事件库构建技术综述

薛 聪[1,2,3], 高 能[1,2,3], 查达仁[1,2,3], 王 雷[1,2,3], 尹芷仪[1,2], 曾泽华[1,2,3]

[1] 中国科学院信息工程研究所信息安全国家重点实验室 北京 中国 100093
[2] 中国科学院数据与通信保护研究教育中心 北京 中国 100093
[3] 中国科学院大学网络空间安全学院 北京 中国 100049

摘要 恐怖事件、突发事件、冲突事件等特定主题事件通常对国家安全带来严重威胁, 记录现实事件的事件库在态势感知、风险预警、应急决策等应用中发挥重要作用, 事件库构建技术随之发展为内容安全技术的重要组成部分。事件库构建技术是一类实现从海量的非结构数据批量生成结构化事件数据的技术, 由于数据环境、表示精度、应用场景的差异, 出现了各类构建技术的相关研究。本文详细介绍了事件库的定义、分类和架构, 按自底向上输出的数据层次, 将事件库构建技术划分为事件检测、事件抽取、事件融合三类关键技术, 并分别对其研究现状和进展进行了全面分析, 总结了事件库的主要应用领域, 最后对事件库构建技术中面临的主要挑战和关键问题进行了探讨。

关键词 事件库; 事件检测; 事件抽取; 事件融合
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2019.03.08

Event Database Construction Techniques

XUE Cong[1,2,3], GAO Neng[1,2,3], ZHA Daren[1,2,3], WANG Lei[1,2,3], YIN Zhiyi[1,2], ZENG Zehua[1,2,3]

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] Data Assurance and Communication Security Research Center, Chinese Academy of Sciences, Beijing 100093, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Specific topic events, such as terrorist attacks, emergencies and conflicts, pose serious threats to national security. Since event database which records real-life activities plays a major role in situational awareness, crisis early warning and contingency decision, its construction techniques have become an important part of content security. The construction techniques of event database achieve mass production of structured event data from massive unstructured data. Because of the differences in data landscape, data precision and application scenario, there have been relevant researches of various construction techniques. This paper describes the definition, classification and architecture of event database in detail. According to the bottom-up data level, the paper divides the construction techniques of event database into three classes of critical technologies, including event detection, event extraction and event fusion, and makes comprehensive analysis on current research status and development of these technologies. Furthermore, the paper summarizes the main application fields of event database. Finally, this work conducts in-depth discussions on major challenges and key issues which the construction techniques of event database confront.

Key words event database; event detection; event extraction; event fusion

1 引言

现代社会, 外交争端、武装冲突、暴恐事件、突发公共安全事件等国内外特定主题事件频繁发生, 对国家和社会安全带来巨大的冲击。与此同时, 国际和社会形势不断变化, 各类事件相互作用, 导致事件发生的诱因日趋复杂, 越来越多的学者发现对特定主题下大量事件的持续监控, 可以发现该类事件的发展规律, 例如美国国防部研究通过对极端组织“伊斯兰国” (ISIS)制造的历史恐怖事件分析发现其活动规律[1]。近年来, 新闻和社交网络等开放信息平台高速发展, 为人们提供了洞察事件的第一手资料, 也使得事件影响随网络迅速传播, 甚至影响事态发展, 例如 2010 年底发生在突尼斯的自焚事件通过社交网络扩散传播, 迅速引起整个阿拉伯地区相继爆发革命事件。然而, 针对大量混杂数据中的事件信息,

通讯作者: 查达仁, 博士, 高级工程师, zhadaren@iie.ac.cn。
本课题得到国家自然科学基金(No. U163620068), 战略合作专项(No. AQ-1704, AQ-1708)项目资助。
收稿日期: 2017-07-11; 修改日期: 2017-10-30; 定稿日期: 2019-02-27

越来越多的学者希望实现这类特殊事件的自动化发现, 获得大量精度高、机器可阅读的事件数据, 并构建出各类结构化事件库。2014-2016 年美国情报先进研究计划局(IARPA)连同多家高校完成了 EMBERS 事件系统[2], 基于新闻、Twitter、Facebook、地图探测等十余种混杂的数据实现了反叛、冲突等 7 大类事件的在线监控, 形成包括发生地点、事件类别、涉事人群等内容在内的精细事件库, 引起俄罗斯、德国等多国政府关注; 在该事件系统基础上开展了涉事团体活动规律挖掘、事件起因发现等多项研究。通过构建事件库, 可形成态势感知、危机预警、风险控制、应急决策等多类安全应用的分析基础。因此, 事件库的构建技术已成为信息内容安全的重要部分。

不同于稳定存在的知识数据, 事件数据记录了瞬时或短期的人类活动及相应影响, 是对现实社会变化的原子性描述。大量事件数据被集成到事件库, 实现了现实世界和人类活动的全面映射。事件库构建技术是一类实现从粗糙的非结构数据到批量生成结构化的事件数据的技术, 相关学术研究和实际应用经过了从零散信息的人工整合到大规模数据的自动萃取集成的发展过程。

第一阶段, 主要由政府主导通过人工搜集事件情报信息: 从 20 世纪 60 年代起, 美国学者就以人工方式构建了国际冲突事件库, 并应用演化预测模型指导军力部署[3]。随后多国研究机构相继构造了政治互动事件库 WEIS、战争行为事件库 BCOW 等, 积累了丰富的结构化社会事件的知识与规范。

第二阶段, 开始推进基于新闻等规范文本的事件库自动构建技术和特定要素的相关编码标准的规范: 1987 年, 美国国防部高级研究计划署(DARPA)资助 MUC 会议着重发展针对军事行动、恐怖主义活动等主题的文本信息提取技术[4], 发展了自动化事件编码系统。随后, 美国堪萨斯大学在 1994 年发布了第一个自动编码的 KEDS 事件库[5]。1998 年后 MUC 会议延伸发展为美国 NIST 组织主办的 ACE (Automatic Content Extraction)测评会议[6], 并从 2005 年起将事件检测和表征作为基本任务之一, 2008 年扩展到实体识别标注的规范, 并发布了英语、汉语、阿拉伯语 3 种语言的用于事件抽取的标注数据, 该数据库至今仍作为事件抽取领域的标准库之一。

第三阶段, 扩充到围绕社交媒体等开放数据中的实时事件检测以及知识类关系数据的获取方法: 2009 年起, NIST 主办的文本分析会议(TAC)、文本检索权威测评会议(TREC)等陆续推动群体消息中的事件发现以及事件时序摘要、精细化事件信息抽取等方向, 并扩展到结构化知识抽取与更新等方向。此外, 美国政府合作机构推进的 SocialRadar[7]、瑞士著名情报公司的 OpenMind[8]等项目陆续围绕社交网络环境中的事件感知开展, 相关研究至今仍是学术研究热点。

与此同时, 大规模自动构建的结构化事件库逐渐成熟, 发展成为全面记录人类现实活动的精细化数据资源, 并推进相关事件分析研究: 2010 年美国国防部公布超过 2.5 亿条记录的 ICEWS 全球冲突事件库[28], 并在此基础上推进早期预警服务; 2011 年, 乔治城大学连同多所科研机构发布 GDELT (Global Database of Events Languages, and Tones)事件库[29], 从超过 100 种语言的新闻、焦点社区和社交网络数据中发现并记录了从 1979 年以来发生的人类主要活动, 掀起智能计算方法用于社会研究的热潮, 仅在 GDELT 平台上的研究项目就有近千个。2013 年 EventRegistry 公司发布全球事件系统, 并跟踪事件主题和热度变化。2014 年起, TAC 会议还将事件分析作为核心任务之一, 包括事件检测和事件要素填充等内容[9]。结合丰富的事件分析研究, 进一步展现出结构化事件库的实际价值。

我国针对中文事件探测和提取的研究经历了十余年的发展逐渐成熟, 哈尔滨工业大学社会计算与信息检索研究中心在中文文本事件检测和实体识别算法等领域发布了语言云[10]等多项成果, 国防科技大学信息系统与管理学院近年来提出多种社交网络中的事件发现方法[11-13]。此外, 国内多家高校和研究机构在非常规突发事件仿真建模、网络数据实时监测方面开展研究, 这些研究成果多针对某类特定的事件应用场景, 对构建基于多源大数据的主题事件库提供了方法支撑。

经过 40 多年的发展, 事件库构建技术在国内外学者推动下已在各类数据场景中涌现了大量的学术成果, 并在实用过程中积累了丰富的优化方法。目前缺少对事件库构建技术全面的综述, 比较相关的综述有 2013 年中国国防科技信息中心的高强[14]和 2016 年鹿特丹大学 Frederik[15]关于事件抽取技术的综述, 但事件抽取方法主要关注数据结构化表示的关键过程, 只是事件库构建中的一个重要部分, 没有考虑事件场景中异构源数据、事件类别差异、事件数据不一致、事件间逻辑关联等一系列问题。本文首次系统分析和总结了事件库构建过程中的各类关键技术, 并基于所在团队的相关研究对事件库的实际应用和构建技术的挑战进行了探讨。

本文的结构如下: 第 2 部分介绍了事件数据的基本概念以及事件库的内涵, 建立事件库构建架构

的基本认识; 第 3 部分围绕从底层数据到精细化事件数据的事件库构建过程, 介绍了事件检测、事件抽取、事件融合的主要方法和研究进展; 第 4 部分介绍了事件库在保障国家和社会安全中的实际应用, 以及构建技术对其他应用技术的影响; 第 5 部分对事件库构建中的难点和关键问题进行了讨论和展望; 最后是结束语。

2 事件库定义、分类与架构

2.1 事件数据的定义

事件是事件库的基本表示对象。对于不同的应用场景, 事件的定义有所差别, 影响力较大的定义由 Gerner 和 Schrodt 于 1994 年给出[5]: 一个事件是指在某个时空点上指定类别实体间的特定相互作用, 它可以用自然语言和电子数据方式进行记录, 包含施事主体和受事客体、施事行为、时间、地点等要素。ACE 项目组在 2005 年将事件研究范围限定在符合 ACE 事件类别和角色列表当中[6], 并引导事件识别的发展。NIST 在 2012 年给出事件的宏观解释[16]: (1)是发生在特定时间和地点的复杂活动; (2)涉及人与人或者人与其他实体对象之间的交互; (3)由许多的人类行为、过程以及活动组成; (4)可被直接观测。基于事件的定义, 本文参考相关文献[17-20], 对与事件相关的概念做如下定义:

事件要素: 又称为事件论元, 指构成事件的基本要素, 具有相应的语义角色, 包括行为、类别、参与实体及属性等内容。不同类别的事件数据其事件要素的描述粒度差别较大, 例如国家建交中关注的实体对象是国别, 而在领导人访问等外事活动中, 关注的实体是具体的人物对象, 如表 1 中“: ”前对应的标签信息; “: ”后对应的信息表示**要素实例**, 即事件要素的文本描述信息。

事件数据: 包含事件语义信息的一条记录, 每个字段分别表示事件不同的要素值, 以“键值对”的形式进行存储和计算, 主要分为元事件和主题事件两类: **元事件**表示在特定时刻发生的一个动作或状态变化; **主题事件**表示一个时间窗内与主题相关的多组动作或状态变化, 可由多个元事件归纳而成。本文没有特别说明时一般指元事件数据。

事件库: 集中存储符合特定规范的结构化事件数据的数据仓库, 事件数据是事件库的组成单元。

基于事件数据的描述性定义, “事件数据”的形式化定义可表示为:

$$e=\left\{(k_i,v_i)\,|\,k_i\in\Re,v_i\in\varphi_{k_i}(t_e)\right\}$$

其中 $\Re$ 表示所有的要素名称集合, 例如涉事实体、参与人数、行为类别、时间、地点名称、经纬度、社会学属性(如伤亡损失、是否属于暴力事件)等, $\varphi_{k_i}(t_e)$ 表示从与事件相关的数据 t_e 中提取的符合要素 k_i 的实例。

值得注意的是, 根据事件库设计和表达粒度要求, 事件要素 k_i 的定义空间差异较大。元事件中通常包括时间、地点、涉事实体、事件各类属性等要素, 通常从文本或知识库中抽取获得; 主题事件则更关心事件主题、持续时间、影响范围等根据事件特征生成的要素类别, 通常由海量数据归纳生成, 事件数据样例如表 1。在事件库中, 事件数据独立存储, 同时又可以通过标记或逻辑关联进行融合, 从而实现可拓展的数据服务。

表 1 事件数据实例

Table 1 The examples of Event Data

类别	相关描述	事件数据样例
元事件	E1: 2016 年 5 月 23 日, 伊斯兰国组织在叙利亚沿海城市杰卜莱制造连续汽车炸弹袭击平民事件, 造成至少 78 人死亡。	"Date": "2016-05-23", "Location" :[{"country" : "叙利亚", "city" : "杰卜莱", "latitude":"35.35N", "longtitude":"35.92E",},], "SourceActor":[{"discription":"伊斯兰国组织", "type":"恐怖组织",},], "TargetActor":[{"discription":"当地居民", "type":"平民",},], "Action":["discription":"汽车炸弹", "type":"恐怖袭击","scale":"-7.8",], "Consequence":["death":{"number":78,},],
主题事件	E2: 叙利亚战争自 2011 年初爆发持续至今, 反政府武装和恐怖活动持续升级。	"character":[isReligious":1, "isTerror":1,], "StartDate": "2011", "Location" : ["Country" : "叙利亚",], "SideA":[{" Actor ":"政府军",}], "SideB":[{"Actor":"伊斯兰国","type":"恐怖",}, {"Actor":"努斯拉阵线","type":"分裂",},], "RelatedActor":{"联合国","俄罗斯","法国",}, "Topic":["Label":{"反政府","内战","阿萨德"},], "character":[{"InfluenceDegree":9.7,}…],

