·名家讲坛·

构建数据密集型科技情报范式*

罗　威
(军事科学院军事科学信息研究中心,北京,100039)

[摘　要] 本报告分析了新时代科技情报工作所处环境发生的变化,提出了构建数据密集型科技情报范式的发展构想,并从业务模式、数据资源、技术手段、协同应用等角度进行了阐述,最后介绍了开展的探索性研究情况。

[关键词] 数据密集型　科技情报　人工智能　大数据
[中图分类号] G250.2　[文献标识码] A　[文章编号] 2095-2171(2021)02-0012-04
DOI:10.13365/j.jirm.2021.02.012

1　挑战与机遇

当前，国际政治经济形势风云变幻，以大数据、人工智能为代表的信息技术迅猛发展，科技情报工作面临的需求环境、信息环境和技术环境正在发生着深刻的变化，挑战与机遇并存。

1.1　需求环境

作为科技发展全面领先的国家，美国高度重视科技情报工作。 2017 年，美国将科技情报纳入《国家安全战略》，指出“几乎所有的现代武器系统都依赖源于科技情报的数据”，确立了科技情报在国家安全中的重要地位。 党的十九届五中全会将科技创新的重要地位摆在了前所未有的历史高度，作为科技工作的“耳目”“尖兵”，科技情报工作可以为科技政策制定、科技创新发展方向选择等提供有力支撑，并正在国家科技创新体系中发挥着越来越重要的作用。

1.2　信息环境

近年来，科技信息呈现爆炸式增长。 以科技论文为例，Scopus 数据库每年新增 300 万篇文献；在生物医学领域，每年有超过 100 万篇科技论文被收入 PubMed 数据库，差不多 1 分钟 2 篇。 此外，随着全球科技竞争加剧，一些重要科技信息的发布受到不同程度的影响，科技情报相关原始信息获取难度加大。

1.3　技术环境

大数据、人工智能等技术的快速发展，为科技情报手段跃升提供了强大支撑。 在信息处理层面，以往处理粒度以单篇文献为主，而有效利用大数据、人工智能等技术，开展大规模机器学习和深度自然语言处理，处理粒度可以细化到段落和句子，处理范围可以扩展到文本、图像、视频等各类信息，处理结果可以形成包含人物、机构、技术、装备等要素的知识图谱。 在情报分析层面，情报研究人员往往先要花费大量时间去搜集、整理信息，再依靠头脑中的知识和经验进行分析研判。 有效利用大数据、人工智能等技术，让计算机完成初步的情报素材分析与整编工作，为脉络分析、情报预警、趋势预测等提供高效技术支撑，可让情报研究人员聚焦于高智力活动，从而大幅提高情报研究工作的效率和质量。

2　发展构想

为了在新时代更好地发挥科技情报工作的

---

* 本文根据罗威副研究员 2020 年 12 月 6 日在 2020 中国情报学年会暨情报学与情报工作发展论坛和第十届(2020 年)全国情报学博士生学术论坛上做的特邀报告内容整理而成。

[作者简介] 罗威,副研究员,硕导,研究方向为大数据技术,数据挖掘。

本文引用格式:罗威.构建数据密集型科技情报范式[J].信息资源管理学报,2021,11(2):12-15.

作用，我们应该直面挑战，抢抓机遇，充分融合数据智能和专家智慧，探索形成数据密集型科技情报范式，即以科技信息大数据资源为基础，以可重用的情报研究业务模式为牵引，以人工智能等新技术领域化应用为手段，通过充分挖掘数据价值和融合专家智慧，形成人机协同的典型科技情报业务解决方案。

2.1 构建高质量大数据资源体系

大数据资源体系是数据密集型科技情报范式的根基。从组成来说，大数据资源体系包含基础信息资源和情报对象库两种数据。其中，基础信息资源包括各类与前沿科技相关的组织规范、融合集成的信息资源，如论文、报告、规划、年鉴、投资、需求等，以篇为基本单元，通过元数据进行组织。情报对象库涵盖相关机构、人员、项目、技术、装备等各类情报对象的基本情况及其发展变化，以知识节点为基本单元，通过属性和关系进行组织。大数据资源体系构建示意图如图 1 所示。

