doi:10.3772/j.issn.2095-915x.2020.04.011

应急科技情报服务平台关键技术研究

开放科学（资源服务）标识码（OSID）

王良熙[1, 2]

1. 福建省科学技术信息研究所 福州 350003；
2. 福建省信息网络重点实验室 福州 350003

摘要：本研究针对应急科技情报服务平台的特点，提出了应急平台的技术框架设计方案，探讨了平台关键技术，包括知识库、探索性数据分析、容器化技术、数据可视化、融媒体技术应用和数据安全等，重点就作为核心技术的知识库构建所涉及的多源异构数据采集、数据清洗、知识抽取、知识推理和推荐算法等进行较系统的阐述，提出相关技术的解决思路和方案，并介绍了平台在应急情报服务的应用案例，为大数据技术在应急平台的应用提供参考。

关键词：应急；响应；知识库；科技情报

中图分类号：TP311.56；G35

Research on Key Technologies of Emergency Science and Technology Information Service Platform

WANG Liangxi[1,2]

1. Fujian Institute of Scientific and Technological Information, Fuzhou 350003, China;
2. Key Laboratory of Information Network in Fujian Province, Fuzhou 350003, China

Abstract: Based on the characteristics of the emergency scientific and technological information service platform, this research proposes the technical framework design of the emergency platform, and discusses some key technologies of the platform, including knowledge base, exploratory data analysis, containerization technology, data visualization, fusion media technology application and data Security, etc., focusing on the multi-source heterogeneous data collection, data cleaning, knowledge extraction, knowledge reasoning and recommendation algorithm involved in the construction of the knowledge base as the core technology，and put forward relevant technical solutions, and the application case of the platform in the emergency information service is introduced, and it provides references for the application of big data technology in the emergency platform.

Keywords: Emergency; response; knowledge base; scientific and technological information

基金项目：中央引导地方科技发展专项“打造区域科技创新智库基础条件与能力建设”（2019L3015）；省属公益类科研院所基本科研专项“分布式科技数据采集与存储平台的研究与应用”（2019R1008-8）。

作者简介：王良熙（1967-），高级工程师，研究方向：信息系统、情报工程、知识挖掘。

引言

当前，大部分科技情报研究机构面向决策支持的课题研究，往往需要动辄数月乃至一年以上，甚至更久的时间。这种常规的应对方式无法满足突发事件发生时，政府机构和社会各界对应急情报服务的迫切性需求，使得科技情报机构在关键时刻难以有效发挥“耳目、尖兵、参谋”作用，无法凸显科技情报的应用价值。因此，科技情报机构有必要构建应急科技情报服务平台（下文简称应急平台），为突发事件中的应急决策提供科学、有效的支持。

苏新宁教授指出：“应急响应情报体系是一个以大数据环境为基、情报技术为力、情报流控制为策、应急决策为标的新型情报体系，适合处理需要快速反应的紧急情况”[1]。在知网中检索“应急响应情报”主题，返回的文献数量不多，仅 58 篇（检索时间 2020 年 7 月 9 日），研究方向众多，涵盖了理论模型、体系架构、知识库到具体行业案例。其中对应用信息化技术特别是大数据技术构建应急情报服务平台的研究很少，文献 [2-5] 主要从宏观层面介绍了在大数据环境下应用数据驱动对情报来源、处理分析和供给、利用进行改进，没有讨论具体的关键技术。而应急平台作为应急响应情报体系的重要支撑，在很大程度上是技术驱动的，也就是以创新的技术来满足服务对象的需求。应急平台的服务对象一般包括政府管理部门、相关专业人员和社会各界三个群体，不同群体对应急科技情报的需求是不同的。在突发灾害事件中，应急平台应能提供关于应急状况的背景情报，如地理数据、地址数据、受灾地普查数据和基础设施概况等，并满足制定应对政策和灾后应急计划的情报需求，包括相应的法律法规、经济相关情报等[6]。突发事件发生时，应急平台的服务对象需求的多样性、急迫性和不确定性，需要依靠数据采集、数据预处理、数据分析、数据可视化、数据安全等技术工具，实现快速响应，快速迭代，并迅捷高效地满足服务需求。基于应急平台技术涉及面广，复杂度高，对关键技术进行研究显得十分迫切和必要。

1 平台特点及架构

1.1 平台的特点

应急科技情报研究涉及应急管理、信息技术、情报学等学科领域，技术的多学科交叉融合，以及应急情报工作特点，对应急平台的技术架构提出了更高的要求，平台应具备以下特点：

及时性：应急背景下，及时性的重要性不言而喻，科技情报研究人员应能快速满足各级各类决策者的多种需求。

全面性：应急平台应拥有各种文献资源、各类媒体和信息资源，将多源情报融会贯通，这也是科技情报工作的一个重要目标。

准确性：应急平台研究的问题是决策者所关注的问题，应具备分析解决相关问题的方法库和知识库，还应具有应对突发事件的常识性知识，增强服务对象对突发事件的认识，以及普及应对突发事件的各种措施。