2.2 事件库的分类

随着事件库发展日趋成熟, 事件库覆盖了包括国家及公共安全领域在内越来越广的应用范围, 表 2 给出了当前主流的事件库, 并按照如下分类标准进行分类。

按构建方式分类, 主要分为人工、自动、半自动等构建方式。人工构建方式依赖专业人员搜集大量情报信息, 由于容易出现主观误差, 因此在该类构

表 2　目前仍在更新的典型事件库

Table 2　The typical event databases which are still being updated

事件库	发布时间	主导机构	简介	构建方式	数据源	表达粒度	主题类别
Correlates of War event data (COW) [21]	1962	美国密歇根大学、加州大学	第一阶段记录了 1816—2001 年以来的国家间军事争端事件集, 第二阶段记录了 2002 年至今的国家间交火、军事合作和协定等事件	人工+半自动	大型事件由专家通过专业文献整理; 小型军事冲突事件由数据搜集师通过新闻采集	混合事件	军事争端
Computational Event Data System(CEDS) [22]	1994	美国堪萨斯大学	1979 年以来中东地区、巴尔干半岛和非洲西部地区的政治类事件, 前身是 Kansas Event Data Set (KEDS)	自动	路透社、法新社等权威机构英文新闻报道	元事件	政治
Wikipedia Worldwide current events[23]	1997	Wikipedia	针对战争、灾害、犯罪、政客活动等分主题, 提取每天的重要事件新闻的事件摘要和关键主题	自动	新闻源	混合事件	综合
Uppsala conflict data sets(UCDP)[24]	2004	瑞典乌普萨拉大学, 国际和平研究所	1946 年以来全球各类战争、武装冲突与和平调解等大型事件数据, 2013 年后公布图谱数据	人工	专家资料库	主题事件	军事冲突
Global Terrorism Database(GTD)[25]	2005	马里兰大学	1970 年来超过 15 万条全球恐怖事件数据, 每个事件包括 170 多个字段	人工	1997 年前的事件数据通过情报服务中心获得; 1998 年后通过新闻及研究报告分析获得	元事件	恐怖主义
Armed Conflict Location and Event Dataset (ACLED)[26]	2010	英国萨塞克斯大学、德克萨斯大学	1997 年以来的非洲、南亚、东南亚等 60 多个发展中国家的超过 20 万条政治暴力、反叛等事件数据, 包括地区、特定主题(如恐怖组织相关)等多类子事件集	半自动	当地新闻报道、人道主义机构发布的报告及研究出版物	混合事件	政治
Social, Political, Economic Event Database (SPEED)[27]	2010	伊利诺伊大学民主研究中心	1945 年以来的城市内乱和政变事件, 以及出于政治意图的个体事件数据, 还包括社会、经济、宗教、民族和政策数据等	半自动	全球新闻媒体及出版物	混合事件	反叛/内乱
Integrated Conflict Early Warning System (ICEWS)[28]	2010	美国国防部、洛克希德马丁公司	2001 年以来全球范围内超过 2.5 亿条暴力、冲突等事件数据	自动	全球范围内超过 6000 个新闻源以及传统的专家类文献	元事件	暴力冲突
Global Dataset of Events, Location, and Tone (GDELT)[29]	2011	雅虎、乔治城大学、宾州州立大学	1979 年来超过 300 种的自然运动, 包括暴动、抗议、外交、灾害等事件数据; 2016 年后发布图片事件据	自动	全球范围内超过 1.5 万个新闻源, 涵盖 60 多种语言	元事件	综合
Social Conflict in Africa Database (SCAD)[30]	2011	美国丹佛大学国际安全研究中心	1990 年以来的非洲和南美地区的国内抗议、暴动、罢工等社会骚乱事件数据	人工	Lexis-Nexis 新闻数据集以及相关主题出版物	主题事件	反叛/内乱
NewsReader[31]	2011	欧盟信息通信技术第七框架项目(FP7-ICT)	为特定组织或范围(公司、科技、社会安全等 40 余个领域)的非结构文本提供标记好的事件数据, 并提供标准事件语料集服务	自动	英语新闻源	元事件	综合
PublicSonar[32]	2012	荷兰代尔夫特理工大学	包括 Twicident 和 CrowdSense 两个子项, 用于识别特定地理区域内的公共安全事件(如火灾、抢劫、聚众等)	自动	Twitter 社交网络	混合事件	公共安全
Event Registry[33]	2013	Event Registry 商业公司	提供各类新闻和新闻事件分析服务, 结构化事件主要针对于分析事件地点、时间、实体和关键词	自动	超过 7 万个新闻源数据, 其中一半为英语数据, 还包括德语、西班牙语和中文数据	混合事件	综合
EMBERS[34]	2014	弗吉尼亚理工大学、美国情报先进研究计划局	提取并预测拉美地区的城市内乱事件	自动	社交网络、新闻数据、名人博客、经济领域的统计数据等	元事件	反叛/内乱

建方式下逐步形成特定的编码规范, 并为机器构建提供了丰富的知识基础, 如 GDELT 使用的编码规范就是基于政治交互事件库 WEIS、冲突与和平事件库 COPDAB 等人工事件库[3]的编码规范形成的。同时

由于数据精度高, 人工事件库仍在使用和更新, 但由于构建速度慢, 更新周期较长; 人工构建方式主要应用在重大事件的统计分析, 并常用作自动构建方式的对比验证集。自动构建方式以事件抽取为核心任务, 从海量混杂数据中快速发现并筛选出准确的事件要素信息, 并形成高质量的事件数据, 是大规模事件库构建的基础。此外, 在实际应用中还出现了半自动的构建方式, 即对自动搜集数据做一些简单分析, 事件编码过程仍依赖人工的方式。本文所指的事件库构建技术均属自动构建方式。

按事件领域分类, 已有军事、政治、叛乱、冲突、恐怖、公共安全事件等多类主题事件库。特定主题的事件库限定了数据来源, 并根据该主题的社会学属性进行解构, 对事件行为、涉事实体等进行更精细的区分。在特殊情况下, 大规模的综合事件库囊括多类主题, 事件类别也可以作为一项事件要素加入事件数据。

按依赖数据源分类, 可分为新闻事件库、社交媒体事件库、跨媒体事件库等。在构建过程中, 由于新闻文本具有结构统一、句法规范等特点, 基于新闻的事件抽取技术发展较为成熟。随着社交网络的普及, 社交媒体中的用户消息成为事件发生后的第一手资料, 同时包含事件发展走向、群众情感、网络传播规模等多类事件信息。社交媒体事件库通过文本挖掘、网络结构分析等技术对大量用户消息进行整合, 发现并还原事件数据。针对短文本中数据错误率高、冗余、缺失等问题, 跨媒体事件库利用融合关联技术将多类新闻源、社交网络以及知识资源的优势进行整合, 提高事件数据质量, 满足更加复杂的分析需求, 已成为事件库发展的新兴领域。

按照事件表达粒度分类, 可以分为元事件库、主题事件库和混合事件库。元事件库以元事件作为独立的记录单元, 各元事件按照统一的结构规范存储和调用; 主题事件库以主题事件作为独立的记录单元; 通过建立主题事件与元事件的关联关系, 可以建立混合事件库, 在主题事件中追踪元事件的演化关系。

2.3 事件库架构

事件库是大规模事件数据的获取和集成平台, 事件库架构(如图 1)包括两大部分, 分别是数据语义精度递增的多级数据架构和自底向上处理数据的构建架构。

事件库的数据层分别为源数据层、事件候选数据层、事件数据层和融合数据层等四级数据层。源数据层包括各类新闻、社交媒体等海量异构源数据; 事件候选数据层指经过数据预处理后与事件相关的源数据文本; 事件数据层指结构化事件数据, 是事件库的数据核心; 融合数据层包括精细化的事件数据以及由事件数据得到的语义更丰富的事件图谱。

图 1 事件库架构

Figure 1 The framework of Event Database

基于四级数据层, 事件库的构建架构主要包括三个组成部分: (1)事件检测, 识别海量数据中的事件线索信号, 提取与事件有关的候选数据; (2)事件抽取, 从候选数据中识别与事件要素相关的数据特征, 抽取事件要素键值对信息, 按照特定的填充规范, 生成结构化事件数据; (3)事件融合, 结合事件的上下文环境以及要素特征, 合并指代同一个现实事件的事件数据, 提高事件数据质量, 并能将具有逻辑关联的事件数据表示为关联图谱等富语义数据, 优化或扩展事件数据的表达空间。

3 事件库构建技术

事件库构建技术解决的核心问题主要围绕: 如何从海量异质数据中快速发现和识别目标事件, 如何准确地抽取与事件相关的要素信息, 怎样建立同类或相关事件的关联和融合机制, 以及如何在提升构建效率的同时降低开销。针对数据场景的差异和事件数据的多粒度表示需求, 根据事件库架构, 事件库构建的核心技术包括事件检测技术、事件抽取技术和事件融合技术三类。

3.1 事件检测技术

事件检测技术(Event Detection)又称事件发现技术, 是事件库感知能力的基础。根据文献[35]的说明, 事件检测的主要任务可分为两类: 一是回溯事件检测, 从新的文档中识别出已知类型的事件; 二是新型事件检测, 从在线数据流中实时发现新事件。两类事件检测任务都需要在海量数据中识别典型或潜在的事件线索信号, 过滤出事件相关的数据, 并对事

件候选数据进行划分。因此, 本文按照所依据的线索信号分类, 分为基于事件触发词、基于文本特征和基于事件模式特征的三类事件检测方法。

3.1.1 基于事件触发词的方法

事件触发词指最能准确识别特定类别事件发生的关键词, 一般以谓词为主, 如表 1 中“袭击”“爆发”等。事件触发词通常由专业人员手工集成, 常见的触发词库包括 ACE 项目公布的事件触发词语料[36], 政治冲突与调解事件编码规范 CAMEO(Conflict and Mediation Event Observations)中的动词词库[37], 基于 WEIS 改进的 Phoenix 政治事件词库[38]等; 领域性较强的事件(如灾害事件库)也可从 WordNet 开放的知识库中获得触发词。

表 3 基于触发词的事件检测方法优化案例

Table 3 The optimization cases of event detection methods based on trigger words

方法描述	案例
触发词同义词库	建立同义词关联规则, 例如“两国和谈”“双边洽谈”等均表示外交合作类事件[39]
配套剔除词匹配	冲突类事件与体育赛事事件都常出现“两国对抗”等类型的词语, 通过添加赛事名称等剔除词过滤掉无关事件[18]
设置事件触发词数量阈值	在社交推文中包含的 3 个以上的事件关键词时标注为与事件相关的数据[40]
包含特殊实体	少数特殊实体(如本·拉登、基地组织等)可以作为与恐怖主题相关的活动线索[3]
停用词分布	网页中可能只出现包含触发词的标题链接, 当网页文本节点中停用词数量越多, 相应内容越重要[41]
元数据特征	优先在标签元数据(如微博 hashtag)[42]或中心标题元数据中的触发词匹配
设置触发词时间属性	常规周期性事件如国际会议、战争纪念活动等, 对相关触发词添加时间属性[3]
后缀树查询方法	后缀树是一种便于字符串快速处理的数据结构。数据流按时间块切分, 对块内触发词(或关键词块)做加权频率统计后存入后缀树, 通过后缀树 Ukkonen 算法能快速识别最近时间内的高频触发词[43]
触发词优先级方法	设置触发词的优先级, 启发式的构造优先决策树, 融合多种方法的优势[44]

基于触发词的事件检测原理是爬取数据时发现与触发词匹配的文本, 将与触发词有关的上下文加入到事件候选数据中, 从而过滤事件无关文本。因此, 在该方法中不用考虑数据量以及文档间的联系, 每份文档的事件检测过程相互独立, 使事件检测和相应的事件抽取过程(详见 3.2.1 节)能够流水执行, 操作方便灵活, 在新闻、专家博客、智库文献等表达规范的数据场景中表现良好。同时, 由于更新触发词受人工构建的限制, 该方法的移植性较差, 并且容易发生事件漏报、误报等错误。在应用中, 通常根据数据场景和事件语境, 结合多种启发式的优化方法使用, 常见方法如表 3 所示。

3.1.2 基于文本特征的方法

在复杂的数据环境中, 事件通常没有统一的规范表述, 不仅在新型事件发现中缺乏相应的触发词匹配, 而且旧触发词在使用中可能发生语义变化, 因此需要通过从数据中得到事件的线索, 即事件的关键特征, 从而实现动态事件检测。基于文本特征的方法是选取有效的文本特征表示方法, 从特定数据场景中学习某个或某类事件的特征空间, 通过适当的文本挖掘和检索算法, 实现事件候选文档的筛选。在该方法中, 文本特征选取的优劣直接影响区分事件的能力和数据处理效率, 因此如何选择出特定事件的文本特征是发现事件的关键。

传统的文本特征选择主要依据文档中词项的统计量, 如互信息、特征词 TF-IDF 值、信息增益等, 通过将检测过程转化为符合相应目标的分类或聚类问题, 发现数据流中的事件。Kjoseph 等[45]在分析“阿拉伯之春”中的冲突事件时, 先构建了该类的关键词库, 根据词项与冲突类别的互信息, 使用贝叶斯分类方法选择出关于冲突类事件的文本。EDCoW 系统[46]针对每个词项信号的 TF-IDF 进行基于滑动窗口的小波变换, 从而实现集中特征的简洁定义; 同时, 根据词项序列的互相关强度, 构造带权的特征信号连接图, 通过实现模块度(Modularity)最优化的子图分解, 得到每个子图对应的事件特征, 实现事件识别。此外, 大规模事件库在特征建模时会融合数据流特性, 例如 EventRegistry 事件库[47]对在线数据流进行实时分组, 根据新闻的标题、标签和涉及的实体名称对文档生成特征向量, 将文档词频作为特征值, 并对源数据流按时间窗进行划分, 同个窗口内计算文档距离并实现在线文档聚类[48], 新文档加入后较前一个时间窗所产生的新簇作为新的事件, 新簇内对应的文本加入该事件候选数据。基于传统特征的事件检测已发展成熟, 文献[49]详细介绍了其他常见的目标数据检测方法, 同样可以应用在事件库构建中。

在近年的事件检测研究中, 结合情境的文本特征选取方法在语义丰富的事件发现中表现出更强的应用价值, 本文介绍 3 种常见方法:

1) 基于事件主题的特征选取。将事件看作特定的主题, 把事件检测转化为文档与主题特征的关联检测, 不仅可以发现更多类别的事件, 还可以融合事件的演变和迁移过程并发现相关事件。事件主题

特征的提取与典型的话题分析类似, 使用词袋假设, 将文档看成不同事件主题上的分布, 并对“文档-主题”“主题-特征词”进行潜在语义分析, 实现文本与主题相关性的判别。文献[50-51]等在新闻数据中应用 LDA 话题模型, 实现了回溯事件检测。同时, 越来越多的研究围绕动态数据场景的时序和并发特性, Wang 等[52]提出基于时序的 TM-LDA 模型, 能够在社交网络消息流中检测到突发事件; Wei 等[53]将推文的发布位置与事件位置的概率依赖关系加入到文本话题模型, 既可以得出推文所属的事件类别, 又能推断事件相关数据的地理覆盖范围。事件的主题特征关联时还应该考虑上下文信息对语义的影响, Wang 等[54]在文档的主题学习过程中融合基于词嵌入向量的语义学习, 构建事件的词义归纳模型, 从而降低事件检测中歧义词的干扰, 实现更紧凑的关键词特征学习。

2) 基于动态检索的特征提取。当已知事件的一部分特征时, 可将该特征作为限制条件和种子信息实现迭代检索, 在每一轮中根据搜索结果中词项的重要程度进行排序, 选择排名较高的词项加入到查询框架中, 直到特征词收敛停止。这种方法可以在学习事件文本特征的同时提取事件候选数据, 还可以通过特征变化的轨迹刻画出事件演变的过程。Naren 等[55]在 Twitter 网络中根据内乱事件的种子特征词, 逐渐发现了“热带之春”事件中的罢工、游行等元事件关键数据。Abel 等[56]则在跨媒体环境中实现了事件候选数据的动态扩充, 首先基于传统媒体(如公共安全服务网站)中的事件记录构造出相应的事件表示框架 *Profile*(*e*), 继而以该框架为种子在社交网络中查询符合条件的推文信息, 并对每条推文构造相应的推文表示框架 *Profile*(*t*), 选择 *Profile*(*t*) 与 *Profile*(*e*) 相似度高的推文加入事件候选数据。动态检索还可以与主题分析方法相结合, 解决词袋假设中缺乏文本间的时空相关性而导致的内容同质化问题。Blei 等[57]提出了基于层次化话题的事件信息检索方法, 通过在话题聚类中嵌入中国餐馆过程, 获得用于数据集成的非参数贝叶斯模型的先验分布, 从而在检索结果中构建出关于事件、话题、文本的层次树。

