题和摘要文本、有关该事件的文章数量、事件日期以及一组与事件相关的概念关键词,属于一种无监督的方法。

对基于CAMEO的事件数据,其基本逻辑是:给定一个句子,编码器在CAMEO动词模式词典中搜索匹配模式。一个模式由一个动词和周围的关键词组成。该模式表示一个特定的行动过程,由事件代码表示。例如,模式"SETOUTVIEWS"中的SET表示一个公开声明类型的事件。找到模式中的匹配后,行为者字典会搜索代表源和目标的匹配实体。在找到必要的信息后,一个事件被PETRARCH编码。如果缺少信息,PETRARCH将忽略该事件。这个事件序列被称为源-行动-目标或SAT格式。基于CAMEO政治事件编码的事件数据构建流程如图1所示。

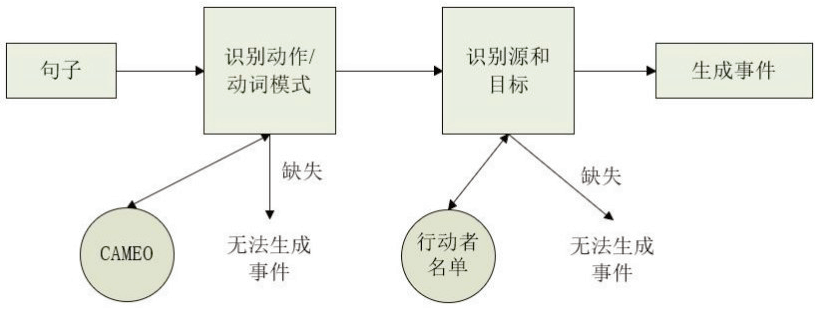

图1 基于 CAMEO 政治事件编码的事件数据

另一大类提取事件的方法则主要基于Universal Dependency（UD），这是一种支持跨语言一致的树库注释的技术。它的总体目标是提供一个通用的类别和准则集合，以促进跨语言的类似结构的一致注释，同时在必要时允许特定语言的扩展。例如对于新闻句子“春节前后，民政部陆续公布了三批250多个涉嫌非法社会组织名单，引起社会广泛关注。”可以生成如图2所示的依赖树：