可读性：在应急期间，决策者很难长时间关注某一个方面的内容，因此，应急平台的产品要具备生动的表达方式，不能过于冗长。

跨界启示：应急平台提供的情报产品，甚

doi:10.3772/j.issn.2095-915x.2020.04.011

至可以超出服务对象通常所关注的范围，这类信息的启示作用是难以估量的[7]。

1.2 平台架构

为了能快速响应，应急平台一般应为轻量级，并将解决方案的组件分隔到不同的层中，每一层内部在抽象级别上基本一致，保持内聚性，与下面各层的耦合关系是很松散的。基于实践经验，我们认为可将应急平台划分为4层，如图1所示。

展示层的主要功能是实现平台数据的传入与输出，在当前技术背景下，展示层应重视数据可视化以及与融媒体结合。

业务层对具体问题进行逻辑判断与执行操作，也是展示层与数据层的桥梁，实现跨层间的数据连接和指令传达。从另一个角度来看，业务层需要构建多领域知识库来为业务决策提供支持。

平台层以软件为核心，为应用服务提供开发、运行和管控环境等中间件功能[8]，鉴于应急平台的特点，平台层的可伸缩性十分重要，自行或外包定制开发一般都难以快速实现，因此应考虑多采用开源软件平台。

图1 应急科技信息服务平台层级图

数据层包括了数据采集、清洗、预处理和存储等功能，面向能支持突发事件应急管理和决策的海量资源，在技术上存在较高的复杂性，在大数据时代，分布式存储成为一种新兴方案。

应急平台在多个层级上，均涉及大量的软件工具、平台，整体框架如图2所示。

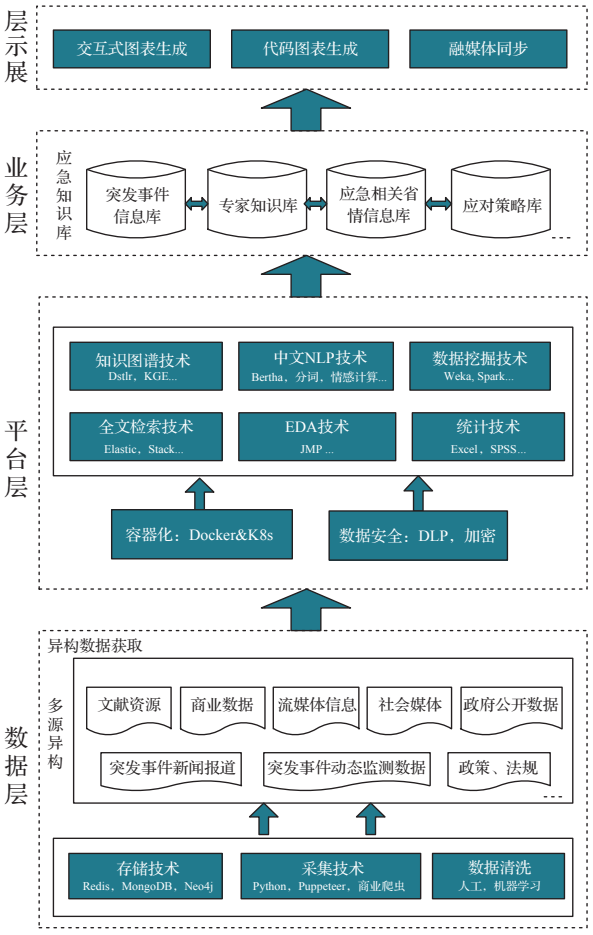

图2 应急科技信息服务体系框架图

2 平台的关键技术

2.1 业务层以应急知识库为核心

信息技术的不断成熟为知识管理的实施起到了巨大的推动作用，情报研究工作中引入知识管理是一种必然，应急平台需要构建多维的知识库，以满足应急响应不同领域的需求。应急事件的信息和知识在应急平台中的作用十分

重要，处于核心地位，不仅为突发情况提供直接的应对知识，更重要是为决策提供支撑。近年来，各科技情报研究机构建设了大量知识库管理系统，如机构知识库、科技人才知识库、各类常识库等，实现了特定领域知识的收集、存储、传播等功能。这些已有的知识库通过检索浏览功能可以迅速给应急平台提供问题答案。但在突发事件的背景下，它们涵盖的范围显然是不足的。随着技术的进步，当前需要构建的是具备层次结构的智能知识库，底层是“事实知识”，中间层是控制“事实”的知识（规则、过程等），最高层是“策略”[9]。为此，必须根据需求快速构建应急知识库。以下小节讨论构建知识库的几个关键技术。

2.1.1 知识库数据来源

数据是科技情报研究与服务的基础，采集全面、高质量、完善的数据是构建应急平台知识库的首要任务，知识库要支持多种类的信息和知识的有序化，对其进行大规模收集和整理、分类保存，并提供相应的检索手段。日常情报研究所涉及的数据包括商业数据、特色数据、政府公开数据、政策库、情报分析产品数据库等，其中一些结构化较好，质量较高。但在应急背景下，还需要采集来自互联网中散布的信息资源，如网页、论坛、社交媒体、维基百科、流媒体信息等，这些信息仅靠人工是很难获得较大数据量的，必须采用爬虫等技术手段批量自动获取。当前商业化、开源爬虫软件技术十分先进，可以满足大多数场景的数据采集需求，也可以采用 Python 等易于使用的脚本语言，自行开发爬虫以满足定制化需求。