3) 基于知识图谱的特征提取。知识图谱中记录了事件相关实体与行为的语义关系, 常见的表示形式包括本体语义网、知识库等, 通过对知识图谱和事件文本的联合学习, 分析特征词所属概念及语义关系, 可突破词袋模型中词项无关性假设的限制, 更精确地发现或过滤事件特征; 同时由于获得了语义映射关系, 便于直接集成事件要素抽取过程。Vanni 等[58]通过危机事件知识图谱(如图 2)分别构造了武装袭击、自然灾难、示威叛乱事件的触发特征语义网, 能直接用于检测包含目标事件各类特征的推文, 同时还可根据知识图谱构建事件抽取规则用于要素抽取。此外, 根据本体语义链进行特征选取还能简化由浅层语义关联构造的混杂网络。例如 Andrzej 等[59]在社交网络环境进行恐怖事件检测中发现, 综合用户关联和消息语义关联生成的 Twitter 网络中存在大量孤立的节点(如媒体名称、特殊语气词等), 根据恐怖事件语义网, 可以过滤掉其中离散无意义的文本特征, 得到事件的关键特征。近年来, 迁移学习算法的进展推动知识图谱在复杂数据场景的广泛应用, Huang 等[60]利用基于知识图谱的主题学习到微博的迁移学习方法 CTrans-LDA, 实现在线数据流的事件检测并实现事件类别标注。

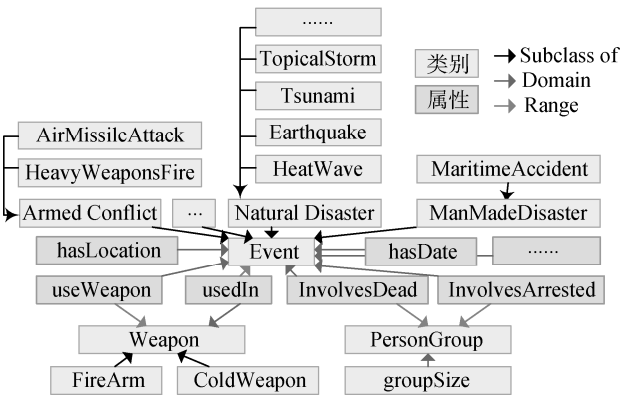

图 2 一种危机事件系统中的知识图谱(节选)[58]

Figure 2 The knowledge graph of a crisis event system[58]

基于文本特征的事件检测方法可用于各类数据来源中, 特别是在社交网络环境内, 能够实现事件的实时感知, 既能实现回溯事件发现, 又能识别出非频繁事件和新型事件; 但由于多采用基于大数据的分析方法而容易遗漏数据量较少的事件。上述两类事件检测方法都是从事件的描述文本出发, 忽略了现实世界中事件发生时动态的关联条件和特定模式。因此, 在事件检测中往往可以结合模式特征, 提高检测效果。

3.1.3 基于事件模式的方法

事件数据记录了现实社会的人类活动和重要变化, 事件发生时往往伴随丰富的模式特征, 例如城市突发事件发生时通常出现人流变化波动点[61], 暴恐事件呈地域性扩散[1], 某地发生重大事件后社交网络消息流量变化[62]等。这些模式特征可以作为发现事件的重要信号, 在非平稳、不均匀海量数据流中

实现事件监控。基于模式特征的事件检测方法就是通过挖掘历史事件发生的模式特征, 当新的数据流中出现该模式特征时触发相应的事件检测模块, 有针对性的执行数据获取程序, 是一种能对事件进行预判的启发式检测方法。常见的可用于检测的事件模式有以下几种。

频繁模式是指一类事件发生时频繁出现的事件或事件要素项集, 常用的发现方法有 Apriori 算法、FP-Growth 算法等。由频繁模式可以生成强关联规则应用到事件监测中, 例如通过恐怖事件的地点要素记录可发现伊拉克地区恐怖事件爆发的地点呈延路网和河流扩散的趋势, 从而限定了事件检测的地理范围, 提高事件发现效率[63]。Johnson 等[64]基于俄罗斯社交媒体 Vkontakte 通过对 ISIS 群组变化规律分析, 发现地区临时恐怖团体数量增长速度超过一定阈值时该地区恐怖事件数量显著增加, 利用该模式可以实现重点时段的监控。近年来, 众多学者关注事件频繁模式挖掘的优化算法, 例如 Hasan 等[65]通过构建并遍历模式树, 实现新模式发现和不定规模的频繁项集的快速统计; Cule 等[66]通过定义事件要素之间的连接紧密度(Cohesion)降低了模式搜索的复杂度。此外, 文献[67,68]等还针对统计频数较低但事件关联时统计依赖的事件模式进行探索, 通过依赖模式来降低发现常规频繁模式的阈值门槛。

序列模式是指从序列数据库中发现频繁的子序列。不同于频繁模式, 序列模式重点关注于发现事件与事件的先后序列关系, 常见的分析方法有 GSP 算法、FreeSpan 算法等。事件的序列模式可以一定程度地反映事件的顺承或因果关系, 通过易于观测的事件辅助复杂事件的检测过程。Souza 等[69]根据特定词汇、语法等多种语言学知识分析出 14 种事件与事件间的序列关系, 如诱因、并列、嵌套、顺承等, 可在事件检测中识别出与其有序列关系的事件。Zhou 等[70]不仅实现事件频繁序列发现, 还根据实时数据流的频率统计算法(如 lossy counting 算法)随滑动窗口动态继承和更新事件序列模式。Mirza 等[71]根据事件的序列关系挖掘序列闭包和反转模式, 进而发现事件的因果关系(如使能、阻碍等)的断言特征, 实现事件的序列识别。

时序模式是指事件时间序列与其它时间序列的相关关系, 常用分析方法有谱分析法、ARIMA 自回归滑动平均模型等, 通过易于监测的时间序列判断目标事件发生的可能性。例如, GDELT 子项目[72]根据阿富汗各省历年的暴乱事件对各地同类事件发生的可能性实现短期预测。在多序列的时序模式发现中, 序列间的时滞影响也可以作为事件的影响要素用于事件发现, 例如, 文献[73]对地区食品物价变化序列与当地冲突事件数量序列联合分析, 通过 Granger 时间序列的因果检验, 发现物价变化与冲突事件爆发的联系。时序模式常用于发现正在发生或即将发生的事件。

周期模式指特定事件发生的周期性特征, 周期可以通过先验知识获得, 还可以通过统计假设检验、离散傅里叶变换等常用算法计算周期。例如, Pietro 等[74]通过高斯过程回归, 发现社交网络中具有周期性的话题标签, 从而发现对应的社会事件的周期性, 例如赛事、纪念日活动等, 一方面用于发现事件的周期, 另一方面在事件发现中降低普通事件带来的干扰。在知道事件的周期模式后, 事件检测任务可变成事件库的常规任务, 降低事件信号发现的难度, 例如 2016 年底的韩国烛光集会事件爆发具有周期性, 通过事件检测过程, 可过滤出抗议主题、抗议人数、地点及规模等。

突发模式指事件在短时间内集中发生但在其他时间发生的可能性较低的情况, 可根据状态机理论、指标系统异常监测、演化突变理论等方法发现。一

表 4 事件检测技术总结

Table 4 The summary of event detection methods

事件检测方法	优点	缺点	适用的数据场景	代表技术与工具	是否可检测新型事件
基于事件触发词的方法	数据源间并行过滤, 可与事件抽取过程组成流水线	跨事件领域的可移植性差, 易漏报误报, 需要人工定期更新触发词库	新闻, 专家博客, 智库文献等	CAMEO 事件词库[37], Phoenix 政治事件词库[38]等	否
基于文本特征的方法	可检测的事件类别灵活, 能反映事件的动态演化特性, 适用于复杂数据环境	易遗漏数据量少的事件, 获取数据的冗余度高	新闻, 社交网络, 跨媒体数据环境	EDCoW[46], EMBERS[55], EventRegistry[47], TM-LDA[52], Ctrans-LDA[60]等	是
基于事件模式的方法	能实时发现正在发生的事件, 可检测非频繁事件	需要配合其他事件检测方法, 灵敏度易受支持度设定影响	新闻, 社交网络, 经济、交通流量等特定检测指标数据	ISIS 群组活动频繁模式[64], SPEP 序列模式预测方法[70], ARIMA[72]等	是

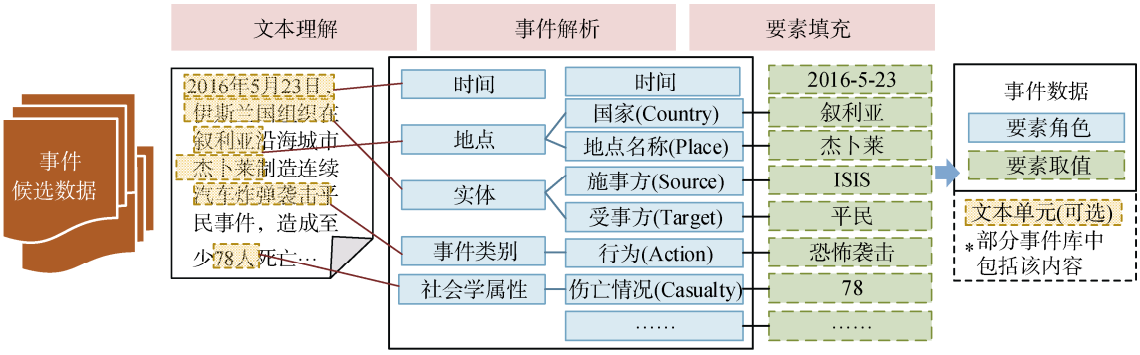

图 3　事件抽取的主要任务

Figure 3　The main tasks of event extraction

般的可将突发事件可以看为低到达率的事件，因此通过训练状态机的转移概率来发现突发事件。He 等[75]在基于 TF-IDF 发现事件文本特征时对于每个特征分配一个正常状态和爆发状态的二元有限状态机,文档流的特征表示中加入状态转移因子，从而发现文档流中的突发特征，用于突发事件发现。突发事件与指标检测进行关联，如人流量指标检测[76]、舆情指标(如讨论热度、某话题在集中地域的参与人数等)检测[77]等，可发现突变时间点；通过搜索指标异常发生前后的数据，提取与事件相关的详细信息。基于演化的突发模式分析，可以发现重大的突变事件，例如 Yasuko 等[78]根据数据流中的在线演化，发现政权更迭事件。突发事件还可以看作周期很大的事件，陈宏等人[79]应用事件周期小波变换方法实现事件可伸缩滑动窗口中的突发数据流检测，并发现事件的突发特征。

基于事件模式特征的事件检测方法能充分发挥事件库的技术优势，将历史事件中发现的模式特征应用到新的事件发现过程；同时，不依赖特定类别的训练语料集，就可以直接用于事件发现，特别是在复杂事件或非频繁事件中能发挥重要作用。该方法在事件检测中可以与其他方法组合使用，系统适用度高，但依赖于事件库应用技术中的模式发现能力，同时模式支持度阈值的设定影响模式应用效果，阈值过低会混入大量不重要的模式，反之可能产生漏报。该方法还在事件的早期预测中发挥重要作用。

表 4 对各类事件检测方法进行了总结和比较，基于事件触发词的检测方法最简单易用，基于文本特征的检测方法应用范围最广泛，基于事件模式特征的方法提升了事件发现的及时性。在大型的事件检测系统中，针对复杂的数据环境往往需要综合应用多种事件检测方法。例如, Twicident 系统[56]在初始事件发现阶段使用基于触发词特征的方法，后根据社交网络中的文本统计特征发现事件相关数据。NewsAnalytics 事件系统[80]则综合了上述三类方法实现跨媒体数据的事件检测，既应用触发词方法在传统媒体中实现事件发现，又应用事件的文本特征和模式特征实现社交网络以及博客、资讯类数据场景中的事件检测。通过事件检测获得的事件候选数据仍属于非结构化数据，需要进一步细粒度的信息抽取得到事件数据。

3.2　事件抽取技术

事件抽取技术(Event Extraction)是指从松散的非结构化信息中抽取事件要素信息并生成精细的结构化事件数据的技术，是事件库构建技术的核心。与事件检测中发现事件关键特征不同，事件抽取需要对完整的事件特征所对应的要素类别进行识别。由于事件数据由不同要素的键值对构成，事件抽取的主要任务包括三方面(如图 3)：一是文本理解，事件的描述文本经过句法成分分析能分割成包括独立语义的文本单元，理解文本单元的语义角色；二是事件解析，识别一则事件数据包括哪些要素单元，如实体、关系、时间、地理信息以及涉事人数、事件类型等属性信息，可由人工设定，也可根据文本理解结果自动生成；三是要素填充，按照要素单元的填充要求，将文本单元转换成符合规范的属性值；后两方面也可以构成事件要素抽取问题。事件抽取过程可形式化表示为解决匹配问题，即 $f: X \to Y$ 其中 X 代表文本输入，Y 表示输出的要素单元空间 $Y \subset \Re$，抽取的关键是学习映射关系 f。根据学习方法分类，可分为基于规则匹配的方法以及基于机器学习的有监督、无监督和弱监督事件抽取方法。

3.2.1　规则匹配方法

基于规则匹配的事件抽取方法是在一些规则模板指导下识别文本中事件要素的启发式方法，主要包括规则模板集合和规则执行引擎两部分：规则模板集合指明了要素单元的上下文约束，融合了领域

知识和语言知识; 规则执行引擎是运用规则进行事件抽取的程序, 执行文本解析及规则匹配算法, 控制处理规则的策略, 将要素识别结果合并为事件数据。基于规则匹配的事件抽取方法在语言表达规范的时事新闻或领域知识积累丰富的单一领域(如政治、军事、灾害事件等)文本中非常有效。

规则模板集合是规则匹配方法的核心, 其获取方式随着事件抽取精度要求的提高不断改进。传统的获取方式主要由人工编写, 领域专家根据事件要素目标的“词法-句法”模式和“词法-语义”模式, 构造特定形式(如正则表达式、词法模式标注)的事件规则模板, 如实体规则、行为规则、事件类别字典等。GDELT 提供的事件数据采用麻省理工学院的 Phoenix 字典库[38]和堪萨斯大学的 CAMEO 字典[37](如图 4), 根据事件句的浅层文本分析结果, 与字典中定义的模式相匹配, 得到事件编码。近年来, 为了适用于更复杂的语言环境, 逐渐发展出新型的规则平台。Apache 基金会推出的 UIMA Ruta 开放规则系统[81], 采用众包方式便于用户动态调用更符合语言场景的规则。此外, 越来越多的研究关注于抽取规则的自动生成方法, 例如 Jiang 等[82]通过相似语句的短语挖掘方法发现要素抽取的元模板发现, 此外文献[83,84]介绍了应用机器学习方法并通过提高分类结果的可解释性构造规则的方法。