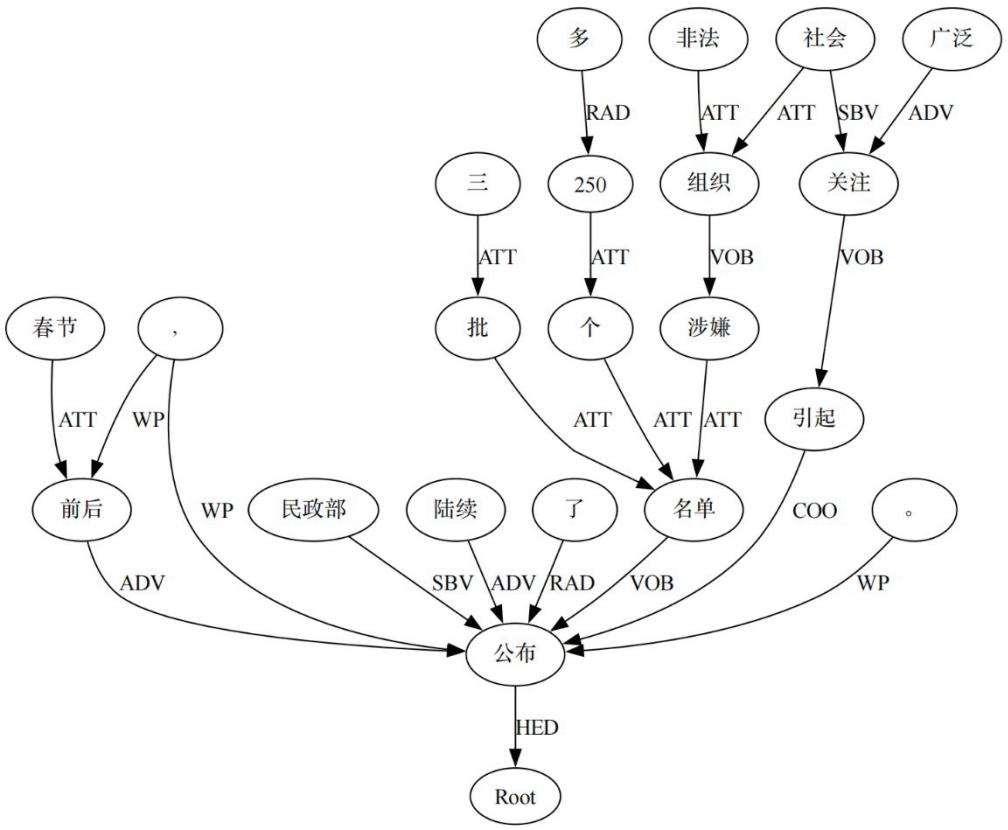

图2 句子依赖树

中文事件数据库的研究相对较少，事件抽取相关研究中广泛使用的ACE2005数据集中包含一定数量的中文数据，Chen等人在此基础上基于语言学特征进行了中文触发词和论元角色分类[8]。而ACE语料之外的中文事件抽取以金融领域为主。DCFEE框架尝试了用远程监督结合已有的专家知识库生成中文的

标注语料,标注的内容除了句子级别的触发词和事件论元,还有文档级别的事件提及[9]。Doc2EDAG框架在文档级的中文金融事件抽取上更进一步,通过基于实体的有向无环图完成抽取任务[10]。研究者还提供了一个大规模真实世界金融事件数据集。在金融之外的其他领域,上海大学语义智能实验室建立了中文突发事件语料库[11],从互联网上收集了地震、火灾、交通事故、恐怖袭击和食物中毒五类突发事件的新闻报道作为语料,经过文本预处理、分析后进行标注和一致性检查,最后保存到语料库中,共计332篇文档。张秀华等利用基于注意力机制的双向长短记忆网络构建新闻事件检测模型[12]。总体而言,现有研究多基于单个领域或有监督技术建立分类模型,缺乏无监督的泛化领域的中文新闻事件研究。

3 中文新闻事件库构建流程

如图3所示,构建中文新闻事件数据的过程可以概括为将原始的新闻文本进行分段、分句、分词处理,然后进行句法分析和语义角色识别,从中提取三元组,再提取动词并转换为向量表示,使用神经聚类框架进行降维和聚类,之后结合人工标注和类别消歧技术形成结构化数据。

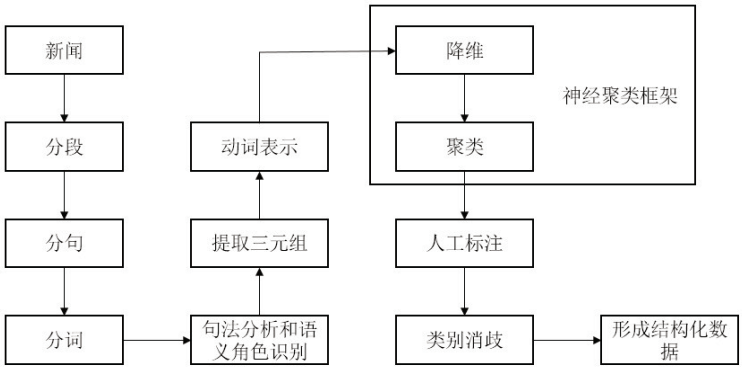

图3 事件数据构建流程

3.1 新闻数据准备

综合考虑时间因素和数据获取的便利性,本文选择了2018年全年的《人民日报》新闻报道作为研究对象。《人民日报》官方网站图文数据库会每日更新当日报纸的电子版数据,并提供一段时间内的历史数据。通过Python语言脚本程序,根据每日报纸网址,获取每日的新闻报道文章列表网址,再根据该列表网址获取每一篇文章的网页HTML源代码,通过正则表达式、标签匹配等方式,去掉无关内容,得到新闻标题、所属版面、新闻正文内容等信息,以文本文件的形式进行存储备用。