知识库汇聚了多源数据的集合，一般应将其进行转换以形成可检索的格式。为了满足多种应用的需求，支持全文检索是十分必要的。以 Elastic Stack 为代表的交互式近实时搜索平台框架可以提供高速检索大数据的能力。此外，根据数据特点和后续处理的需求，也可以将数据以 Key-Value 数据库、文档数据库、图数据库等方式进行存储，典型的开源软件平台有 Redis、MongoDB、Neo4j 等，其中图数据库在构建知识图谱方面具有一定优势。

2.1.2 数据清洗

一般的数据分析过程中，必须在数据清洗过程中检查数据的一致性，处理无效值和缺失值，对数据进行重新审查和校验，也就是把“脏”的“洗掉”。值得注意的是一些虚假信息会给后续的分析挖掘带来困扰，所以在清洗的同时，也要进行可信度分析。大量实践经验表明，数据清洗可能会占到数据分析过程的 50%~80% 的时间。

传统上，数据采集应该遵循严格的规则，并根据预定模式进行格式化。这个过程被称为 ETL（Extract-Transform-Load: 提取 - 转换 - 加载），应急平台面对数据的洪流，往往没有时间和能力对其进行梳理来筛选出需求的信息，也不应苛求数据格式的统一性和完整性。因而，对数据进行采集和清洗更应该在日常下功夫，引入基于人机交互的数据清洗方法和在线 - 离线相结合的数据融合处理技术，或者将其后置到分析过程中一并完成，如归一化、验证及其他操作，这种范式被称为 ELT（提取 - 加载 - 转换），更强调数据传递的及时性。两种范式的对比如图 3 所示。

doi:10.3772/j.issn.2095-915x.2020.04.011

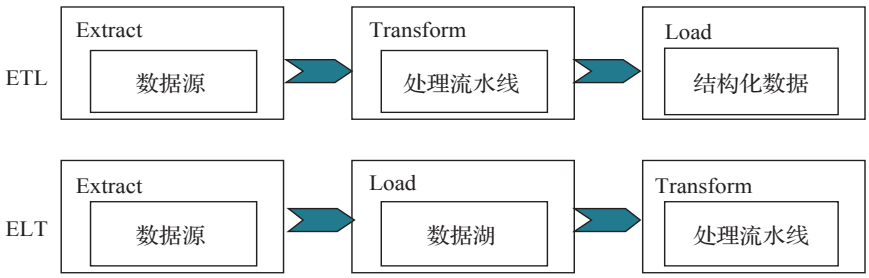

图 3 两种范式对比

大部分情形下，部署应急平台的研究机构很难获得海量高置信度的标注数据，因为其成本很高。另一方面，应急平台的任务时效要求高，所需迭代次数多。因此除了平时的人工积累，需要引入机器学习的“弱监督学习（Weakly Supervised Learning）”的工具或平台，它适合拥有海量未标注样本却只有少量标注样本的场景。第一类是不完全监督（incomplete supervision），数据训练集的一个小子集有标注，其他数据未标注；第二类是不确切监督（inexact supervision），数据只有粗粒度的标签；第三种是不准确的监督（inaccurate supervision），标注存在错误，或者数据本身难以分类[10]。

MIT 的 Curtis Northcut 和 Google 的 Jiang lu 在文献[11]中提出了“置信学习（Confident Learning）”框架，就是弱监督学习的一个分支实现，用来识别错误标注，进行“带噪学习（noisy label learning）”，并开源了 CleanLab 软件，支持多种图像、NLP 处理任务，支持自定义模型，适合 scikit-learn, PyTorch, TensorFlow, FastText 等机器学习框架。

2.1.3 知识抽取

知识抽取（Knowledge Extraction）的目标在于从海量数据中抽取出三元组信息，包括“实体 - 关系 - 实体”以及“实体 - 属性 - 实体”等两类，可划分为术语抽取、实体抽取、关系抽取、事件抽取、共指消解等多个子任务。应急平台收集的大量中文文本数据，相比英文等字母语言数据，在计算机处理和分析的过程中存在很大的差异，也形成了独特的难点。例如中文的“词”之间没有分界符，需要使用中文分词 (Chinese Word Segmentation) 技术将连续的字序列按一定的规范重新组合成词序列。传统中文分词算法主要有基于字符串匹配的算法、基于统计的算法和基于理解的算法等三类，这些技术大多比较成熟，也达到了一定的精度，但在实践中仍存在许多问题，如未登录词，指代消岐等。这就需要依赖大量的人工标注来提高准确率，而相对于应急平台的快速响应特点，人工手段的效率显然满足不了要求。

2018 年以来，Google 发布了 BERT 模型，在各项 NLP 任务如 NER（命名实体识别）、CLS（分类）、QA（问答）、MT（机器翻译）中表现十分抢眼，它针对中文给出了基于字的模型，不再需要进行中文分词，并提供了预训练模型（否则需要巨大的算力和数据量才能取得较好效果）[12]。为了将中文词的信息引入模型，百度以及哈工大为代表的工业界和学界也推出