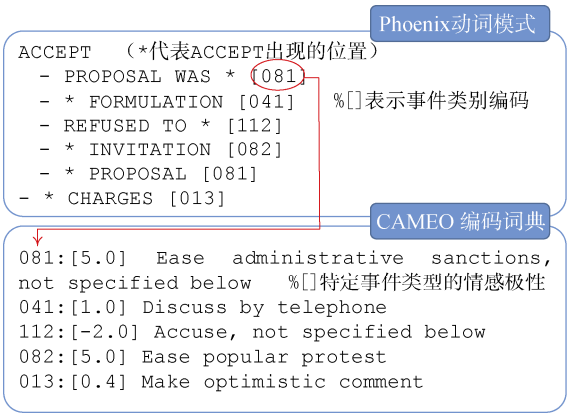

图 4 GDELT 事件库在事件抽取中采用的匹配规则

Figure 4 The typical matching rules for event extraction in GDELT event database

规则执行引擎决定了规则匹配方法的执行逻辑和运行效率, 提供了一系列匹配算法和优化策略。规则执行引擎可采用正向演绎的规则匹配算法(如常见的 Rete 算法[85]), 从事件文本出发, 运用要素约束规则, 找到满足约束的要素实例; 也可以采用反向归纳的算法(如 FOIL 算法[86]), 从特定类别的事件描述规则出发, 在文本集合中分别提取匹配的事件要素。在实际的规则匹配事件抽取系统中, 正向演绎的执行逻辑最为常用, 如 GDELT 开源的 PETRARCH[41]、OPEA 下的 TABARI[38]均采用该方法; 反向归纳方法可同时应对多要素跨文本匹配的情形, 如 REEE 系统[87]对 100 多类要素模板归纳生成统一的要素抽取模块, 同时关联多个新闻源实现抽取。当规则集合不断增大时, 匹配时间面临指数级增长, 并且容易出现规则冲突的情况, 因此需要恰当的规则优化策略。OFEE 系统[88]应用模糊推断方法, 实现过滤、抽取、分类等多种规则的集成应用。李培峰等[89]在中文事件抽取规则匹配中加入了同文档内关联事件的联合学习机制, 根据富要素实例推导贫要素实例, 要素识别和要素角色分配等方面的精度明显提升。规则执行引擎在设计时还会融合软件开发中的常用方法, 提高规则可重用性和可配置性, 在实际应用中越来越成熟。

基于规则匹配的事件抽取方法通过构建事件和要素抽取规则来模仿人的推理方式, 在扩展性、简洁性、调试性和运行速度上都具有良好表现, 至今仍是商用事件库中最常用的事件抽取方法。但是, 该方法忽略了事件爆发后海量数据的潜在关联, 难以适应各类事件表达需求的细粒度差异, 因此, 借助机器学习方法实现事件抽取逐渐成为研究热点, 并应用于大规模事件库构建。

3.2.2 有监督方法

有监督的事件抽取方法使用要素角色标注好的事件语料作为训练集, 采用机器学习方法构造要素标注模型并学习模型参数, 实现新输入的文本到特定要素类别的预测; 其核心是文本单元的要素角色分类, 构造模型时需要考虑上下文分类结果。

序列标注模型是最常用的有监督事件要素标注方法, 能表示观测序列和状态序列的整体关联特征, 通过将输入文本作为观测序列, 输出的要素类别作为状态序列, 在事件抽取中常用算法包括最大熵马尔可夫模型(MEMM)、条件随机场模型(CRF)等。MEMM 模型通过将最大熵模型和隐马尔可夫模型(HMM)相结合, 克服最大熵模型中的类别独立性假设和隐马尔可夫模型中的观测独立性假设的限制, 将上下文信息引入到模型的学习和识别过程中。McCallum 等[90]最早将 MEMM 模型应用在事件抽取中, 大幅度提升了 ACE 项目中的事件抽取任务的识别精度, 在此工作基础上文献[90,91]分别在暴力事件和恐怖事件抽取中进行优化, 实现事件主体、主体关系、事件类别等要素的抽取。CRF 模型则对整个序列的联合概率统一建模, 避免了 MEMM 模型中

的标记偏置问题, 并且可以融合复杂、可重叠和非独立的特征进行训练和推理。Li 等[93]以 Foursquare 数据得到个体活动的标记数据, 训练 CRF 模型从而可以实时抽取出事件的地理位置要素。陈箫箫等人[94]将词语的上下文特征加入 CRF 模型观测特征模板, 实现微博中的开放域事件抽取。

近年来, 新型的特征表示方法能够融入文本的结构特征和语义特征, 进而提升要素分类的学习效果, 并越来越广泛的用于解决事件抽取问题。构造特征联合模型是一种常用的特征优化方法, 文献[95]在事件句范围内通过联合观测单元与状态单元中的词法、句法、实体类型等 25 种本地特征和共现、关系指向等 8 种全局特征训练感知机分类模型, 要素划分效果较传统分类器显著提高。基于树核的事件抽取算法能根据要素类别和文本单元在句法分析树中的树结构相似性, 构造恰当的树核表示并在训练集中学习合适的分类器, 从而融合结构特征, 高源等人[96]就应用了基于卷积树核的 SVM 分类方法实现事件实体要素和关系要素抽取。此外, 深度神经网络可用于自动学习文本的语义特征, 在标注语料充足的情况下, 该类事件抽取方法较传统方法有大幅度改进。Zhang 等[97]在已标注的中文文本中, 通过动态监督训练深度信念网络模型(Deep Belief Network), 从而得到词语的深层语义特征, 实现了突发事件的要素标记分类(如图 5)。除了上述特征优化方法, 多分类器的 Boosting 方法也可以从分类结果中提高要素抽取的精度。

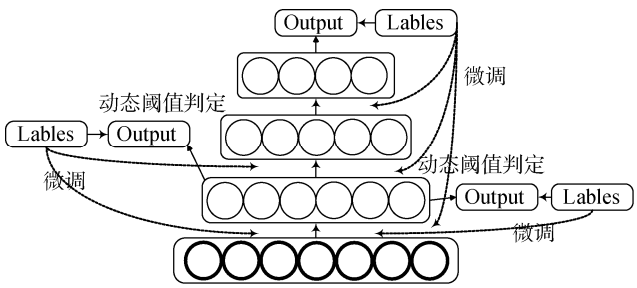

图 5 一种动态监督下的深度信念网络模型[96]

Figure 5 The dynamic-supervised Deep Belief Network model [96]

此外, 为了提高事件抽取对新数据的适应能力, 在训练中能够反映事件信息随时间的语义变化(例如“烛光集会”原多发生在哀悼事件, 后用于反政府示威, 2016 年底又特指“倒朴”抗议活动), 学者们对渐进式学习方法进行探索, 应用于事件中关键实体和关系要素识别, 通过渐进式的知识更新, 不必在样本更新时重新对全部数据进行学习, 降低了训练中的时间和空间需求。吴广财[98]通过构建新旧知识融合模型在基于 HMM 的实体识别模型中实现参数的增量学习。斯坦福大学开放的 DeepDive 结构化知识库构建系统在关系学习中迭代入新构造的训练样例, 通过远程监督发现实体关系[99], 以及事件数据中关系类要素(例如人员与组织、施事与同谋、行为与事件类别等), 还可以通过关系学习实现地点、时间等多粒度要素的自动补全。

有监督的事件抽取方法能够直接获得事件文本单元到事件要素的分类结果, 由于充分利用了事件的先验知识, 事件抽取精确度较高。然而标注训练数据需要花费大量人力, 对于网络动态数据, 新型事件要素往往由于缺乏事件要素的标注语料而被遗漏, 此外事件爆发前后往往伴随大量同质数据产生, 经常采用不需要或只需要少量标注数据的无监督和半监督方法实现事件抽取。

3.2.3 无监督方法

无监督的事件抽取方法根据文本潜在的内部特征(如句法依存关系、文本相似性、语义相似性等), 将输入文本中相似文本单元划分到同一类(簇)中, 再建立该类(簇)中文本与要素特征的关联关系, 找出中心文本单元对应的要素角色; 其核心是同类文本单元的划分方法, 常用算法包括数值聚类算法、子图分割、非负矩阵分解方法等。

数值聚类算法的目标是根据文本差异的合理度量, 将概念差异较大的输入文本划分到不同的簇, 簇的中心即为关键文本单元, 再辨别对应的要素角色。Zheng 等[100]先对一个事件相关的微博集合分析句子成分, 得到主语词簇集合、谓语词簇集合和宾语词簇集合, 然后对处于不同句子成分中的词语簇进行链接, 得到多个＜主语词簇, 谓语词簇, 宾语词簇＞的组合, 这些组合即代表了涉事主体、实体关系、涉事客体的事件语义要素。Liu 等[101]通过加权无向二分图提取事件关键词, 并应用层次聚类算法发现关键词中的实体, 在实时新闻中发现实体要素。在聚类过程中, 文本关联度可以根据语境做调整, 例如 Ji 等[102]综合事件要素在文档层和聚类中的频率表示方法, 在“文档-触发词”“句子-触发词”“文档-角色”等六类置信加权频率基础上计算三类边际效用, 基于上述度量方法构建关键要素推断规则, 实现跨文档的要素抽取。

文本单元间的句法依存关系构成图结构, 特定类别的语义角色具有相似的拓扑特征, 通过拓扑挖掘发现划分子图, 子图内的文本单元表示同类语义角色。例如, 一些事件库将事件要素类型划分为描述

型、数值型和断言型, 其中描述型要素的取值通常是一系列文本, 如事件摘要、事件状态等; 数值型要素可用数值进行量化, 如事件时间、死亡人数等; 断言型要素则表示事件性质, 如是否发生武力冲突、财产损失等。文献[40]发现不同类型的事件要素对应的文本在句法依存关系中存在差异, 描述型要素一般为星形结构, 中心为键值, 连接节点为相应的文本描述; 断言型要素则会产生对特定词的强二元连接; 数值型要素的键名与数字强连接; 基于上述拓扑特征进行子图分割, 可以识别出子图对应文本生成的键值对。

非负矩阵分解方法可以将高维数据近似成多个低维因子的乘积的形式, 从而描述出数据中的潜在结构。在事件抽取中, 基于文本间或要素项间的关联, 可构造文档和词项的加权非负矩阵, 通过矩阵分解得到词项的隐式表示向量, 以及文本的划分矩阵和要素特征的划分矩阵以及样本簇与特征簇的关联关系。Hao 等[42]构造了词项与新闻文档和社交网络文本分别关联的双层模型, 通过加权矩阵区分缺失词和观察词, 并利用不同文档中同类标签、相似实体和已知的时序关联, 使新闻和推文两个加权矩阵分解中的词向量计算相互影响, 从而得到相似文档与相似词项, 实现了预标记要素信息的扩充, 再根据词项所属的句法标记找到相应的要素角色。

无监督的事件抽取方法, 能够突破标注数据限制, 提高对实时数据抽取任务的适应能力, 扩展事件要素类别的多样性, 实现新型事件要素的提取和标注。但是, 由于该方法依赖大量的指代同一事件的相似数据, 无法实现只有少量候选文本的事件抽取, 容易遗漏较少被提及的关键要素信息; 与此同时, 对于“同物异谱”或“异物同谱”现象需要更严格的限制和区分, 往往需要进一步判别或解释后, 才能得到可靠的要素标记。因此, 近年来在无监督的事件抽取方法中需要依赖启发式标记判别[103]或人工交互等修正方法, 以便提高标记准确率; 同时为了发挥无监督方法在发现新事件关键特征的优势, 在事件数据中增加更多类别的描述性要素字段。

3.2.4 弱监督方法

弱监督的事件抽取方法能够从少量标注、弱标注、高噪声的事件样本或相关知识库中训练学习器, 对缺少标注的事件候选文本进行关键要素识别, 既缓解了有监督方法中需要大量精确标注数据的困难, 又解决了无监督方法不精确的问题; 其核心是建立合适的标记推断模型。近年来, 越来越多的学者针对丰富的标记数据场景提出新型事件抽取方案, 下面介绍几种典型的弱监督事件抽取方法。

对于标注不充分的情况, 通常采用半监督方法实现标记扩充, 常用算法包括 Bootstrapping 自举算法、基于图的标签传播算法、归纳学习算法等。Bootstrapping 自举算法根据标注数据中的实体对来发现语义模板, 再针对新语料进行迭代抽取以发现新的实体对, 但该方法容易出现语义漂移问题; 为解决该问题, Roman 等[104]将多学习器引入自举程序, 完善文档相关度, 从而获得更精确的模板; Carlson 等[105]通过该方法发现事件要素互斥和类型检查约束的耦合模式, 从而提高要素填充的准确率。基于图的标签传播算法根据样例之间的几何结构构造图, 用图的结点表示样例, 利用图中的邻接关系将有类标签的样本向无类标签的样例传播, 典型算法包括随机游走、Laplacian 变换、LP-ZGL 方法等[106]。归纳算法实现交叉类的集成学习, Hong 等[107]利用直推学习方法构造跨实体的事件描述文本分类器和语义角色分类器, 根据实体出现在特定事件的情景推断共现实体和实体类别, 例如“平民”出现在恐怖类事件时往往对应“伤亡人数”这一事件要素, 同时推断出应存在袭击者等信息, 进而从其他文档内容中填充该要素单元。此外, 文献[108]中列举了其他用于标记分类的半监督算法, 可用于解决要素标记问题。

除了对于已标注数据的标记扩充, 开放的知识数据、特定类别的网站数据等也可以作为弱标注数据提供知识信息, 可采用迁移学习、主动学习、多标记多示例学习等方法实现文本中的事件要素抽取。Wei 等[109]提出由长文本到短文本分类的迁移学习模型, 通过隐式语义分析方法, 分析事件候选数据中的关键词标签, 并以关键词为节点构造无向图, 通过拉普拉斯特征映射到低维空间, 从而在该特征空间内表示短文本, 将样本分类器通过最小互信息计算拓展到短文本分类器。Angeli 等[110]将主动学习方法应用到远程监督算法中, 基于多示例多标记关系抽取算法从 Freebase 和 TAC KBP 知识库的实体关系推断文本中的实体关系, 进而映射到相应的语义单元。此外, 弱标记数据学习中还可结合标记传播算法, 华盛顿大学的 FIGER 系统[111]利用 Wikipedia 数据, 通过自动序列标注算法对文本自动标注, 实现名词性实体发现; Navigli 等[112]根据 Wordnet 的语义网信息, 面向事件句出现的词项生成所有可能的语义连接子图, 应用局部连接度和全局连接度最优搜索算法, 选择出最优子图的语义路径作为各词项单元的正确语义, 从而实现准确的语义角色映射。

弱监督的事件抽取方法能够充分利用网络中的各类结构化标记信息和知识数据推断要素间关系, 大大减少对标记样本的依赖度, 并提高了事件抽取方法对环境的适应能力; 特别是在学习名词类、动作类要素时具有良好效果。但是, 由于事件数据属于短期数据, 通过长期关系进行的要素角色推断仍然容易出现偏差。随着弱监督事件抽取方法研究的发展, 跨媒体、多源数据融合场景下的事件抽取在事件库构建技术中蕴含巨大潜力。

表 5 对事件抽取技术进行了总结。在实际应用中, 一条完整的事件数据需要对不同的事件要素进行抽取。由于要素类别的语义差异与识别精度的变化, 在进行事件抽取时会采用多类抽取方法的组合。例如 EventRegistry 事件库[47]在事件时间和地点要素抽取时采用规则匹配, 实体抽取采用有监督方法, 通过聚类方法识别出人员伤亡、事件性质等要素属性等, 此外还包括事件摘要、关键词等描述性属性。Storybase 事件库[113]先采用有监督方法抽取所有命名实体, 然后再用规则匹配方法对实体对应的要素名称进行编码。文献[114]则介绍了一种针对社会活动事件的跨媒体抽取方法, 先用规则匹配和有监督方法从咨询网站中抽取社会事件的时间、地点、类别等属性, 再基于知识类数据通过半监督方法从 Facebook 中提取人员、内容等属性, 并对关键要素内容进行核实。因此, 事件抽取技术在不同数据场景和任务类别中具有较高的灵活性。