由于《人民日报》的内容除了一般意义上的新闻之外,还包括评论、生活类文章及一些热点专题的深入报道。因此为了保证提取工作的顺利进行,首先对新闻进行基础筛选,基础筛选主要根据抓取新闻中的"版名"字段进行。2018年《人民日报》共涉及72个版名,其中既包括"共商友好合作大计共绘发展美好蓝图·2018年中非合作论坛北京峰会特别报道"这样的专题版名,也包括"文件"、"理论"这样的解读性内容,还有"副刊"、"读书"一类的文艺版面,以及"广告"、"视觉"等文字内容较少的版面。经过研究和咨询专家,考虑不同文体的适用性,过滤筛选去掉了部分版面的文章,最后保留了包括"要闻"、"社会"、"政治"、"国际"、"综合"、"体育"等在内的26个版面,共计27988篇文章,以此作为本文构建中文新闻事件数据库的基础语料。

3.2 事件三元组提取

事件三元组是一种原子级别事件表示的形式,简单来看就是由一个动作及对应的施事、受事组成。但为了保持语义的完整性,施事和受事并不仅仅是句子的主语词和宾语词,还要对其进行扩展。这些生成的事件三元组会作为构建事件类型编码系统的原材料,并作为备选事件进入下一步处理流程。

三元组的提取主要通过依存句法分析和语义角色标注完成。依存句法分析(Dependency Parsing)的目标是分析句子中各成分之间的依赖关系,这种依赖关系就体现了句子的语法结构。一个句子中所有单词的依赖关系构成一棵句法树,根节点是中心谓词。

语义角色标注可以完成对句子的浅层语义分析，找出其中的谓词-论元结构。本文对清理完毕的新闻进行分段、分句、分词、词性标注和语义角色标注，对于语义角色标注出的结果，将施事、谓语、受事作为基本元素构建三元组，对于未能标注的句子，通过依存句法分析找出主语、谓语和宾语，然后在此基础上进行扩展。

例如“文化部主办了今年的文化交流展会”一句话的提取结果为“ORG_主办_展会”，其中“ORG”代表施事的主语，即本句话中的机构名“文化部”，“展会”是受事宾语，“主办”是谓语动词，连接施事主语和受事宾语；又如“文化部组织的文化交流展会树立了中国的形象”一句话提取结果为“SVO_树立_形象”，其中“树立”是谓语动词，“形象”是受事宾语，“文化部组织的文化交流展会”以表示主体的从句的形式作为施事主语，同时，该从句本身也可以概括为“ORG_组织_展会”的结构。

3.3 事件类型编码

本文使用HDBSCAN（Hierarchical Density-Based Spatial Clustering of Applications with Noise）聚类，是一种基于密度的层次化空间聚类，它将向量空间中的点到k个最近邻居的距离定义为核心距离，较小的核心距离的点被视为密集点，然后在可达性基础上构建最小生成树，最后压缩树形成类别。与传统聚类相比，HDBSCAN聚类的优势是不容易受到噪声点的影响，并且可以适应密度不同的类。

根据前述的方法，生成的备选事件三元组共计595028个，如此细粒度的原子级别事件是无法构成具有可用价值的数据库的。因此我们还需要对这些三元组进行进一步的分析建模，以形成事件类别编码。事件类别编码基本思路是以动词作为三元组的核心，使用word2vec预训练词向量作为动词表示，使用UMAP (Uniform Manifold Approximation and Projection)进行降维后通过HDBSCAN层次密度聚类算法聚类，并联合UMAP降维和HDBSCAN聚类进行网格化参数搜索。UMAP使用余弦距离，HDBSCAN使用欧几里得距离，评价指标设置为轮廓系数（Silhouette Score, SS），每个词的轮廓系数SS计算公式如式（1）：

$$SS = \frac{b - a}{\max(a,b)} \tag{1}$$

其中a代表该词与同一类别中所有其它词之间的平均距离，b表示该词与下一个最近的类别中所有词之间的平均距离，聚类方案的轮廓系数表示为所有动词轮廓系数的平均值。最终UMAP的邻居数量设置为3，维数设置为8，《人民日报》2018年新闻报道中提取的备选事件三元组共计280个事件类别，最终轮廓系数为0.4953。