ERINE、全词覆盖的中文 BERT 预训练模型等，在多个中文数据集上取得了更好的效果。在应急平台中，应尝试应用此类新技术，尤其是实体识别和实体对齐，在判研科技人才、科研机构等场景中，同名专家是否是同一个人，一个机构的曾用名、别名、简称等，可能大大影响分析的结果。近两年的大量研究和测试表明，BERT 模型应用到 NER 中，可以利用小数据集进行训练，大幅减少了人工标注的需求，训练时长、准确率都达到了工业级别，十分适合在应急平台知识库中应用。

几年前，知识抽取主要利用语义 Web 技术、中文 NLP 技术和启发式规则对中文数据进行处理 [13]。近年来，AI 技术高速发展，并得到了广泛的应用，深度学习的神经网络模型（循环神经网络、卷积神经网络、Transformer 等）在知识抽取方面表现突出。例如，通过远程监督（Distant Supervision），将纯文本与现有的知识图谱进行对齐，可以实现自动标注大规模的训练数据集 [14]，在 ACL2019 上，微软的艾伦人工智能实验提出的 COMET 模型，在自动构建知识图谱上的效果已经接近人类 [15]。

2.1.4 面向知识图谱的知识推理

应急平台通过自动知识抽取将数据转换为三元组形式，然后构建合适的数据模型，形成标准的知识表示，这是“已知”的知识。在此基础上，可以通过知识推理产生“新的”知识，仍以三元组形态呈现 [16]。知识图谱由语义网络发展而来，包含着许多三元组，就当前的技术而言，数据来源存在大量干扰，知识抽取的精度也不能尽善尽美，得到的知识图谱并不完备，还可能存在矛盾知识。因此，需要通过知识推理对其进行改进，即知识图谱补全和去噪，这样才能提高后续应用的性能，如应急领域的垂直检索、问题解答等。

知识推理的传统方法包括基于规则和本体推理两种，能满足准确率较高的要求，但可计算性差。随着分布式表示、神经网络等技术的发展，面向知识图谱的知识推理发展出新的推理方式，如单步推理和多步推理。ProjE、MT-KGNN、HNM、R-GCN、IRN、DCN 等模型从减小参数规模、利用图谱的属性信息、结合单词语义、应用图的特性、共享记忆组件等方面对提高知识推理的精度、语义理解起到了促进作用 [17、18]。在应急平台知识库中应用，能大幅提高检索的质量，并对政策研究、潜在解决方案发现等有很大的帮助。

2.1.5 知识库的推荐算法

应急平台的知识库支持用户画像和内容推荐，这样一方面能加强科研人员的业务契合度和研究协作性，另一方面能更为准确地了解服务对象的需求，以此为其推荐或定制对应的科技情报服务，进行信息服务精准服务，提高应急服务的有效性。近年来，基于协同过滤的推荐算法应用十分广泛，结合用户的基本信息和行为信息如访问记录、停留时间等进行分析，以此建立用户画像库，分析其潜在的科技情报服务需求 [19]，对其内容偏好进行建模，向其推送相似或相关的其他内容。但这类方法依赖历史数据，效果与用户历史数据量的大小和准确性直接相关，并且对用户偏好的演变无法及时响应。而在应急平台中，由于大多数任务具有紧迫性，用户可能需要追踪多领域的知识，兴趣点在短时间内变化很快，这样推荐算法就不

doi:10.3772/j.issn.2095-915x.2020.04.011

能很好地适配。

知识图谱在表达实体关系方面具有天然的优势，也很容易展示出上下位的实体，因此，通过知识图谱结合路径和内容的推荐算法具有一定的优势。例如，应急平台检索某个科研机构，则“供职”于该机构的科研人员作为下位实体可能被算法选中进行推荐；属性相同或者相关的其他实体也可能被推荐，例如检索科研人员时，推荐其论文或项目的合作者。目前基于知识图谱的混合推荐算法有基于元路径相似、基于信息相似等，知识图谱的路径中包含了“隐含语义”，因而这类推荐算法的可解释性更好[20]。

2.2 平台层技术

2.2.1 容器化

在应急平台上，应该尽量使用开源框架或成熟的第三方框架，减少开发量，把重点聚焦在业务逻辑上。由于应急平台的特殊性，有时需要频繁发布部署新版本的应用程序，这就需要应用容器化技术，使应用程序能够简单快速地发布和更新，而无需停机。容器将应用程序的代码、运行时、系统工具、系统库和配置打包到一个实例中。与虚拟机相比，它十分轻量，占用的空间小，启动速度快。容器无需指定资源数量，它能动态使用服务器上的资源，具备高弹性。因而，应急平台可以采用容器技术部署密集的应用环境，以充分利用物理服务器的资源，这对于应用的快速部署、快速执行意义重大。

容器技术的典型代表是 Docker，它把容器化技术推向了一个高潮。随着容器的大规模应用，人们需要一个强大的管理平台对容器进行管理，2014 年，来自 Google 公司的 K8S（Kubernetes）开源迅速崛起，目前已成长为容器管理领域的事实标准，它的基本架构如图 4 所示。在主控上（master）的 etcd 用来保存整个集群的状态，API Server 提供资源操作（认证、授权、访问控制、API 注册和发现等）的入口，Controller Manager 负责维护集群的状态，Scheduler 负责资源的调度；在工作节点（node）上由 Kubelet 维护容器的生命周期，提供容器的运行时（runtime）等[21]。