表 5 事件抽取技术总结

Table 5 The summary of event extraction methods

事件抽取方法	方法类别	代表算法	优势	特点
规则匹配方法	正向演绎方法	Rete 算法	各文本独立处理, 提高并行化	扩展性、简洁性、可调试性强, 依赖规则模板更新
	反向归纳方法	FOIL 算法	可得到结构字段统一的事件数据	
有监督方法	序列标注模型	最大熵马尔可夫模型	克服最大熵模型中的类别独立性假设和隐马尔可夫模型中的观测独立性假设的限制	依赖充足的已标注事件样本, 抽取精度高, 对动态数据缺乏可扩展性
		条件随机场模型	避免部分标记偏置问题, 可以融合复杂、可重叠和非独立的特征进行训练和推理	
	特征融合方法	特征联合模型	文本单元与要素角色的文本特征进行全局关联	
		结构特征模型(如树核特征)	可利用句法分析树结构相似性进行分类, 提高抽取准确率	
		动态监督的深度信念网络	标记数据充足情况下, 利用隐式特征学习进行分类的精度高	
无监督方法	数值聚类算法	渐进式增量学习	可更新抽取模型以适应事件要素的语义变化	无需标注数据, 依赖大量同类数据, 关联要素单元需要启发式判别
		基于相似性的聚类	最基础的聚类方法, 聚类中心即为要素单元的语义中心	
		在线聚类方法	解决实时数据流中的键值对聚类问题	
		关联度可调的聚类	通过调整不同要素关系间的聚类强度, 提高语境的适应性	
	子图分割	句法依存关系拓扑图分割	可按照描述型、断言型、数值型等要素类别进行抽取	
	非负矩阵分解	加权文本矩阵分解	便于跨文档要素抽取以解决上下文缺失引起的要素识别困难	
弱监督方法	标记扩充方法	Bootstrapping 自举算法	最简单实用的解决标记不充分问题的方法	依赖少量标注、弱标注、高噪声事件样本, 提高抽取方法对数据环境的适应能力
		基于图的标签传播算法	利用已知样例的邻接关系, 减少标记扩充中的语义偏移问题	
	弱标记学习方法	归纳学习算法	根据实体出现在特定事件的情景, 推断共现实体和实体类别	
		迁移学习算法	将长文本中的学习结果扩展迁移至短文本要素识别中	
		多示例多标记学习算法	融合知识库中的实体关系进行事件要素识别	

3.3 事件融合技术

事件抽取实现了获取结构化事件数据的目标, 然而从异源信息中抽取的事件数据可能包含大量冗余、冲突甚至错误信息, 同时事件数据间的关系是扁平化的, 不能反映出部分事件涌现时现实空间中天然的相关性, 因此需要事件融合技术(Event Fusion)实现以下功能: 一是提高事件库的数据质量, 包括降低事件数据的冗余度和稀疏度, 提高准确率, 保证实时性等; 二是增强事件库所能表达的语义信息, 包括事件之间现实的关联和演化关系, 主题事件和

元事件的层次性, 不同时空粒度空间中组织事件数据的适应能力等。针对上述功能, 本文介绍事件合并技术和事件关联技术。

3.3.1 事件合并

事件合并技术又称为事件共指融合, 即将指代同个现实事件的多条事件数据合并为一条事件数据。起初事件合并主要通过规则方法实现, 随着事件内容逐渐复杂, 主要通过解决如下任务实现: 1)发现共指事件, 通过识别事件间共指关系, 实现同类事件数据聚合, 与传统相似性聚类不同, 相似事件并不等价于共指事件, 一些信息互补的事件数据在合并中发挥更大作用, 因此共指关系发现是实现事件融合的基础; 2)要素对齐, 在合并具有共指关系的事件数据时保留最有价值的事件要素信息, 补充缺失要素内容, 将要素值以规范格式输出, 从而提高事件数据质量, 并丰富事件数据的表示维度。

事件共指关系的发现方法可分为基于描述逻辑和基于统计学习两类方法。基于描述逻辑的共指关系发现主要利用事件的要素结构约束或特殊的事件性质约束, 满足约束条件的事件即划分为共指事件。例如, Schrodt 等[18]在 KEDS 政治事件库中的共指发现中定义了优先度规则, 首先检测是否存在来源的引用嵌套, 再基于事件要素的重合度判断事件共指。Lu 等[119]在规则方法基础上设计了针对多类表示事件共指特征的规则路径筛选器, 提高检测效率和准确率。Zheng 等[120]在约束条件中考虑了事件的情感极性、时态、情态和普遍性四个事件性质, 提出共指事件在事件性质一致性上的约束。基于统计学习的方法主要根据事件结构特征和上下文文本特征, 采用合理的聚合过程对共指事件进行聚类。例如, Bejan 等[117]提出在事件聚合过程中考虑到事件要素会随时间发生漂移, 在聚类算法中加入狄利克雷过程, 实现在事件融合中自动选择信息增量最大的特征, 针对事件数据流的动态特性实现聚合。Yang 等[121]考虑到同源和异源的共指事件在聚合过程中的差异, 采用了两层中餐馆过程, 实现共指事件聚类。

由于事件抽取方案日益复杂, 事件数据的差异逐渐增大, 混合的共指事件发现方法成为结构化数据融合的发展趋势。Chen 等[115]根据数据的结构特征, 通过比较树核相似性, 可获得事件数据局部结构的相似性, 实现共指事件聚类。Wang 等[116]将事件的上下文关联、事件数据的信息量以及语义相似性综合考虑, 提出核心相似度度量方法, 形成共指关系的度量方案。此外, 针对重要事件中爆发一系列子事件情况, 混合的共指发现方法将学习到的数据关系(重合、包含、互补和无关等)与合并规则相结合。例如 Araki 等[118]在生成恐怖袭击事件数据时, 首先发现逮捕、枪击、破坏等一系列相关的子事件, 再根据规则实现子事件中的关键信息融合(如图 6)。

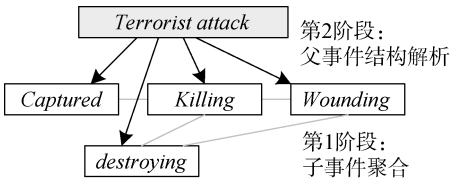

图 6 一类恐怖袭击事件数据合并方法

Figure 6 The event combination method for particular terrorist incidents

事件要素对齐主要针对要素信息的冗余、冲突和内容缺失等问题, 实现要素内容的规范输出。其中, 解决要素冗余问题主要通过同义词判别算法, 发现共指事件中同类要素。传统的同义词消解主要基于同义词词典, 随着词嵌入向量技术发展, 基于表示距离的同义词发现方法成为趋势, 例如 Qu 等[122]在结构化知识数据中利用属性类别共现频率和模式分析模块更新学习词嵌入向量, 进而通过词向量距离发现同义词。解决要素冲突问题主要通过合理的事件要素筛选机制, 例如最大频率优先法[56]、数据质量优先法等。文献[116]引入了事件数据的完整度和可信度两类评价指标, Naren 等[55]又提出衡量生成事件数据的时效性方法, 出现要素冲突时选取事件质量较高的字段作为要素标准取值。解决要素内容缺失问题主要通过基于已获取字段的要素推理方法, 例如事发地点要素取值为“北京”, 通过推理获得事件所属国别要素“中国”。常见的推理方法包括基于已获取要素的实体扩充方法[123]、基于本体语义网的查询方法[124]等。此外, 一些研究还探索与其他事件库构建技术相结合的要素对齐, 从而提升事件合并精度。例如, Lee 等[125]提出冗余要素与事件共指联合的迭代消解方法, 构建实体簇和事件簇所构成的团间双向信息反馈的聚类策略, 实现共指关系发现和实体要素对齐。Wu 等[126]则在事件抽取时直接构造要素的词嵌入向量, 在融合阶段根据语义距离实现事件同类实体要素的对齐。

本文提到的事件合并主要针对抽取后的结构化数据的融合问题, 表 6 对事件合并中的关键任务进行了总结。更广义的融合技术往往应用在事件库构建所涉及的各类数据中, 从事件检测阶段实现文档的相似内容发现, 到事件抽取中文本中的代词、同义词指代消解等技术, 再到实现共指事件合并技术,

出现了诸多以同质数据链接为目标的技术, 并形成MUC、B^3、CEAF、BLANC、CoNLL F1 等多类评价标准[127]。

表 6 事件合并技术总结

Table 6 The summary of event combination methods

主要任务	类别	代表技术与算法	特点
事件共指关系发现	基于描述逻辑的方法	要素重合度检查, 规则路径筛选, 事件性质一致性检查等	启发式经验方法, 适用于静态的相对完整的事件数据
	基于统计学习的方法	动态聚合方法, 分层聚合方法	融合文本特征, 适用于事件数据不断增加的动态场景
	混合方法	多因素混合的相似度度量方法,子事件融合方法等	融合更多共指事件特征, 面向复杂事件合并场景需求
事件要素对齐	面向冗余要素的方法	同义词关联方法, 联合迭代消解方法	提高事件要素表示的规范度
	面向冲突要素的方法	频率优先法, 数据质量优先法	提高可信度和准确度
	面向缺失要素的方法	实体扩充方法, 语义网查询方法	提高事件数据完整度

3.3.2 事件关联

现实世界中很多事件具有天然的逻辑关系, 常见的关系如下: a) 顺承关系, 如示威事件后发生冲突事件, 上海外滩新年活动到发生踩踏事件等, 通常具有明显的先后重叠和事件演化的规律; b) 并列关系, 通常为某个时空区间内连续发生的多起同类事件, 例如 2015 年巴黎系列恐怖袭击事件、“阿拉伯之春”连续爆发民主活动等, 通常这些事件具有相似的特征; c) 因果关系, 如灾难事件后发生救援事件、萨德入韩事件后国内抵制乐天等, 先发生的事件是后续事件的直接原因。在事件库中, 通过将一系列具有逻辑关系的事件数据关联融合, 可生成蕴含更丰富语义信息的事件图谱(Event Graph)[128], 并作为整体带入到特定主题事件相关研究中。事件数据的关联方式包括基于共享要素的关联和基于事件序列的关联两类, 分别对应以事件要素为节点和以元事件为节点的两类事件图谱表示方式。

基于共享要素的事件关联通常先采集特定区域或时间窗口内的事件数据, 然后以共享实体要素或关系要素为关联纽带, 生成以事件要素为节点的事件网络图谱, 如情节链、事件二部图等。常见的以事件要素为节点的事件图谱包括情节链、事件二部图等。其中, 情节链(Storyline)由具有时空关联的一系列事件数据根据其顺承关系串联生成, 事件要素作为节点按出现的时序关系排列, 事件要素间的边则表示对应的事件行为。Santos 等[129]根据事件数据分析了美国波士顿马拉松爆炸案中涉事实体的时空序列, 并组织为情节链, 如图 7a 内的 S_1、S_2、S_3; 通过多组相似的情节链可通过实体类别间的规则发现泛化表示为情节模板, 如模板 *P*; 还可通过不同地理区域内的情节链进行对比, 发现事件对邻近区域影响。文献[130]详细介绍了根据事件数据生成情节链的方法。事件二部图(Event Bipartite Graph)则可将事件表示为任意两类要素的多组关系的组合, 如恐怖组织和地点集合, 边代表某组织在某地区发动恐怖事件, 多个二部图组还可合成事件网[131], 每条穿过不同要素类别的完整路径代表一个事件, 如图 7b 中路径 1 和路径 2 均表示在纪念广场发生的事件。基于事件二部图的事件图谱表示方法在预测任务[55]、特定模式分析[132]、数据管理[133]等分析任务中应用广泛。

常见的以事件为节点的事件图谱包括序列图、影响网络等。事件序列图(Event Sequence Graph)将具有逻辑关联的事件按照时间序列进行串联, 从而可以反映事件的演化过程, 并常用于特定模式分析。事件序列通常由领域专家与事件系统交互生成, 近年来基于判别式或生成式模型的自动化事件序列生成方法研究成为热点。判别式方法以目标事件为关联依据, 学习模型对目标事件相关事件进行正确划分, 再根据时序顺序进行关联, 例如, Ning 等[134]采用多示例多目标分类方法, 生成了 2013 年阿根廷债务危机引发抗议游行的事件序列(如图 7c)。生成式方法假设主题事件演化的状态空间, 通过构造概率图模型推断事件空间所代表的事态变化, 并找到状态序列的代表性子事件组成事件序列, 例如 Qiao 等[135]基于 GDELT 结构化事件数据生成占领运动中的演化事件序列。此外, 事件的因果关系也可通过序列形式表示在因果图[136]中, 影响网络(Influence Net)就是一类反应因果关联强度的事件图谱, 它能将事件关联关系泛化为促进事件发生的正向影响和阻碍事件发生的负向影响, 可用于风险因素发现和事件转折点发现。影响网络可通过人工的启发式方法生成, 也可表示为贝叶斯信念网, 通过特定因果逻辑约束进行构造。例如, Haider 等[138]以 1999 年东帝汶危机中各事件的影响网络为例(如图 7d, “+”表示正影响, “-”表示负影响), 在 CAST[137](CAusal Strength)因果推断方法上提出了基于事件行为集合的边约减法, 通过近似比较边移除前后的变量联合概率变化, 获得影响方向和极性, 生成影响网络。

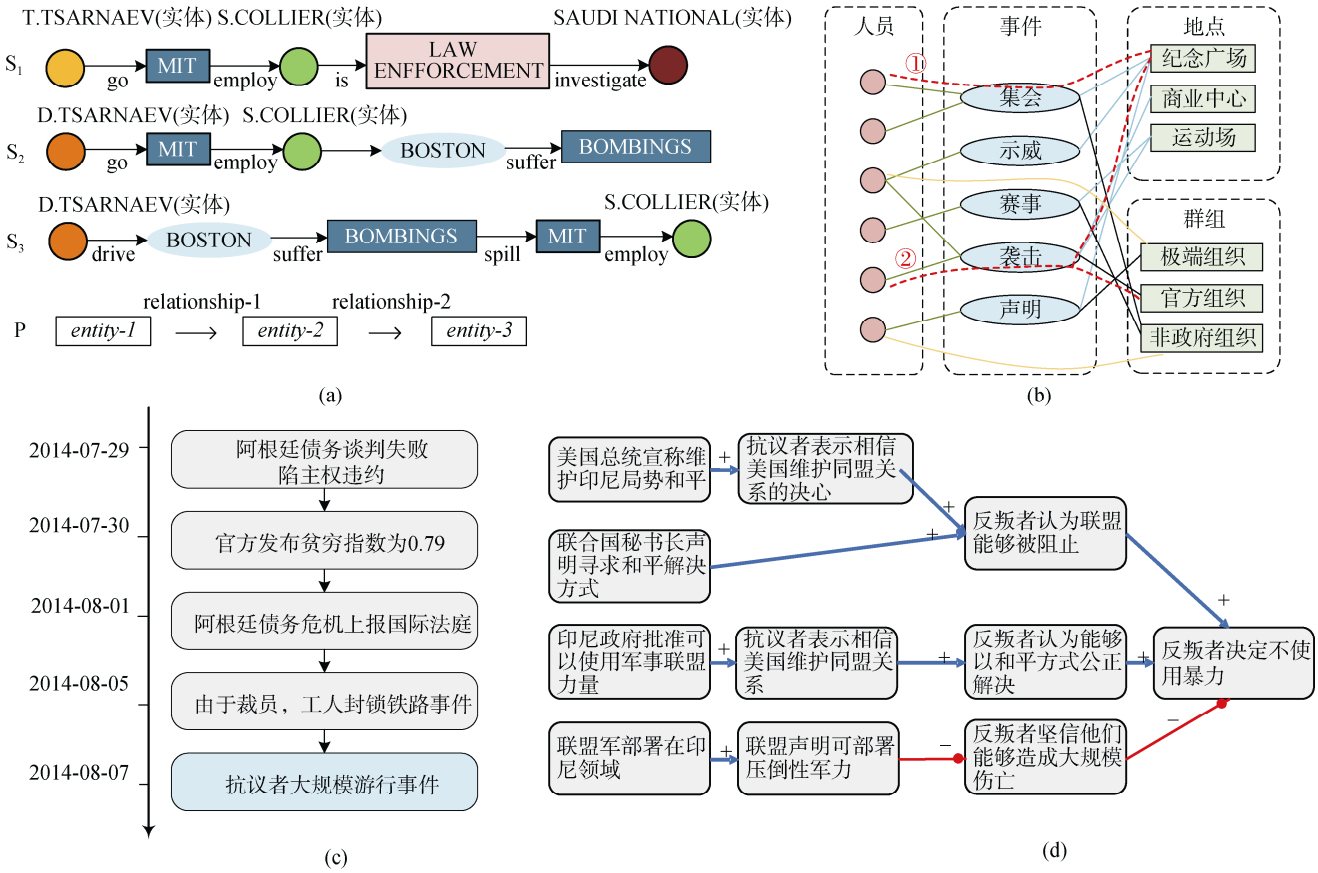

图 7　事件图谱表示方法(a 为情节链, b 为事件二部图, c 为序列图, d 为影响网络)