3.4 类别层次化

考虑新闻报道的主题题材具有集中性，往往一篇新闻报道文章涉及的社会主题不会太多，会相对独立地集中于报道经济、时政、灾害、社会事件等，不会出现混杂的情况，每一类主题报道所使用的词语会具有相对集中的趋势。为了进一步降低事件类别的维数，从更宏观的角度来概括具体的新闻报道文章，本研究提出“事件共现”的概念，即同时出现在某一主题新闻报道中的事件的集合。

定义事件共现距离作为类别合并的标准，两个类别的共现率$Event_{corr}$计算公式如式（2）：

$$Event_{corr} = 1 - \frac{\sum corr_{v_i v_j}}{count_1 * count_2} \tag{2}$$

其中$corr_{v_i v_j}$ =

$$\begin{cases} 1, \text{ 如果} v_i \text{和} v_j \text{有在同一篇文章中共同出现} \\ 0, \text{ 没有共同出现} \end{cases},$$

$count_1$和$count_2$表示两个类别中的词数。

首先计算280个事件类型在《人民日报》2018年新闻报道中的共现矩阵（矩阵为280*280维），然后再次聚类，即得到一级分类。

3.5 事件重要性得分

借鉴关键词提取中的YAKE算法[13]，构建事件动作重要性得分KAS（Key Action Score），对前一章得到的三元组中谓语动词进行排序，从而得到一篇文章中的主要事件动作。

KAS考虑五个特征，具体计算方法如下：

（1）动作权重$Action_{tf\text{idf}}$，使用动词$tfidf$值作为动作的权重：

$$Action_{tf\text{idf}} = \frac{Verb_{freq}}{\ln\left(\frac{Doc_{num}}{Verb_{df}} + 1\right)} \tag{3}$$

其中$Verb_{freq}$代表对应动词的词频，Doc_{num}表示文档数量，$Verb_{df}$表示动词的文档频数；

（2）动作依赖关系$Action_{dep}$，如果三元组的谓语动词是句子中的根节点或者根节点的同位语，更可能是关键事件动作：

$$Action_{dep} = \frac{Verb_{head\_count}}{\log(1 + Verb_{freq})} \quad (4)$$

其中 $Verb_{head\_count}$ 表示对应动词作为句子根节点或根节点同位语的次数,除以词频进行标准化;

(3)实体搭配 $Action_{ner}$,如果三元组的主语和宾语中包含命名实体,更可能是关键事件动作:

$$Action_{ner} = \frac{Subject_{is\_ner} + Object_{is\_ner}}{\log(1 + Verb_{freq})} \quad (5)$$

对于对应动词的所有三元组搭配,$Subject_{is\_ner}$ 为主语中有命名实体的频次,$Object_{is\_ner}$ 为宾语中有命名实体的频次,除以词频进行标准化;

(4)三元组位置 $Action_{position}$,考虑到新闻的写作特点,事件动作越靠近文章开头重要程度越高;

$$Action_{position} = \ln(\min(Verb_{position}) + 1) \quad (6)$$

其中 $Verb_{position}$ 表示包含该动词的文章中句子的索引;

(5)语义角色参数数量 $Action_{srl}$,谓语动词识别到的语义角色参数数量越多,说明动作含义越完整,就越可能是关键事件动作;

$$Action_{srl} = \frac{\ln(Verb_{srl\_num} + 1)}{\text{mean}(Verb_{srl\_sum}) + \text{std}(Verb_{srl\_num})} \quad (7)$$

其中 $Verb_{srl\_num}$ 表示对应动词语义角色识别出的参数数量总和,$\text{mean}(Verb_{srl\_sum})$ 和 $\text{std}(Verb_{srl\_num})$ 分别表示文章中所有动词的语义角色参数总量的均值和标准差。

得到这五个特征后,按照公式(8)计算每个候选动词的分数:

$$\text{KAS} = \frac{Action_{position}}{(Action_{ner} + Action_{srl}) \times (1 + Action_{dep}) \times Action_{tf\text{idf}}} \quad (8)$$

计算出的KAS得分越小,说明该事件动作在新闻中越重要。对得分进行升序排序,从而得到事件重要性排名。

4 结果分析

4.1 事件分类列表

通过前述技术处理,从2018年《人民日报》新闻报道文本中得到的事件二级类别共280类,对这280类进行人工识别与编码,标注对应的事件类别。表1列出前10类及对应的前10个关键词作为示例:

表1 事件二级分类中前10类的关键词示例

类别编号	类别名称	关键词
001	表示务工返乡相关事件	创业、打工、务工、返乡、发家、经商、回乡、起家、谋生、探亲
002	表示排练演出相关事件	编排、排练、吹奏、编配、登台、演出、执棒、演奏、编创、巡演
003	表示租赁关系相关事件	租用、购置、闲置、轮候、承租、出租、租赁、合租、转租、招租
004	表示勘察勘测相关事件	勘查、勘察、勘定、航测、试采、勘测、测绘、勘探、踏勘、查勘
005	表示交通相关事件	通行、过境、驶入、途经、分流、开进、放行、驶出、驶离、停靠
006	表示对外资助相关事件	资助、援建、筹措、募集、捐赠、筹款、捐助、捐款、募捐、援助
007	表示学习教育相关事件	学习、攻读、深造、自学、选修、涵育、教育、进修、选学、修读
008	表示出游访问相关事件	走访、参观、住宿、外出、相聚、团聚、相约、游览、出行、聚会
009	表示服役参军相关事件	服役、备战、参军、转业、跋涉、深潜、兼程、退伍、集训、当兵
010	表示投资收入相关事件	投资、升值、盈利、亏损、分红、收益、价值、复盘、变现、对赌

从上表可见,本文构建的事件分类涵盖了常见新闻报道中的时事、政策、经济、体育、农业、军事、历史、税务、勘探、社会、民生、交通、建筑、犯罪、文化、艺术、医疗等众多方面内容,从宏观的报道内容概览到微观的重点细节叙述和描写,可以构成一份完整的结构化新闻事件数据。

通过专家和标注员分别人工核对原始新闻与事件关键词的符合度、意义表达的准确度和可信度,本文的事件数据提取方法可靠,所提取的事件主题关键词可以用来判断新闻报道文章的事件类型分类。

4.2 层次化一级分类

在前述事件的二级分类的基础上,本研究根据关键词在不同文章中同时出现的概率,借鉴距离计算的思想,使用统计算法进行再次聚类,结合新闻学题材分类的特点,辅以主观判断,提取出事件的一级分类,表2所示是编号为01的一级分类及所辖二级分类示例。

表2 事件二级分类聚类成一级分类示例

一级类别编号	一级类别名称	二级类别编号	二级类别名称	关键词
01	与案件相关的社会事件	043	表示警方办案相关事件	取缔、整治、收缴、捣毁、缴获、破获、抓获、逮捕、盗窃、侦查
		162	表示监督督办相关的事件	监督、问责、巡察、明察暗访、督导、督办、督察、暗访、督查、纠治
		264	表示审判相关的事件	判决、起诉、宣判、审判、索赔、调解、投诉、改判、指控、审理

编号01的一级分类是与案件相关的社会事件，其下的二级分类包括编码043的表示涉案或警方侦查过程相关的事件、编码162的表示与案件的监督督查相关的事件和编码264的表示与案件的审判判决相关的事件。此示例中，一级分类可以认为是对各个组成它的二级分类的概括汇总，二级分类是一级分类事件进程的各个组成部分。此类新闻的报道一般以时间和事件的发展过程为顺序，本文的分类结果是基本符合这一顺序的。编码043的分类主要用来概括事件的基本事实情况。编码264的分类主要用来概括警方侦查侦办结束后人民法院对案件进行审判判决的事件。在新闻报道中，会有一定的篇幅对案件判决结论进行叙述，一方面对事件本身给出最后的定论，一方面是起到普法和警示的作用。编码162的分类主要用来概括与案件相关的以人民检察院为主体的监督督查相关事件，多出现在腐败类案件、重大审判案件等事件中。《人民日报》报道的涉案类社会新闻事件多为重大、有影响力的事件，实际中，此类事件也多会有监察机关介入，对此类内容的概括也是整个事件概括的组成部分之一。以上几个二级分类的事件可以单独出现，也可以组合出现，以不同的事件发展程度的组成部分构成对一级分类事件的概括。