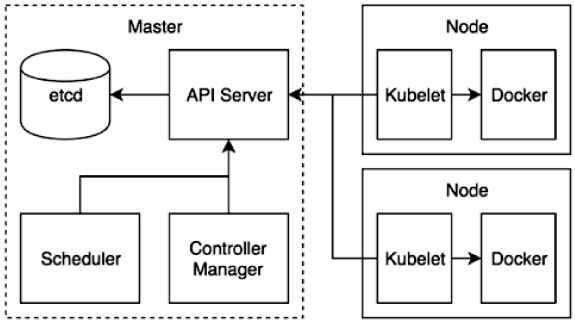

图 4 K8S 的基本架构

在应急平台上使用 K8S 来对各类容器进行管理是一种自然选择，它的概念较多，学习曲线门槛较高，但相较于其带来的管理方便性，投入学习成本是值得的。目前也出现了许多开源的集成化的企业级容器管理平台，大幅降低了部署容器管理的复杂度，其代表是 Rancher。

2.2.2 探索性数据分析

传统上，在对数据进行初步处理后，根据数据分布（可能只是假设而非真实）我们选择某个或某些模型来对数据进行分析，在很多情况下，这会带来“偏见”，让数据结果更显著来支持预设的立场。实际上，在应急平台上，可能已经有了一定量的数据，但研究者对这些数据并不了解，又缺乏足够的行业背景知识，

无法识别数据的特征，往往感到无从下手，不知该采用什么模型。在这种情况下，应尝试进行探索性数据分析（Exploratory Data Analysis，EDA）。EDA 是一种数据分析方法，由美国统计学家 John Tukey 于上世纪 60 年代提出，它在尽可能少的先验假定下进行探索，通过制作图表、计算特征量等方法探索数据的结构和规律 [22]。其目标一般为：

（1）根据观测到的现象，提出解释成因的假设；

（2）对假设进行统计评估；

（3）选择适当的统计工具和技术；

（4）根据初步探索结果为下一步收集数据提供凭据。

从以上目标可以看出，它分为探索和验证两个阶段，前者探求线索和证据，发现数据中隐藏的“知识”，而后一个阶段评估这些证据，更精确地分析具体情况。应急平台对数据进行初步分析时，常规的统计分析可能效果不明，可以采用 EDA 探索数据的模式与特点，再根据情况选择和调整合适的分析模型，同时也能发现数据相对于常见模型的偏离。更进一步采用以显著性检验和置信区间估计为主的统计分析技术，就可以科学地评估发现的数据模式。同时，EDA 十分强调直观和数据可视化，使分析者能一目了然地发掘数据中隐含的有价值的“知识”[23,24]。

应急平台的使用者并不一定都具备统计学基础，因此应选择交互性强且易于上手的统计发现软件工具或平台，如 SAS 发布的 JMP 等。

2.2.3 数据安全

应急平台产生的情报产品的内容具有特殊性，在相当程度上可能有一定的敏感性，因此其数据安全也是十分重要的。

从数据来源上看，应尽量采集经相关部门授权的、可公开的数据，并进行相应的脱敏处理。如果在数据中包含了一些个人隐私信息，需要用户主动同意隐私协议，且用户有权随时撤销授权，在获取个人信息数据时要秉持最小化原则。在技术层面上，应采取措施防止数据泄露，区分敏感数据，对数据进行分层管理，在传输与存储数据时，若有必要应进行加密。对于核心数据，应采用透明加密方式来控制数据的安全，在这种模式下，文件会被强制加密，而格式和形态不会发生变化。如果没有经过安全通道，或者没有合法身份验证或访问权限，都无法访问该文件。

数据安全保护，对外应确保应用安全，对内应加强数据泄露防护（Data Leakage Prevention, DLP）。同时，必须在制度、机构、人员等方面加强管理，以避免发生安全事件。DLP 的商业化解决方案很多，也有开源的 OpenDLP 等软件。

2.3 展示层技术

2.3.1 数据可视化

为了清晰有效地传递信息，应急平台应结合数据可视化技术，采用统计图形、图表、信息图表和其他工具在视觉上传达信息，这样能使得复杂的数据变得易于理解，更容易启发传统的描述性统计难以提供的洞见。数据可视化可以在应急平台的全流程上应用：从信息抽取预处理到网络型数据的分析和推理，到最终情报分析结果的判读等。在对原

doi:10.3772/j.issn.2095-915x.2020.04.011

始数据进行分析提取时，可视化能呈现出数据之间的关系和趋势，这样分析人员能更好地进行发掘和利用；知识图谱技术的深入发展，也在一定程度上提高了数据可视化的应用，在科技情报中，许多实体隐含的关联，通过网络图等形式进行展示，显得直观简洁；最终情报产品的展示，通过可视化方式展示，能让服务对象产生大体直观的印象，有利于进一步的情报分析和判读。[25-27]

特别值得注意的是，在科技情报研究中结合地理信息、时间序列，并以可视化方式呈现，对于应急平台的服务对象来说，更容易发现其中发展规律，为制定相关应急对策提供有力的支撑。图 5 展示了应急平台上产出的一张抗疫期间福建省抗疫相关论文与其他地区机构的合作关系图，直观呈现了地域特点。