Figure 7　The representation methods of event graphs

(a. Storyline; b. Event Bipartite Graph; c. Event Sequence Graph; d. Influence Net)

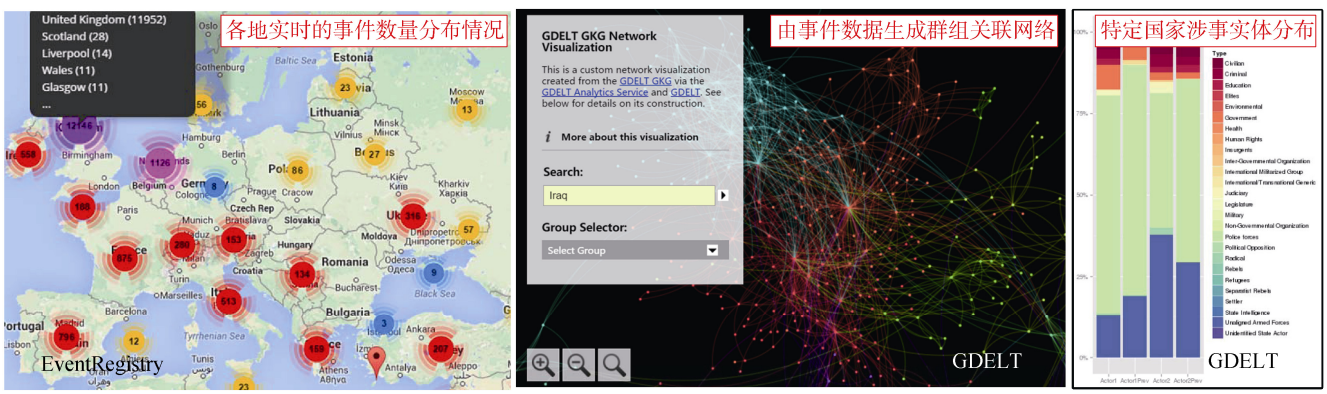

图 8　典型的感知系统截图

Figure 8　The typical event awareness systems

表 7 对比了两类事件数据关联方式所反映的事件间逻辑关系，关联融合生成的事件图谱可反映更丰富语义的数据视角。在大规模事件库中，事件图谱的生成与表示更加灵活，文献[139]展示了基于 GDELT 事件库生成多类事件图谱的实例，如针对特定类别事件的实体关系图、针对特定实体的事件序列图、针对特定时刻的事件图谱等，从而实现事件发展过程的可视化展示和深层挖掘。事件数据组合为事件图谱形式后，还可以用语义网形式进行表示，从而进一步提高数据的可扩展性。文献[140]实现将美国政治暴力事件库转化为语义网数据，并建立该领域的知识词表，同时进一步地实现了将英国、爱尔兰等国同类事件转化为语义网数据。

表 7 事件关联方式对比

Table 7 The contrast of event correlation methods

事件数据关联方式	典型事件图谱	可反映的事件间逻辑关系		
		顺承关系	并列关系	因果关系
基于共享要素的关联	情节链	√		
	二部图		√	
基于事件序列的关联	序列图	√	√	
	影响网络	√		√

4 事件库的应用

事件库作为同类现实事件的集中映射, 提供了具有时空特性的精细化信息集成的方式。事件库数字化表示了各个涉事实体状态变化和交互过程, 并应用于态势感知、事件推演、风险预警和响应等多个安全服务领域。

在态势感知方面, 基于事件库可实现量化分析和可视监控, 各国学者和国防机构纷纷将事件库作为态势分析的重要部分。文献[141]根据 UCDP 冲突事件库获得各国的冲突事件爆发趋势, 同时结合国家发展指标(如经济状况、人口数据等), 发现冲突事件对国家各领域发展的影响。Gao 等[142]根据 GDELT 新闻事件库分析了国家之间的互动强度, 并定量分析了国际安全及不稳定程度与政治互动强度的相关性。此外, 利用事件库中精确的时空信息, 可以灵活实现可视化态势感知系统(如图 8), 在地域监控和趋势感知方面提供更高层的观察视角, 仅 GDELT 上的相关项目就有上百个, 如抗议和冲突事件地图、各国领导人口碑变化趋势等。

在事件推演方面, 事件库蕴含了各类事件要素间的关联关系, 通过智能分析方法, 可以发现事件演化规律。美国军事分析学者根据 ISIS 恐怖组织的历史事件库, 通过概率时序逻辑的规则发现方法, 总结出事件演化中袭击方式、活跃地点等事件要素变化的因果关联[1]。英国国防专家根据 ICEWS 事件数据中涉事实体关联, 分析了反政府组织的作战网络演变规律, 并应用于泰国国内安全局势分析[143]。此外, 国防科技大学学者根据事件库中的事件时间、参与者、事件类型构建并挖掘酝酿和爆发阶段的频繁子图, 发现香港占中事件中的关键行为[144]。

在风险预警和响应方面, 一方面事件库构建中伴随事件发现过程, 能够提前发现风险事件及影响规模; 另一方面, 基于大量历史事件积累, 可以实现场景的动态仿真, 为应急决策提供支持。EMBERS 事件系统[55]通过包括五种事件预测方法的混合模型, 从社交媒体中找出可能发生的国内动乱事件并在事件库中发布。此外, 美国军方研发的虚拟战略预测工具 V-SAFT[137,138]通过不断抽取事件训练基于 Agent 的仿真模型, 模拟国家动荡局势中实体的变化过程, 从而实现罕见重大事件预测, 并能通过追踪策略部署后系统状态的变化, 理解和评估决策的有效性。

除了上述领域, 事件库作为一种精细的知识数据资源, 在情境搜索、深度问答、舆情分析等技术中蕴含巨大潜力。情境搜索关注搜索行为的时间、地点、背景等要素, 通过关联事件库可以理解搜索目标的意图, 例如用户搜索“ISIS”, 能够给出关键事件、组织架构、同盟组织、局势概况等一系列内容。深度问答需要大量结构化知识数据建立问题到查询, 查询到答案的语义映射关系, 当用户询问“某地区发生的最严重的恐怖袭击事件”时, 根据事件库建立查询语句(地区名称, 伤亡人数, 财产损失, 持续时间), 根据事件数据还原生成事件摘要给出答案。此外, 事件数据为网络舆情分析提供了清晰的焦点, 对于形成恶性社会影响力的事件可综合历史事件评判分析, 提出更合理的监管方案。

5 讨论与挑战

事件库是一类具有现实语义的精细化数据, 事件库的质量直接决定上述应用中的实现效果。事件库的数据质量不仅与构建技术的选取有关, 还与数据源选取、要表达的要素内容、领域知识的应用、具体实现的灵活性等因素密切相关, 本节对这些问题和挑战逐一展开探讨; 此外, 还对事件库的质量评价方法以及其它场景下的事件检测技术进行了简要介绍。

1) 事件数据源的选择。事件库构建过程中会维护一张巨大的数据源列表, 用于获取足够的数据量。然而, 集成的数据源越多, 同时意味着带来数据处理和分析能力的挑战; 另外, 不同数据源的数据质量参差不齐, 重大事件发生后通常会出现大量重复报道, 文献[142]还发现敏感人物不重要的琐事反而吸引了大量新闻媒体; 特别是事件检测中, 研究[147]发现数据源中的事件内容具有长尾效应, 并且即使是同个新闻源的覆盖范围也具有不稳定性。文献[148]证明了数据源集成并不是越多越好, 数据源最优选择问题是 NP 完全问题, 因此学者们开始研究数据集成中的一些近似优化方法。针对事件库构建场景, Dong 等[149]提出了从数据源覆盖率、新鲜度和准确度三个方面动态衡量数据源质量的方法。EventRegistry 事件库则选择了可交互的数据源分类选择系统

SourceSight[150], 根据事件类型动态选择合适的数据源。这些方法提供了数据源选择和集成的优化方法, 但在事件库构建中数据源的选择仍然是影响事件库性能的重要因素。

2) 事件情景要素的挖掘。通常事件抽取主要用于获得事件候选文本中包含的要素信息。目前越来越多的事件库开始探索在事件数据中加入某些潜在的情景要素信息, 例如事件造成的民众情感波动、事件影响的时空范围等。Joseph 等[151]提出了利用影响控制模型, 根据事件中的实体和行为推断出该事件的 EPA 情感极性, 即评价(evaluation, 事件的好坏)、效能(potency, 实体的情感强度)和动力(activity, 实体的情感活跃度)。Raimundo 等[129]利用事件实体之间和事件主题之间的关联推断事件的地理空间的影响范围。情景要素信息为实际应用提供了更丰富的素材, 但需要依赖整个事件的数据环境进行挖掘; 同时在选取情景要素时, 对于有争议的要素, 例如民众游行事件的好坏定性、引起的经济波动等, 应避免加入到事件数据, 影响数据的现实客观性。

3) 领域知识的应用与更新。在事件库构建的各个环节中均依赖领域知识的积累。由于事件涉及的要素类别众多, 除了常用的文本标注字典或语料库外, 每个要素类别均需要特定的领域知识或推断规则。例如通过时间数据知识库将文本中的“昨天”“3 天后”结合发布日期推断出准确日期, 地理信息能通过地名库中的经纬度进行推断。由于知识在现实世界中并非一成不变, 知识图谱技术的发展能提供更丰富和新鲜的知识数据, 从百科类、签到类、点评类等互联网应用中的数据提取出的相关知识数据亦可以用于事件库构建。Liu 等[152]整理了突发危急事件中常用的知识图谱, 实现了常见的危机事件的分类和涉及实体。此外, 事件库的富语义、结构化等特性也可以用于知识更新, 诸多学者基于结构化数据展开知识更新方法的研究, 例如 Trivedi 等[153]提出了基于时序特征的知识演进模型, 从而实现知识更新。在大规模事件库的构建过程中, 知识更新方法仍处在探索阶段。

4) 事件库设计的灵活性。事件数据是事件库的直接输出, 目标数据的表示粒度是影响事件库灵活性的关键因素。要素结构过于简单的事件库会出现大量相同的事件数据, 过于精细会造成内容稀疏, 甚至出现错误。因此, 事件库构建过程中应考虑该类事件的发生密度, 并权衡数据粒度和数据偏差, 定期进行稀疏性检测, 指导构建技术的调整。除此之外, 事件库构建中的各项关键技术在落实到业务模块流水线时, 还融合了人机交互接口, 在模型中融合人的智慧, 例如文献[56,146]中的事件系统提供数据源范围、时间段、事件热度、特定地点等设置接口, 从而提高了构建事件库的灵活性。

5) 事件库的质量评价。由于事件数据与现实事件对应, 事件数据的真实值只与现实事件有关, 即每个事件要素的真实值与信息源的记录差异无关。因此, 衡量事件库的数据质量优劣需要标准事件集。EMBERS 事件系统[55]通过与人工事件库 GSR 对比, 建立综合时间、地点、要素差异个数等因素的事件库评价方法。然而, 事件库领域类别差异较大, 很难为每个事件库都构造标准事件库。尽管如此, 有学者通过比较同类事件库的方式, 间接地评价事件库[155], 在使用中提供了一定参考价值。事件库质量与选取的构建方法直接相关, 相关技术的有效性可以分别进行评价。对于事件各类要素抽取的准确率, TREC 等评测机构提供了相关测试集进行评估。还可在各类模型方法中通过参数调节, 找到最优模型, 例如在聚类算法中调节分离出新类的参数条件, 异常事件模式定义中对比不同阈值定义的表现、事件抽取中文本单元与要素单元的相似度上界等。通过优化事件库构建中的各个关键模块, 提升事件库的数据质量。

6) 其他场景中的事件检测和抽取。除了从文本中获取事件信息外, 传感器、图像、视频等数据为事件分析提供了更丰富的素材。基于交通流量或用户 GPS 信息可以获得人群聚集区域, 结合社交网络内容分析, 可预判聚众示威[129]或突发踩踏事件[156]。基于新闻图像分析, 可获得事件中嫌疑人特征及环境特征信息, GDELT 事件库从 2016 年 2 月起增加新闻图片数据库, 形成综合的事件数据平台。此外, 视频中的事件检测近年来发展迅速, TRECVID[157]自 2010 年起将多媒体事件检测和事件监控列入事件检测关键任务并公布事件样本集, 根据特定行为的样本视频判断视频流中的特殊行为、人员信息等, 还可以提取事件的关键帧作为证据采样[158]。事件库在未来发展中逐渐融合上述多模态的事件信息, 为理解现实世界提供更全面的视角。

6 结束语

社会安全事件分析是内容安全的重要组成部分, 通过构建事件库可将新闻媒体、社交媒体、开放知识库等数据生成细粒度、高可用性的结构化事件数据, 实现对现实世界的数据映射。在未来发展中, 事件数据的获取能力就代表感知全球变化的能力。本

文从事件数据的内涵出发, 对事件库构建关键技术的研究和发展现状进行了全面分析和总结, 并对事件库的实际应用以及构建技术中的挑战和关键性问题特别展开探讨。

结构化事件库的构建技术, 一方面有利于形式化描述和解释突发安全事件, 揭示事件的内在机理和规律, 预测类似事件的发展趋势; 另一方面有利于全面量化评估国家或地区的政治安全风险, 以及预测国家或地区的安全发展态势。随着自动化构建技术的发展, 事件库在社会计算和国家安全领域发挥关键作用。

参考文献

[1] A Stanton, A Thart, A Jain, P Vyas, A Chatterjee and P Shakarian. “Mining for Causal Relationships: A Data-Driven Study of the Islamic State” *in Proc. ACM Conf. on Knowledge Discovery and Data Mining.(SIGKDD’15)*, pp. 2137-2146. 2015.

[2] S Muthiah, P Butler, R.P. Khandpur and P. Saraf, “EMBERS at 4 years: Experiences operating an Open Source Indicators Forecasting System” in *Proc. ACM Conf. on Knowledge Discovery and Data Mining.(SIGKDD’16)*, pp.205-214, 2016.

[3] P.A.Schrodt and D.J.Gerner. “Analyzing International event data: a handbook of computer-based techniques,” University of Kansas, 2000.