综合以上示例可以看出，提取出的一级分类分为两个不同的类别：一个类别是指一级分类是对各个组成它的二级分类的概括汇总，二级分类是一级分类事件进程的各个组成部分；另一个类别是指二级分类是平行并列存在的，共同属于同一个一级分类的题材。通过对相关新闻报道文章原文的回顾，梳理相关文章的报道逻辑、写作特点，结合新闻学对报道题材的划分，可以认为以上的分类具备科学性、准确性、实用性。

4.3 事件重要性与关联网络构建

排名后需要对文章中的事件进行筛选，为了减少人工干预，采用多项式函数拟合得分曲线，然后计算拐点作为保留标准。例如，2018年1月22日“要闻”版的《美破坏南海稳定是不识时务的妄动》新闻中，首先按照前一节所述方法提取出的三元组共计27个，然后计算重要性得分，结果如表3所示（为简略起见仅列出动词）。

表3 示例新闻的事件三元组KAS得分表

编号	动词	KAS得分	编号	动词	KAS得分
1	进入	0.4033	15	推动	1.6872
2	干扰	0.4553	16	损害	1.9339
3	达成	0.5492	17	解决	1.9719
4	进行	0.5756	18	探讨	1.9889
5	受到	0.6751	19	保持	2.2593
6	建立	0.7486	20	违背	2.4586
7	表示	0.9584	21	启动	2.6519
8	维护	1.0274	22	动摇	2.7238
9	开展	1.0930	23	利用	2.7970
10	拥有	1.1196	24	炫耀	2.9183
11	举行	1.2729	25	期盼	4.2574
12	造成	1.3854	26	得出	6.0359
13	予以	1.4475	27	臆想	7.3060
14	宣布	1.4726			

事件三元组KAS得分拟合曲线如图4所示。可以看到，拐点出现在n=21处，因此根据重要性保留前21个三元组。

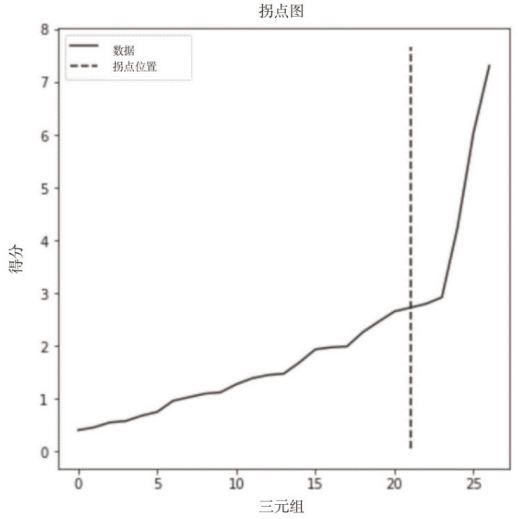

图4 示例新闻的事件三元组KAS得分拟合曲线

如果说事件三元组提取解决了新闻事件“Who”

和“What”的问题,那么事件参数提取就是要解决“When”和“Where”的问题。能够定位到新闻中事件发生的时间和地点,对新闻事件数据库的分析有着非常大的意义。但是,中文本身是一种分析语言,亦即并不是通过词语本身的形态变化,而是通过词序和虚词表示语法意义。同时,汉语有相对丰富的时态(Aspect)标记,而缺乏显性的时制(Tense)标记[14],这也给中文新闻中的时间提取造成了困难。

对于这两类数据的提取,主要采用模式匹配和预训练模型相结合的方法。其中对时间的提取以模式匹配为主,共设计300余条规则模式,示例如:({Lunar}(\s*))?(((({SimpleYear}|{DateYearInChinese})年)(\s*))?{Month}(\s*){DateDayInChinese}((\s*|,|,){WeekDay})?表示匹配(农历)?(XXXX年)?X月X日(星期X)?的格式。提取出的日期标签主要分为“单一日期”和“日期范围”两类,前者可直接附加进数据库,后者则来自于类似“2017年全年”这样的表述,对于日期范围的处理,统一设置为范围结束的日期,也就是说,“2017年全年”会被设置为“2018-01-01”。由于新闻中一般对时间书写较为规范,规则匹配准确率较高,经人工抽样审查,准确率超过70%,尚可接受。对于地点则以命名实体识别为基础,在识别出的地名实体基础上进行排序。分别将距离三元组核心动词最近的时间和地点匹配给三元组。