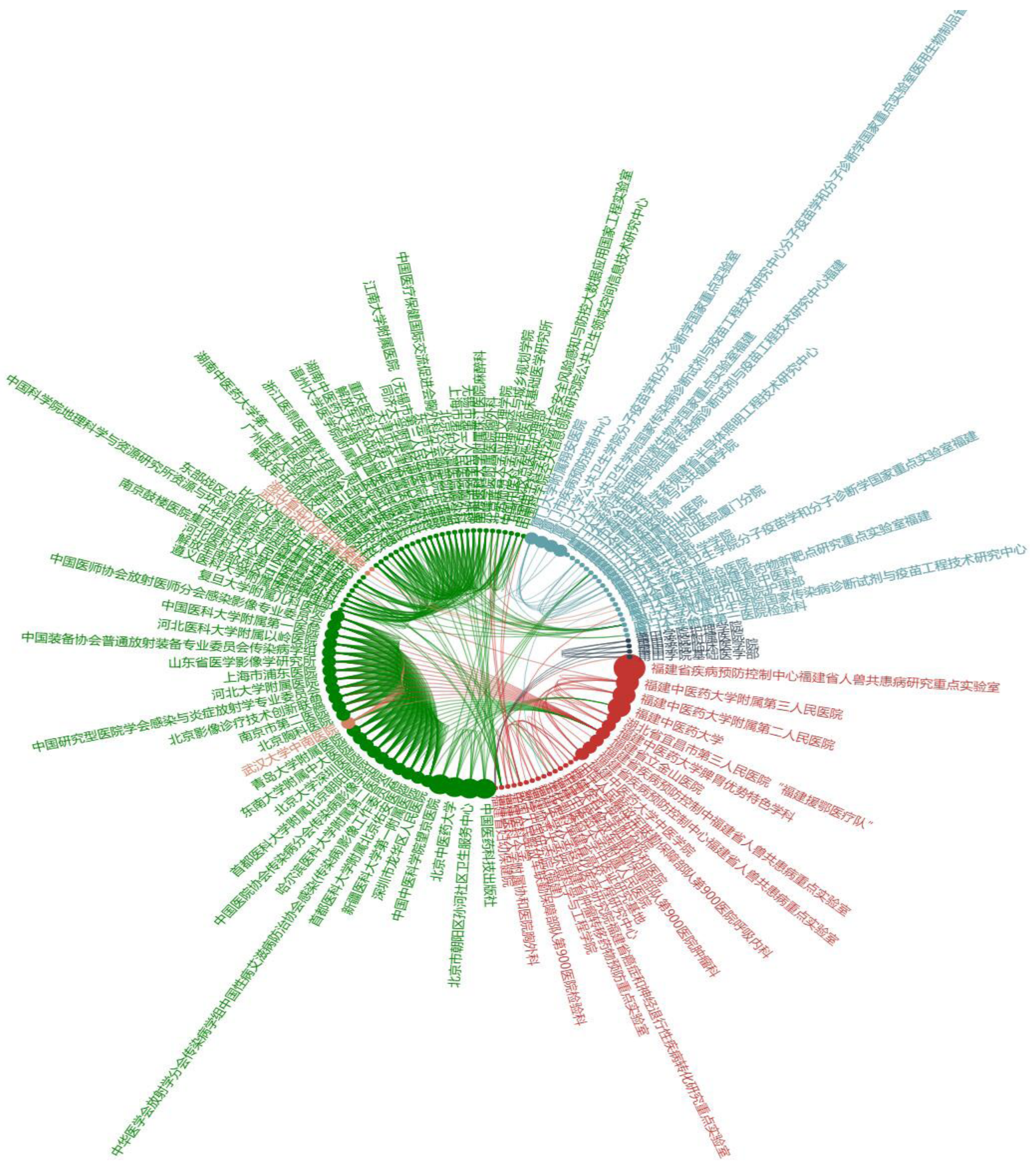

图 5 抗疫期间福建省抗疫相关论文与其他地区机构的合作关系

2.3.2 结合融媒体

近年来，移动互联网对人们生活、工作的各个领域产生了深远的影响和改变，应急平台也应适应互联网发展的变化，照顾用户使用习惯，结合多种融媒体，如微信、微博、客户端等，为服务对象提供多样便捷的信息获取渠道。以往，科技信息研究的产品多数以纸质的形式提供给服务对象，其阅读情况、关注兴趣点、是否分享等情况多数难以获知。若能结合融媒体，则应急平台能更多掌握相关信息，主动发现目标用户，向其推送更及时精确的应急科技情报产品，这样才能实现应急情报服务与用户之间的信息流双向流动。

整体上看，应急平台的服务对象特点与一般科技情报的用户特点有相当大的差别，更需要个性化、专业化的应急科技情报服务。不同的用户层次对应急科技情报的需求可能存在较大差异，因此，其用户行为数据采集路径不同，也要用不同方法进行处理。用户行为数据的维度越高，就越能准确地刻画用户画像，越能实现应急科技情报产品与其精准对接[28]。