[4] R Grishman and B Sundheim. “Message understanding conference-6: A brief history,” in *Proc. of the 16th conference on Computational linguistics(COLING)*, vol.1, pp.466-471, 1996.

[5] D.J.Gerner, P.A.Schrodt and R.A.Francisco, “Machine coding of event data using regional and international sources,” *International Studies Quarterly*, vol.38, no.1, pp: 91-119, 1994.

[6] “ACE Annotation Tasks and Specifications,” https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications, 2008.

[7] Mathieu J, Fulk M, Lorber M, et al. “Social radar workflows, dashboards, and environments”. *MITRE CORP BEDFORD*, 2012.

[8] “OpenMIND” http://www.3i-mind.com/wp-content/uploads/2015/02/3iM-Bro-OpenMIND-LEA-09Feb2015.pdf, Feb.2015.

[9] “TAC KBP” https://tac.nist.gov/2014/KBP/Event/index.html,2014.

[10] “语言云(语言技术平台云),” http://www.ltp-cloud.com/, 2015.

[11] H.X Diao, G Xu, J Xiao. “An improved new event detection model,” *Information and Automation. Communications in Computer and Information Science*, vol 86, pp.431-437, 2011.

[12] H Zhang, G Li and X Xu, “Modeling association of news events on term network,” *Journal of National University of Defense Technology*, vol.36, no.4, pp.169-176, 2014.

[13] Y.N. Li, Y Tao and Wang J N. “A New Online New Event Detection Algorithm Based on Event Merging and Event Splitting,” *Applied Mechanics and Materials*, vol.513, pp.2024-2030, 2014.

[14] Q. Gao and H.L. You, “Research on Event Extraction,” *Information Studies:Theory and Application*, vol.36, no.4, pp.114-117 (in Chinese),2013
(高强, 游宏梁. 事件抽取技术研究综述. 情报理论与实践, 2013, 36(4): 114-117.)

[15] F Hogenboom, F Frasincar, U Kaymak and F.D. Jong, “A survey of event extraction methods from text for decision support systems”. *Decision Support Systems*, vol.85, pp.12-22, 2016.

[16] P. Over, G. Awad, M. Michel and J. Fiscus, “Trecvid 2012–an ´overview of the goals, tasks, data, evaluation mechanisms and metrics,” in *Proc TRECVID*, 2012.

[17] J. Kalita, "Detecting and Extracting Events from Text Documents." arXiv preprint arXiv:1601.04012, 2016.

[18] P.A Schrodt and J Yonamine. “Automated coding of very large scale political event data.” in *New directions in text as data workshop*, Harvard. 2012.

[19] J Pustejovsky, J.M.Castano and R Ingri, “TimeML: Robust specification of event and temporal expressions in text,” *New directions in question answering*, vol.3,no.3, pp.28-34, 2003.

[20] Ahn D, “The stages of event extraction,” in *Proceedings of the Workshop on Annotating and Reasoning about Time and Events. Association for Computational Linguistics*, pp.1-8, 2006.

[21] “COW,” http://www.correlatesofwar.org/data-sets, May. 2017.

[22] “CEDS” http://eventdata.parusanalytics.com/, 2008.

[23] “Wikipedia Worldwide current events,” https://en.wikipedia.org/wiki/Portal:Current_events, June.2017.

[24] “Uppsala conflict data sets(UCDP),” http://ucdp.uu.se/, June.2017.

[25] “GTD,” https://www.start.umd.edu/gtd/, June.2017.

[26] “ACLED,” http://www.acleddata.com/, July.2017.

[27] “SPEED” http://www.clinecenter.illinois.edu/data/speed/, 2012.

[28] “ICEWS” http://www.lockheedmartin.com/us/products/W-ICEWS.html, June.2017.

[29] “GDELT,” http://www.gdeltproject.org/, July.2017.

[30] “SCAD,” https://www.strausscenter.org/scad.html, June.2017.

[31] “NewsReader,” http://www.newsreader-project.eu, June.2017.

[32] “PublicSonar,” http://publicsonar.com/solutions, June. 2017.

[33] “Event Registry,” http://eventregistry.org/, July.2017.

[34] “EMBERS,” http://dac.cs.vt.edu/research-project/embers/, 2016.

[35] J. Allan, J. Carbonell, G. Doddington and J. Yamron. “Topic detection and tracking pilot study,” In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[36] Linguistic Data Consortium, “ACE (Automatic Content Extraction) English Annotation Guidelines for Events,” version 5.4.3 2005.07.01 edition. 2005.

[37] Gerner, Deborah J, “Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions.” *International Studies Association, New Orleans*, 2002.

[38] “WEIS Event Codes,” http://eventdata.parusanalytics.com/data.dir/weis.html, 2015.

[39] P.A. Schrodt and D.J. Gerner, “Analyzing the dynamics of international mediation processes,” in *Annual summer meeting of the Political Methodology group*, pp.19-21, 2001.

[40] S Chakraborty. “Big Data Analytics for Development: Events, Knowledge Graphs and Predictive Models[Ph.D.dissertation],” New

York University, 2015.

[41] "PETRARCH2", http://petrarch.readthedocs.io/en/latest/#, Dec 2016.

[42] H Li. "Moving from news to social media: Unsupervised knowledge enrichment for event extraction[Ph.D.dissertation]," Rensselaer Polytechnic Institute, 2015.

[43] T. Snowsill, F. Nicart, M. Stefani, T. De Bie, and N. Cristianini. Finding surprising patterns in textual data streams. In *Cognitive Information Processing, 2nd International Workshop*, pp.405–410, 2010.

[44] P A Schrodt, "Cameo: Conflict and mediation event observations event and actor codebook". Pennsylvania State University, 2012.

[45] K Joseph, K M Carley and D Filonuk, "Arab Spring: from newspaper," *Social Network Analysis and Mining*, vol.4, no.1, pp.1-17, 2014.

[46] G Ifrim, B Shi and I Brigadir, "Event detection in twitter using aggressive filtering and hierarchical tweet clustering" in *Second Workshop on Social News on the Web* (SNOW'14), April 2014.

[47] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry:learning about world events from news," in *Proc. of ACM conf. on World wide web companion(WWW'14)*, pp.107–110, 2014.

[48] C.C Aggarwal and P.S Yu. "On clustering massive text and categorical data streams,". *Knowledge and information systems*, vol.24, no.2, pp. 171-196, 2010.

[49] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Computational Intelligence*, vol. 31, no.1, pp. 132–164, 2015.

[50] W. Cui, S. Liu and L. Tan, "TextFlow: towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol.17, no.12, pp.2412–2421, 2011.

[51] A. Ahmed, Q. Ho and J. Eisenstein, "Unified analysis of streaming news," in Proceedings of the 20th international conference on World wide web(WWW'11), pp.267-276, 2011.

[52] Y. Wang, E. Agichtein and M. Benzi. "TM-LDA: efficient online modeling of latent topic transitions in social media," in *Proc. ACM conf. on Knowledge discovery and data mining (KDD'12)*, pp.123-131, 2012.

[53] W. Wei, K. Joseph, W. Lo and K.M. Carley. "A Bayesian Graphical Model to Discover Latent Events from Twitter," in *Ninth International AAAI Conference on Web and Social Media(AAAI'15)*, pp.503-512, 2015.

[54] J. Wang, M. Bansal, K. Gimpel, B. Ziebart, and T. Clement, "A sense-topic model for word sense induction with unsupervised data enrichment," *Transactions of the Association for Computational Linguistics*, vol.3, pp.59–71, 2015.

[55] N. Ramakrishnan, P Butler and S Muthiah. "'Beating the news' with EMBERS: forecasting civil unrest using open source indicators." In *Proc. ACM conf. on KDD*, pp.1799-1808. 2014.

[56] F. Abel, C. Hauff and G.J. Houben, "Semantics+ filtering+ search= twitcident. exploring information in social web streams," *ACM conf. on Hypertext and social media(HT'12)*, pp.285-294, 2012.

[57] D.M. Blei, T.L. Griffiths and M.I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol.57, no.2, pp.7-39, 2010.

[58] V. Zavarella, H. Tanev and R. Steinberger, "An Ontology-Based Approach to Social Media Mining for Crisis Management," in *SSA-SMILE ESWC*, pp.55-66, 2014.

[59] A Najgebauer, R Antkiewicz and M Chmielewski, "The prediction of terrorist threat on the basis of semantic association acquisition and complex network evolution," *Journal of Telecommunications and Information Technology*, pp.14-20, 2008.

[60] W Huang, T Wang and W Chen, "Category-Level Transfer Learning from Knowledge Base to Microblog Stream for Accurate Event Detection," in *22nd International Conference on Database Systems for Advanced Applications (DASFAA'17)*, pp. 50-67, 2017.

[61] E. Andrade, S. Blunsden, and R. Fisher. "Modelling Crowd Scenes for Event Detection." In *Proc. of International Conference on Pattern Recognition (ICPR'06)*, pp.175–178, 2006.

[62] S Amer-Yahia, S Anjum and A Ghenai, "MAQSA: a system for social analytics on news," in *Proc. of ACM Conf. on Management of Data (SIGMOD'12)*, pp.653-656, 2012.

[63] "New GDELT 2.0 API Interactive Maps," http://blog.gdeltproject.org/new-gdelt-2-0-api-interactive-maps-adm1/, 2017.

[64] N.F Johnson, M Zheng and Y Vorobyeva, "New online ecology of adversarial aggregates: ISIS and beyond". *Science*, vol.352, no.6292, pp.1459-1463, 2016.

[65] A. Hasan, K. Teymourian and A. Paschke, "Probabilistic Event Pattern Discovery," in *annual International Web Rule Symposium (RuleML'15)*. pp.241-257, 2015.

[66] B. Cule, L. Feremans and B. Goethals, "Efficient Discovery of Sets of Co-occurring Items in Event Sequences," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases(ECML-PKDD'16)*, pp.361-377, 2016.

[67] F. Liang, S Ma and J.L. Hellerstein, "Discovering fully dependent patterns," in *Proc. of the 2002 International Conference on Data Mining (SDM'02)*, pp.511-527, 2002.

[68] S Ma and J.L. Hellerstein, "Mining mutually dependent patterns," in *Proc. IEEE Conf. on Data Mining (ICDM)*. pp.409-416, 2001.

[69] J.D. Souza and V. Ng, "Classifying Temporal Relations with Rich Linguistic Knowledge." in *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT)*, pp.918–927, 2013.

[70] C. Zhou, B. Cule and B. Goethals, "A pattern based predictor for event streams," *Expert Systems with Applications*, vol.42, no.23, pp.9294-9306, 2015.

[71] P. Mirza, "Extracting Temporal and Causal Relations between Events," in *ACL Student Research Workshop*, pp 10-17.2014.

[72] J.E. Yonamine, "Predicting Future Levels of Violence in Afghanistan Districts Using GDELT," *GDELT Project Manuscript*, 2013.

[73] S Chakraborty, A Venkataraman and S. Jagabathula, "Predicting Socio-Economic Indicators using News Events," in *Proc. ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD'16)*, pp.1455-1464, 2016.

[74] D Preotiuc-Pietro and T. Cohn, "A temporal model of text periodicities using Gaussian Processes," *Int'l Conf. Empirical*

Methods on Natural Language Processing (EMNLP'13). pp.977-988, 2013.

[75] Q He, K Chang and E.P. Lim, "Bursty feature representation for clustering text streams" in *Proc. of the 2007 SIAM International Conference on Data Mining (SDM'07)*, pp.491-496, 2007.

[76] J. Zhang, Y. Zheng and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," in *31th AAAI Conf. on Artificial Intelligence (AAAI'17)*. pp.1655-1661, 2017.

[77] C Zhang, G Zhou and Q Yuan, "Geoburst: Real-time local event detection in geo-tagged tweet streams" in *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR'16),* pp.513-522, 2016.

[78] Y. Matsubara and Y. Sakurai, "Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences," in *Proc. ACM Knowledge Discovery and Data Mining (KDD'16)*, pp.1045-1054, 2016.

[79] H. Chen and W. Chen, "Analyzing bursty feature for event detection," *Application Research of Computers*, vol.28, no.1, pp.117-120 (in Chinese), 2011.
(陈宏, 陈伟. 基于突发特征分析的事件检测. 计算机应用研究, 2011, 28(1):117-120.)

[80] "NewsAnalysis," http://cs.nyu.edu/~sunandan/v2.0/news.html.

[81] "Apache UIMA Ruta (Rule-based Text Annotation)," https://uima.apache.org/ruta.html, Apr. 2017.

[82] M. Jiang, J. Shang, T. Cassidy, X. Ren, L.M. Kaplan, T.P. Hanratty and J. Han, MetaPAD: Meta Pattern Discovery from Massive Text Corpora. in *Proc. ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD'17)*, pp.877-886, 2017.

[83] D Martens, B.B Baesens and T.V. Gestel, "Decompositional rule extraction from support vector machines by active learning", *IEEE Transactions on Knowledge and Data Engineering*, vol.21, no.2, pp.178-191, 2009.

[84] P.P Ruiz, B.K Foguem and B. Grabot, "Generating knowledge in maintenance from Experience Feedback," *Knowledge-Based Systems*, vol.68, pp.4-20, 2014.

[85] C L.Forgy "Rete: A fast algorithm for the many pattern/many object pattern match problem," *Artificial intelligence*, vol.19, no.1, pp.17-37, 1982.

[86] J.R. Quinlan, "Learning logical definitions from relations," *Machine learning*, vol.5, no.3, pp.239-266, 1990.

[87] C. Aone and M. Ramos-Santacruz, "REES: a large-scale relation and event extraction system" in *Proc. of ACL Conf. Applied Natural Language Processing (ANLP)*, pp.76-83, 2000.

[88] C.S. Lee, Y.J. Chen and Z.W. Jian, "Ontology-based fuzzy event extraction agent for Chinese e-news summarization," *Expert Systems with Applications*, vol.25, no.3, pp.431-447,2003.

[89] P.F. Li, G.D. Zhou and Q.M. Zhu, "Semantics-Based Joint Model of Chinese Event Trigger Extraction," *Journal of Software*, vol.27, no.2, pp.280-294, (in Chinese), 2016.
(李培峰, 周国栋, 朱巧明. 基于语义的中文事件触发词抽取联合模型[J]. 软件学报, 2016, 27(2):280-294.)

[90] A. McCallum, D. Freitag and F.C.N. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," in *Int'l Conf. on Machine Learning(ICML'00)*. pp.591-598, 2000.

[91] J. Piskorski , H. Tanev and P.O. Wennerberg, "Extracting violent events from on-line news for ontology population," *10th International Conference on Business Information Systems(ICBIS'07)*. pp.287-300, 2007.

[92] A. Sun and R. Grishman, "Cross-domain bootstrapping for named entity recognition," in *Proc. ACM SIGIR 2011 Workshop on Entity-Oriented Search*, pp.33-40, 2011.

[93] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. ACM conf. on Research & Development in Information Retrieval(SIGIR'14)*, pp.43-52, 2014.

[94] X.X. Chen and B. Liu, "Extracting Open Domain Events in Microblogs", *Computer Applications and Software*, vol.33, no.8, pp.18-22 (in Chinese), 2016.
(陈箫箫, 刘波. 微博中的开放域事件抽取. 计算机应用与软件, 2016, 33(8):18-22.)