前面虽然已经完成了事件三元组的筛选和时间地点的抽取,但对每篇新闻而言,事件依然是孤立的细粒度三元组,为了更好地对单篇新闻中的事件分布进行呈现,本文使用网络来构成事件结构。对于每个三元组,取主语和宾语中的核心词汇,与动词抽象成为的类别建立关系,分别为主语核心词→动词类别和动词类别→宾语核心词,然后再为主语和宾语分别构建修饰关系,也就是修饰词→核心词的连接,最后令已被赋予时间和地点标签的标签分别指向核心动词类别。上一节所述《美破坏南海稳定是不识时务的妄动》一文形成的事件关联网络及其2-核心如图5和图6所示,其中颜色表示使用社区发现算法计算出的类别:

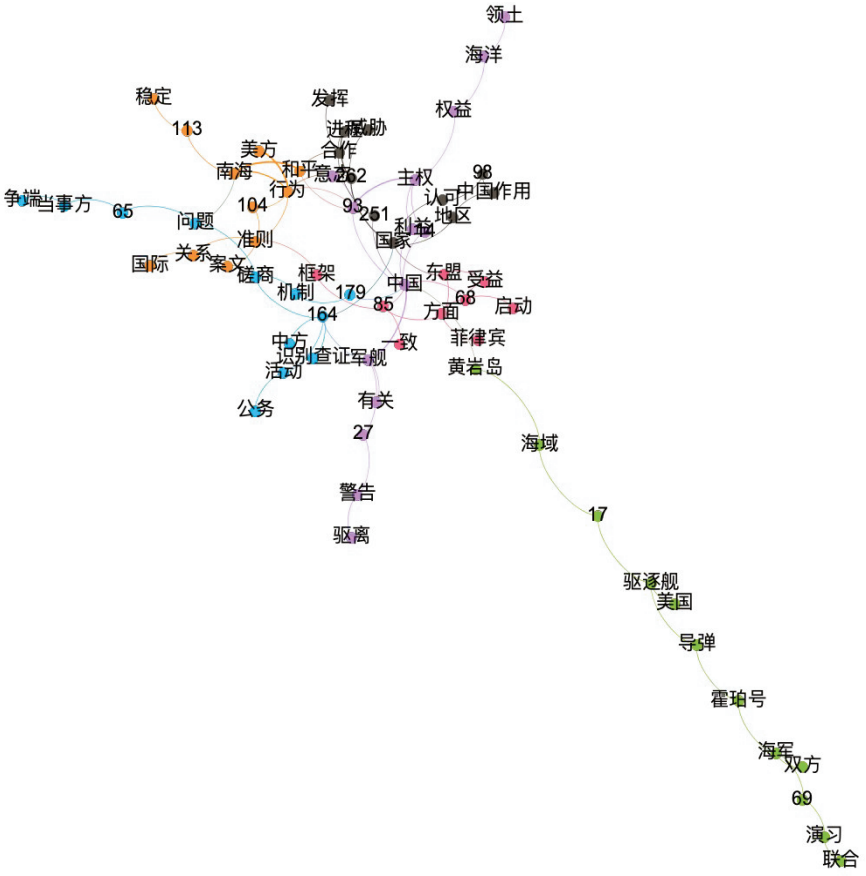

图5 示例新闻的事件关联网络

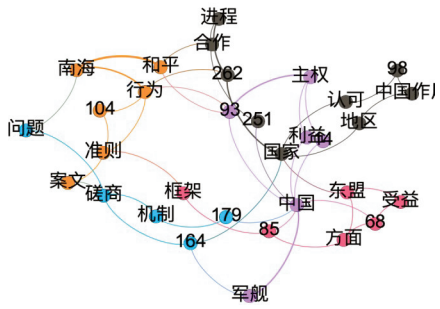

图6 示例新闻事件关联网络的2-核心

可以看到，这篇新闻中的事件被分成了8个互相交织的主题，分别描述了美方挑衅、中方应对、宣布主权、呼吁和平以及作为背景事件的南海行为准则及地区国家关系等。

5 总结与讨论

本文针对无监督的中文新闻事件数据库构建任务，首先通过提取三元组构成原子级别的备选事件，然后将动词作为事件的根本，通过词向量生成动词的向量表示，通过智能技术提取出新闻事件的二级分类，而后结合关键词共现率算法总结归纳一级分类，实现了事件数据构建基本流程。在新闻业务实践中，词语是新闻事件报道的主要构成元素，本文通过对词语在不同事件上的倾向特征计算，将词语元素进行拆分和聚类，实现了事件主题关键词的概括和提取。词语作为新闻事件的符号式表现形式，是新闻事件意义的载体，关键词语组合成的事件分类便构成了事件的意义建构。本文形成的新闻事件数据库符合新闻学的理论和业务实践要求，能够真实、公正、全面、客观地再现新闻事实主旨，是一种科学性的处理方法，得到的结果具备可靠性、准确性、实用性，是对自然语言处理技术的应用和发展。通过计算机智能程序手段和技术在新闻传播学领域的探索，获得了用程序化方法解读和理解新闻语言文本的途径和方法，具备科学实用的理论研究和社会实践价值。

在备选事件三元组基础上通过KAS得分评估事件三元组重要性，然后通过模式匹配抽取事件发生时间，通过命名实体识别抽取事件发生地点，将词语和动作类别连接起来构成网络作为新闻文档中的事件表示分布。在未来的研究中，该网络表示可以通过社区侦测等相关算法探测更高一级抽象的主题，也可以通过核心组件发现的方法再次进行事件筛选，还可以通过随机游走（Deep Walk）的方式生成事件向量，输入到其它相关问题的分析之中。

参考文献（References）：

[1] Hopp F R, Schaffer J, Fisher J T, et al. iCoRe: the GDELT interface for the advancement of communication research [J]. Computational Communication Research, 2019(1): 13-44.

[2] Barroso del Toro A, Tort-Martorell X, Canela M A. How shareholders react to sustainable narratives about leading European energy companies? an event study using sentiment data from the global database for events, language and tone (GDELT)[J]. Applied Economics, 2022, 54(30): 3482-3494.