文献 [29-30] 的研究指出，当前我国科技政务、科技期刊的新媒体获得了一定的发展，主要在信息推送方面表现较好，但互动性和服务能力还相当薄弱，尤其微博的表现更差。结合目前微信普及情况，应急平台融合新媒体应将重点放在微信公众号上。

此外，数据可视化的动态效果也只有在融媒体上才能充分地呈现，纸质产品的效果将大打折扣。

3 应用案例及效果

2019 年底的这一次重大突发公共卫生事件，对我国乃至世界公共卫生安全带来极大挑战。政府科技主管部门主要承担了科研攻关、科技企业复工复产等工作任务，亟需获取及时、全面、可靠的应急决策情报支持。为此，福建省科学技术信息研究所组织了跨学科的科研协同攻关团队，充分发挥情报研究、数据挖掘等业务专长，构建了应急科技情报服务平台，推出《科技战疫专报》应急情报服务产品，强化面向政府科技管理部门的应急科技情报服务，为突发事件的精准决策提供重要情报支撑。

该应急平台集成了多个原有的软件工具 / 平台，包括一套自研分布式数据采集系统，以 MongoDB、Elasticsearch、Neo4J 等数据库作为存储系统，还有一套自研知识库系统，一套 Superset 开源可视化系统。此外根据工作任务目标，还动态整合新加入了多个功能模块，以在不同层面上满足需求。这些应用 / 服务都采用 Docker 技术来部署，由 Ranchers 对 Docker 和 K8S 集群进行管理和调度。应急平台重视和强调快速反应特性，下面以福建省生物医药产业的数据挖掘分析为例，对应急平台的实际工作流程做简要介绍。

生物与新医药是福建省重点培育发展的新兴战略产业之一，在此期间，科技管理部门需要了解本省高新技术企业尤其是生物 / 制药 / 医疗等直接相关高企的发展状况，为科技政策调整提供支持。应急平台的技术人员立即采用 Python 编写爬虫模块在较短时间内采集了文献信息 2 千余条，政策信息 1 千余条，科技项目信息 3 万余条，科技人员信息 1 万余条，企业信息 8 千余条，企业知识产权信息 20 余万条，然后从以往存储的福建省各科技项目知识库中

doi:10.3772/j.issn.2095-915x.2020.04.011

检索出相关高企申报的项目 140 多项。以上数据经软件自动和人工清洗后，先进行探索性数据分析，初步了解了福建省相关高企的知识产权产出特征及其在科技项目申报中的主要领域。研究人员调整了部分模型后，对知识产权产出与企业注册资本、所属行业等的相关性进行分析，采用 Jupyter Notebook 简单编写代码进行数据可视化（图 6）。图 7 为福建省生物 / 制药 / 医疗高新技术企业申请的科研项目名称词云图。最后，在应急平台提供的数据分析基础上，研究人员提炼总结了福建省生物 / 制药 / 医疗相关高企的发展特点，提出了支持生物安全关键技术研发，把福建省生物医药产业打造成为具有一定竞争力的新兴支柱产业的相关政策建议，撰写报告提交给科技管理部门，整个流程不到 4 天时间。

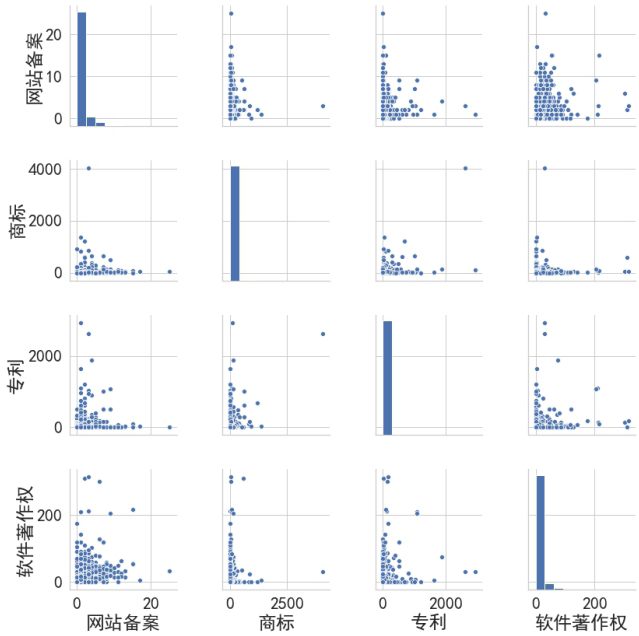

图 6 Jupyter Notebook 可视化

图 7 福建省生物 / 制药 / 医疗高企申请的科研项目名称词云

4 结语

构建应急科技情报服务平台，是科技情报机构提升核心能力的重要基础性工作。平台构建涉及多学科、多种技术的交叉应用，本文对其中的多源异构数据采集、数据清洗、探索性数据分析、中文 NLP 处理、知识库构建、容器化技术、数据可视化、融媒体技术应用和数据安全等部分关键技术技术内容进行了研究探讨，提出相关技术的解决思路和方案，并以实例进

行了说明。可以看出，应急平台数据规模大、来源多样、实时性要求高，所需的技术栈十分复杂，把各类大数据技术应用于其中是未来发展的必然方向。情报研究机构和人员应该树立大数据意识，加大对技术方面的投入与研发，以提升科技情报研究和服务能力。