[95] Q. Li, H. Ji and L. Huang, "Joint Event Extraction via Structured Prediction with Global Features," *Association for Computational Linguistics*, no.1, pp.73-82, 2013.

[96] Y. Gao, Z.Y. Xi and B.C. Li, "Argument Extraction Algorithm Based on Convolution Tree Kernel", *Journal of Chinese Computer Systems*, vol.37, no.4, pp.722-725 (in Chinese), 2016.
(高源, 席耀一, 李弼程, 等. 基于卷积树核的事件论元角色抽取方法[J]. 小型微型计算机系统, 2016, 37(4): 722-725.)

[97] Y. Zhang, Z. Liu, and W. Zhou, "Event recognition based on deep learning in Chinese texts," *PLoS ONE*, vol.11, no.8, pp.1–18, 2016.

[98] G.C. Wu, "Application and Research of HMM Incremental Learning Algorithm in Chinese Named Entity Recognition [Master dissertation]," South China University of Technology (in Chinese), 2011.
(吴广财. HMM 增量学习算法在中文命名实体识别中的应用研究[硕士学位论文],华南理工大学, 2011.)

[99] J. Shin, S. Wu and F. Wang, "Incremental knowledge base construction using deepdive," in *Proceedings of the VLDB Endowment*, vol.8, no.11, pp.1310-1321, 2015.

[100] L Zheng, P Jin and J Zhao, "A fine-grained approach for extracting events on microblogs," in *Int'l Conf. Database and Expert Systems Applications (DESA'14)*, pp.275-283, 2014.

[101] M. Liu, Y. Liu, L Xiang, X. Chen and Q Yang, "Extracting Key Entities and Significant Events from Online Daily News," in *9th Int'l Conf. Intelligent Data Engineering and Automated Learning (IDEAL'08)*, vol.5326, pp.201-209, 2008.

[102] H. Ji and R. Grishman, "Refining Event Extraction through Cross-Document Inference," *ACL*, pp.254-262, 2008.

[103] C. Zhang, S. Soderland and D.S. Weld, "Exploiting parallel news streams for unsupervised event extraction," *Transactions of the Association for Computational Linguistics*, vol.3, pp.117-129, 2015.

[104] R. Yangarber, "Counter-training in discovery of semantic patterns," in *Proc. Annual Meeting on Association for Computational Linguistics(ACL'03)*, vol.1, pp.343-350, 2003.

[105] A. Carlson, J. Betteridge and R.C. Wang, "Coupled semi-supervised learning for information extraction," in *Proc. ACM Int'l Conf. on Web Search and Data Mining. (WSDM'10)*, pp.101-110, 2010.

[106] X. Zhu, Z. Ghahramani, and J. Lafferty. “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proc. Int’l Conf. on Machine learning (ICML’03)*, pp.912-919, .2003.

[107] Y Hong, J Zhang and B Ma. “Using cross-entity inference to improve event extraction,” in *Proc. Annual Meeting of the ACL Human Language Technologies(HLT’11)*, vol.1, pp. 1127-1136, 2011.

[108] J.W. Liu, Y. Liu and X.L. Luo, “Semi-Supervised Learning Methods”, *Chinese Journal of Computers*, vol.38, no.8, pp.1592-1617 (in Chinese), 2015.
(刘建伟, 刘媛, 罗雄麟. 半监督学习方法. 计算机学报, 2015, 38(8): 1592-1617.)

[109] F.M Wei, J.P. Zhang, Y. Chu and J. Yang, “FSFP: Transfer learning from long texts to the short,” *Applied Mathematics & Information Sciences*, vol.8, no.4, pp.2033-2044, 2014.

[110] G Angeli, J Tibshirani and J Wu, “Combining Distant and Partial Supervision for Relation Extraction,” in *Proc. Int’l Conf. Empirical Methods on Natural Language Processing (EMNLP’13)*, pp.1556-1567, 2014.

[111] X. Ling and D.S. Weld, “Fine-grained entity recognition,” in *Proc. 26th AAAI Conference on Artificial Intelligence (AAAI’12)*, pp.94-100, 2012.

[112] R. Navigli and M. Lapata. “An experimental study of graph connectivity for unsupervised word sense disambiguation,” *IEEE transactions on pattern analysis and machine intelligence*, vol.32, no.4, pp.678-692, 2010.

[113] Z.H. Wu, C. Liang, and C. L. Giles, “Storybase: Towards building a knowledge base for news events,” in *Proc. Int’l Joint Conference of Natural Language Process(IJCNLP’15)*, pp.133-138, 2015.

[114] H. Li, H. Ji and L. Zhao, “Social event extraction: Task, challenges and techniques” in *Advances in Social Networks Analysis and Mining (ASONAM’15)*, pp.526-532, 2015.

[115] B. Chen, J. Su and S.J. Pan, “A Unified Event Coreference Resolution by Integrating Multiple Resolvers,” *Proc. Int’l Joint Conference of Natural Language Process(IJCNLP’11)*, pp.102-110, 2011.

[116] Wang R. Information-based Event Coreference[Ph.D.dissertation]. University of Illinois, 2015.

[117] C.A.Bejan and S.Harabagiu, “Unsupervised event coreference resolution,” *Computational Linguistics*, vol.40,no.2,pp.311-347, 2014.

[118] J. Araki, Z.Z. Liu and E. Hovy. “Detecting subevent structure for event coreference resolution,” *Proc. Int’l Conf. on Language Resources and Evaluation(LREC’14)*, pp.4553-4558, 2014.

[119] J. Lu and Ng V. Event Coreference Resolution with Multi-Pass Sieves, *International Conference on Language Resources and Evaluation(LREC)*, 2016.

[120] Z Chen, H Ji and R. Haralick, “A pairwise event coreference model, feature impact and evaluation for event coreference resolution,” in *Proc. of the workshop on events in Emerging Text Types, Association for Computational Linguistics (eETT’09)*, pp.17-22, 2009.

[121] B. Yang, C. Cardie and P. Frazier. A hierarchical distance-dependent bayesian model for event coreference resolution, arXiv preprint arXiv:1504.05929, 2015.

[122] M. Qu, X. Ren and J. Han, Automatic Synonym Discovery with Knowledge Bases, in *Proc. ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD’17)*, pp.997-1005, 2017.

[123] J. Shen, Z. Wu, D. Lei, J. Shang, X. Ren and J. Han, SetExpan: Corpus-based Set Expansion via Context Feature Selection and Rank Ensemble”, in *Proc. of European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD’17)*, 2017.

[124] X. Rong, Z. Chen, and Q. Mei, Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion, in *Proc. of the Ninth ACM International Conference on Web Search and Data Mining (WSDM’16)*, pp.645-654, 2016.

[125] H. Lee, M. Recasens and A. Chang, “Joint entity and event coreference resolution across documents,”in *Proc. Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning(EMNLP-CoNLL)* pp.489-500, 2012.

[126] J.L. Wu and W.Y. Ma, “A Deep Learning Framework for Coreference Resolution Based on Convolutional Neural Network,” in *Proc. IEEE International Conference Semantic Computing (ICSC’17),* pp.61-64, 2017.

[127] JAE Hovy and T. Mitamura, “Evaluation for partial event coreference,” *Association for Computational Linguistics (ACL’14)*, pp. 68-76, 2014.

[128] G Glavaš and J. Šnajder, “Construction and evaluation of event graphs,” *Natural Language Engineering*, vol.21, no.04, pp. 607-652, 2015.

[129] R.F. Dos Santos, A. Boedihardjo and S. Shah, “The big data of violent events: algorithms for association analysis using spatio-temporal storytelling,” *GeoInformatica*, vol.20, no.4, pp.879-921, 2016.

[130] D. Kumar, N. Ramakrishnan and R.F. , Helm, “Algorithms for storytelling,” *IEEE Transactions on Knowledge and Data Engineering*, vol.20, no.6, pp.736-751, 2008.

[131] T.A.N. Pham, X Li and G Cong, “A general graph-based model for recommendation in event-based social networks,” in *Proc. IEEE Int’l Conf. Data Engineering (ICDE’15)*, pp.567-578, 2015.

[132] P. Wang, P. Pattison and G. Robins, “Exponential random graph model specifications for bipartite networks-A dependence hierarchy,” *Social networks*, vol.35, no.2, pp.211-222, 2013.

[133] W. Yang, R. Li and P. Li, “Event Related Document Retrieval Based on Bipartite Graph,” in *Int’l Conf. on Web-Age Information Management (WAIM’16)*. pp.467-478, 2016.

[134] Y. Ning, S. Muthiah and H. Rangwala. “Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning,” in *Proc. ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD’16),* pp.1095-1104, 2016.

[135] F. Qiao, P. Li, Zhang X, et al. “Predicting Social Unrest Events with Hidden Markov Models Using GDELT,” *Discrete Dynamics in Nature and Society*, 2017.

[136] O. Handel, D.H. Biedermann and P.M. Kielar, “A system dynamics based perspective to help to understand the managerial big picture in respect of urban event dynamics,” *Transportation Research Procedia*, vol.2, pp.669-674, 2014.

[137] J.A. Rosen and W.L. Smith, Influence Net Modeling with Causal Strengths: An Evolutionary Approach, in *Proc. of the Command and Control Research and Technology Symposium*, pp.25-28, 1996.

[138] S. Haider and S.A. Raza, "Complexity reduction of influence nets using arc removal," *Journal of Intelligent & Fuzzy Systems*, vol.28, no.4, pp.1849-1859, 2015.

[139] P. Delias and I. Kazanidis, "Process Analytics Through Event Databases: Potentials for Visualizations and Process Mining," *Int'l Conf. on Decision Support System Technology (ICDSST'17)*. pp.88-100, 2017.

[140] R. Brennan, K.C. Feeney and O. Gavin, "Publishing social sciences datasets as linked data: a political violence case study," *Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH'13)Workshop at 36th ACM SIGIR Conference,* 2013.

[141] M.D. Ward, N.W. Metternich and C.L. Dorff, "Learning from the past and stepping into the future: Toward a new generation of conflict prediction," *International Studies Review*, vol.15, no.4, pp.473-490, 2013.

[142] J Gao, K.H. Leetaru and J. Hu, "Massive media event data analysis to assess world-wide political conflict and instability," *Social Computing, Behavioral-Cultural Modeling and Prediction*, vol.7812, pp.284-292, 2013.

[143] N.W. Metternich, C. Dorff and M Gallop, "Antigovernment networks in civil conflicts: How network structures affect conflictual behavior," *American Journal of Political Science*, vol.57, no.4, pp.892-911, 2013.

[144] F Qiao, P Li and J Deng, "Graph-Based Method for Detecting Occupy Protest Events Using GDELT Dataset" in *Proc. IEEE Int'l Conf. Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC'15)*, pp.164-168, 2015.

[145] I.S. Lustick."Making Sense of Social Radar: V-saft as an Intelligent Machine." https://pdfs.semanticscholar.org/9a60/90518472ef34dd60de8a91170e036d1dc373.pdf, 2015.

[146] R J. González, "Seeing into hearts and minds: Part 1. The Pentagon's quest for a 'social radar'," *Anthropology Today*, vol.31, no.3, pp. 8-13, 2015.

[147] X.L. Dong and D. Srivastava, "Big data integration," *Synthesis Lectures on Data Management*, vol.7, no.1, pp.1-198, 2015.

[148] X.L. Dong, B. Saha and D. Srivastava, "Less is more: Selecting sources wisely for integration," in *Proceeding of the VLDB Endowment. (VLDB'12)*, vol.6, no.2, pp.37-48, 2012.

[149] T. Rekatsinas, X.L. Dong and D. Srivastava, "Characterizing and selecting fresh data sources," in *Proc. ACM Int'l Conf. on Management of data(SIGMOD'14)*, pp.919-930, 2014.

[150] T. Rekatsinas, D. Srivastava and X.L. Dong, "SOURCESIGHT: Enabling Effective Source Selection," in *Proc. ACM Int'l Conf. on Management of Data (ICMD'16)*, pp.2157-2160, 2016.

[151] K. Joseph, W. Wei and M. Benigni, "A social-event based approach to sentiment analysis of identities and behaviors in text," *Journal of Mathematical Sociology*, vol.40, no.3, pp. 137-166,2016.

[152] S. Liu, C. Brewster and D. Shaw, "Ontologies for crisis management: A review of state of the art in ontology design and usability," in *Proc. Int'l Conf. on Information Systems for Crisis Response and Management (ISCRAM'13)*. pp.349-359, 2013.

[153] R. Trivedi, M. Farajtabar and Y. Wang, "Know-Evolve: Deep Reasoning in Temporal Knowledge Graphs," arXiv preprint arXiv:1705.05742, 2017.

[154] A. Sheth, A. Jadhav and P. Kapanipathi, "Twitris: A system for collective social intelligence," *Encyclopedia of Social Network Analysis and Mining(ESNAM'14)*, pp.2240-2253, 2014.

[155] M.D. Ward, A. Beger and J. Cutler, "Comparing GDELT and ICEWS event data," *Analysis*, vol.21, pp.267-297, .2013.

[156] Y. Zheng, H. Zhang and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," in *Proc. ACM Int'l Conf. on Advances in Geographic Information Systems.(SIGSPATIAL'15),* pp.2-11, 2015.

[157] "Trecvid TREC Video Retrieval Evaluation (TRECVID)," http://trecvid.nist.gov/, 2017.

[158] C. Gan, N. Wang and Y. Yang, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR'15).* pp. 2568-2577, 2015.

[159] Q. Liu, Y. Li and H. Duan, "Knowledge Graph Construction Techniques", *Journal of Computer Research and Development*, vol.53, no.3, pp.582-600 (in Chinese), 2016.
(刘峤, 李杨, 段宏,等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53(3):582-600.)

[160] M.D. Ward, N.W Metternich and C.L. Dorff, "Learning from the past and stepping into the future: Toward a new generation of conflict prediction," *International Studies Review*, vol.15, no.4, pp. 473-490, 2013.

[161] C. Cioffi-Revilla, "Introduction to computational social science: principles and applications". Springer, vol.10, 2017.

薛聪 于 2012 年在电子科技大学信息安全专业获得学士学位。现在中国科学院大学网络空间安全专业攻读博士学位。研究领域为社会计算、事件分析。研究兴趣包括社会预测、事件抽取与模式挖掘、网络安全等。Email: xuecong@iie.ac.cn

高能 于 2006 年在中国科学院研究生院通信与信息系统专业获得博士学位。现为中国科学院信息工程研究所研究员, 现任信息安全国家重点实验室副主任。研究领域为网络安全、系统安全。研究兴趣包括: 安全事件分析、社交网络隐私。Email: gaoneng@iie.ac.cn

查达仁 于 2010 年在中国科学院研究生院信息安全专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络安全、密码工程。研究兴趣包括: 密码工程与应用, 网络体系结构与安全防护。Email: zhadaren@iie.ac.cn

王雷 于 2012 年在中国科学院研究生院信息安全专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络信息安全。研究兴趣包括: 数据智能处理、密码工程与应用、身份认证技术。Email: wanglei@iie.ac.cn

尹芷仪 于 2010 年在武汉大学计算机应用技术专业获得博士学位。现任中国科学院信息工程研究所客座硕士生导师。研究领域为社会计算、网络空间安全。Email: llinshi@163.com

曾泽华 于 2015 年在中国科学技术大学信息安全专业获得学士学位。现在中国科学院大学网络空间安全专业攻读博士学位。研究领域为事件分析、内容安全。研究兴趣包括: 社交网络中的事件分析。Email: zengzehua@is.ac.cn