[3] Bernard G, Suire C, Faucher C, et al. Tracking news stories in short messages in the era of infodemic[C]//13th International Conference of the Cross-Language Evaluation Forum for European Languages, 2022: 18-32.

[4] Leetaru K, Schrodt P A. Gdelt: global data on events, location, and tone, 1979-2012[C]//ISA Annual Convention, 2013(2): 1-49.

[5] Schrodt P A, Analytics P. Comparing methods for generating large scale political event data sets[C]//Text as Data Meetings, 2015.

[6] Nardulli P F, Althaus S L, Hayes M. A progressive supervised-learning approach to generating rich civil strife data[J]. Sociological Methodology, 2015, 45(1): 148-83.

[7] Leban G, Fortuna B, Brank J, et al. Event registry: learning about world events from news[C]//Proceedings of the 23rd International Conference on World Wide Web, 2014: 107-110.

[8] Chen C, Ng V. Joint modeling for Chinese event extraction with rich linguistic features[C]//Proceedings of COLING, 2012: 529-544.

[9] Yang H, Chen Y, Liu K, et al. Dcfee: a document-level Chinese financial event extraction system based on automatically labeled training data[C]//Proceedings of ACL 2018, System Demonstrations, 2018: 50-55.

[10] Zheng S, Cao W, Xu W, et al. Revisiting the evaluation of end-to-end event extraction[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 4609-4617.

[11] 刘炜，王旭，张雨嘉，等.一种面向突发事件的文本语料自动标注方法[J]. 中文信息学报，2017，31(02)：76-85.

[12] 张秀华，云红艳，贺英，等.基于注意力机制的新闻事件检测研究与应用[J].计算机与数字工程，2021，49(06)：1143-1147+1280.

[13] Campos R, Mangaravite V, Pasquali A, et al. YAKE! keyword extraction from single documents using multiple local features[J]. Information Sciences, 2020(509): 257-289.

[14] 顾阳.时态、时制理论与汉语时间参照[J]. 语言科学，2007(04)：22-38.

编辑：赵志军