参考文献

[1] 苏新宁等 . 应急响应情报体系 : 理论、技术与实践 [M]. 北京 : 科学出版社 , 2019.
[2] 蒋勋 , 朱晓峰 . 基于政府大数据能力建构的智库应急情报服务 [J]. 社会科学文摘 , 2020(4):118-120.
[3] 蒋勋 , 张志祥 , 朱晓峰 , 等 . 大数据驱动智库应急决策的情报架构 [J]. 情报理论与实践 , 2019(5):25-32.
[4] 陈迎欣 , 李烨 . 智慧城市大数据背景下应急管理情报体系构建 [J]. 价值工程 , 2019(33):290-291.
[5] 张运良 , 丁思媛 , 高雄 . 突发事件评论集中的情报甄别方法初探 [J]. 情报工程 , 2020(2):21-35.
[6] 宋丹 , 高峰 . 美国自然灾害应急管理情报服务案例分析及其启示 [J]. 图书情报工作 , 2012(20):79-84.
[7] 武夷山 . 超出“广快精准特”的情报产品要求 [EB/OL]. [2020-5-13]. http://blog. sciencenet. cn/home. php?mod=space&uid=1557&do=blog&id=1232923.
[8] 梁晓琴 . 基于云计算的政企互动电子政务模式研究 [D]. 天津 : 天津大学 , 2013.
[9] 星朝 . 知识库系统的一些思考 [EB/OL]. [2019-05-16]. https://www.cnblogs.com/jpfss/p/10876180. html.
[10] Zhou, Z H. A brief introduction to weakly supervised learning[J]. National Science Review, 2018(5):44-53.
[11] Northcutt, Curtis&Jiang, Lu&Chuang, Isaac. Confident Learning: Estimating Uncertainty in Dataset Labels[J]. 2019, arXiv:1911. 00068.
[12] 望江小汽车 . 2018 年最强自然语言模型 Google BERT 论文全文中译 [EB/OL]. [2019-2-19]. https://juejin. im/post/5c6b5975e51d452704714573.
[13] 车海燕 , 冯铁 , 张家晨 , 等 . 面向中文自然语言文档的自动知识抽取方法 [J]. 计算机研究与发展 . 2013, 50(4):834-842.
[14] 韩旭 , 高天宇 , 刘知远 . 知识图谱从哪里来：实体关系抽取的现状与未来 [EB/OL]. [2019-11-19]. https://zhuanlan. zhihu. com/p/91762831.
[15] Bosselut, A, Hannah R, Maarten S, et al. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction[C]. ACL, 2019.
[16] 艾瑞咨询 . 2020 年中国知识图谱行业研究报告 [EB/OL]. [2020-4-15]. https://www. shangyexinzhi. com/article/1686182. html.
[17] 官赛萍 , 靳小龙 , 贾岩涛 , 等 . 面向知识图谱的知识推理研究进展 [J]. 软件学报 . 2018, 29(10): 2966-2994.
[18] 张仲伟 , 曹雷 , 陈希亮 , 等 . 基于神经网络的知识推理研究综述 [J]. 计算机工程与应用 . 2019, 55(12):8-19.
[19] 曾冠桦 . 基于科研知识库的文档推荐系统的设计与实现 [D]. 哈尔滨 : 哈尔滨工业大学 ,2018.
[20] 高松 . 基于知识图谱的合作者推荐系统设计与实现 [D]. 大连 : 大连理工大学 , 2019.
[21] CodingCattwo. Kubernetes(K8s) 初 探 [EB/OL]. [2019-8-28]. https://www. jianshu. com/p/13f7c5deec55.
[22] Tukey, John W. Exploratory Data Analysis[M]. Addison-Wesley Publishing Company, 1977.
[23] John Behrens. Principles and Procedures of Exploratory Data Analysis[J]. American Psychological Association. 1997, 2.
[24] 陈昌春 . 所谓探索性数据分析 Exploratory Data Analysis 更宜译作“试探性”[EB/OL]. [2013-2-18]. http://blog. sciencenet. cn/blog-350729-662859. html.
[25] 钱虹 . 面向技术创新生态的科技情报服务平台建设——以陕西省科技情报综合服务平台为例 [J]. 中国科技资源导刊 , 2020(5):83-88.
[26] 曾文 , 车尧 , 张运良 , 等 . 服务于科技大数据情报分析的方法及工具研究 [J]. 情报科学 , 2019, 37(4):92-96.
[27] 杨岩 , 姚长青 ; 董诚 , 等 . 区域科技创新可视化平台开发 [J]. 地理空间信息 , 2019, 17(8):1-4.
[28] 王益成 , 王萍 . 基于用户动态画像的科技情报服务推荐模型构建研究 [J]. 情报理论与实践 , 2019(4):83-88.
[29] 方延风 . 我国省级科技政务新媒体发展现状研究 [J]. 科技成果管理与研究 , 2020(4):12-15.
[30] 俞敏 , 吴逊眉 , 武瑾媛 . 基于移动端的科技期刊新媒体内容多平台发布策略研究 [J]. 编辑学报 , 2020(3):307-313.