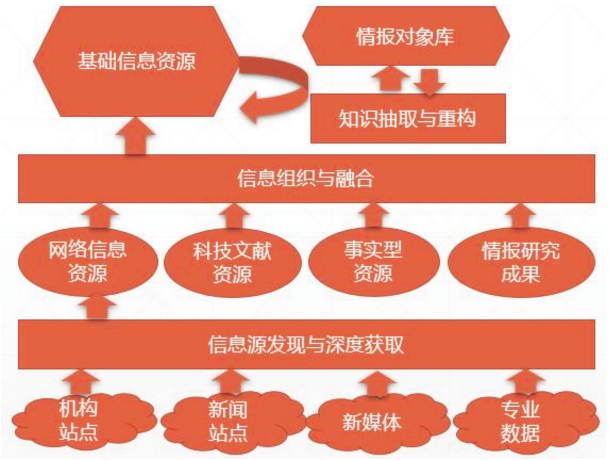

图 1　大数据资源体系构建示意图

构建高质量大数据资源体系，需要解决好三个问题：一是信息源研究，重点研究互联网重要科技网站和新媒体账号等信息源，以及各类机构内部出版物，为及时高效搜集高质量信息提供支撑。二是知识组织体系构建，既需要传统的知识组织工具，如新兴技术分类与主题词表，对杂乱无章的信息进行序化，也需要针对情报对象库研究建立领域本体，对多来源知识进行规范化组织。三是信息深度揭示，重点实现信息的碎片化、标签化、关联和还原。其中，碎片化将信息揭示粒度细化到节、段落、句子甚至三元组；标签化按国别地区、军兵种、行业、技术领域、事件类型、语义功能等不同纬度，实现对碎片化信息的多层次分类；关联解决实体名称的别名与歧义问题，实现基于语义的信息与知识汇聚；还原主要是通过对碎片化信息的汇聚，恢复事物的原貌。

2.2 提炼可重用的情报研究业务模式

科技情报研究是高智力活动，其流程方法无一定之规，导致当前对过程和成果的把控不够精准，数据与技术工具支持的程度不够高。我们认为，可以参考软件工程的理念，探索科技情报工程方法路径，针对不同类型的情报研究任务，制定信息搜集、分析研判、成果产出的流程规范和最佳实践。这是个非常重要的问题，试想如果情报研究过程都是凭着感觉走，一千个人有一千种流程，那要实现数据密集型科技情报范式就无从谈起。这就要求我们对动态情报研究和专题情报研究等任务进行精细分类，根据不同类别任务特点，基于实践经验进行总结提炼可重用的业务模式。一种针对前沿创新科研项目的技术溯源业务流程如图 2 所示。

当然，在科技情报工作中应用大数据、人工智能等技术，不仅仅是实现已有流程的信息化、自动化，还可能会重塑现有的情报研究业务模式，这就需要情报研究人员与技术人员通力合作，针对具体任务大胆探索，并进行抽象泛化。总之，需要通过不断的实践探索

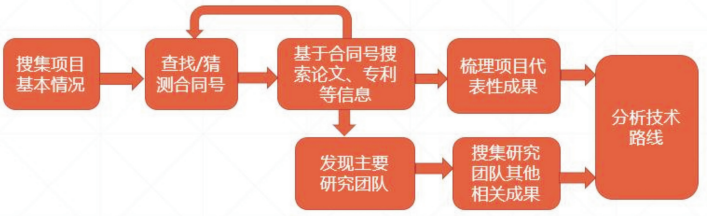

图 2　一种针对前沿创新科研项目的技术溯源业务流程

和总结提炼，最终形成新时代科技情报研究的理论方法体系。

2.3　开发先进可用的技术与模型

技术与模型是构建数据密集型科技情报范式的驱动力。 应围绕科技情报的特色问题，开发先进可用的技术和模型。 一是围绕基础信息资源构建，基本问题是明确与前沿科技相关的信息资源有哪些，以及如何进行高效的获取与组织，涉及信息源辅助发现、信息深度获取、自动组织等关键技术与模型；二是围绕情报对象库构建，基本问题是明确与前沿科技相关的知识内容，以及如何从大数据中抽取、重构和融合这些知识内容，涉及机器阅读、信息抽取、知识图谱等关键技术与模型；三是围绕情报智能分析，基本问题是如何从海量数据中获取有价值的信息和结论，以及如何与情报研究过程进行深度结合，涉及信息评估、脉络生成、技术预测、技术溯源等关键技术与模型。

在技术研发层面，重点要将通用计算机领域技术问题科技信息化，如研发针对科技信息的知识抽取、情感分析、自动问答、自动综述等技术。 为此，要利用各种已标注的数据或采用人工标注方式，建设基础语料库，吸引计算机学科力量共同解决技术问题。

在模型研发层面，要注重充分利用已有的指标与模型，在此基础上根据科技情报研究关键环节的应用需求进行优化完善和补充开发，这需要联合计算机技术、科学学等多学科力量共同完成。

2.4　形成面向典型场景的人机协同应用

科技情报研究需要大量依靠专家智慧，在目前甚至以后相当长的时间内，都无法实现完全的自动化、智能化。 要构建数据密集型科技情报范式，关键就是要秉持融合数据智能和专家智慧的理念，形成系列面向典型场景的人机协同应用。 其中，数据智能主要体现在应用技术手段对科技信息大数据进行挖掘，识别热点、突变、关联、时序等有价值的情报线索；专家智慧主要体现在由专家定义问题、形成假设、寻找证据、分析研判、解读结果等。作为美国情报界从事高风险、高回报技术的创新机构，美国情报高级研究计划局（IARPA）也将下一代情报愿景描述为自动化（Automation）＋智慧的人（Smart People）。

要形成面向典型场景的人机协同应用，当前较为可行的途径是实现人机协同流程管理与特定环节的数据驱动。 前者基于可重用的情报研究业务模式；后者则需要针对特定需求，应用先进可用的技术和模型，对高质量大数据资源体系进行挖掘和分析。 其中，需要着力解决两个关键问题：一是合理界定专家与计算机的工作边界，各取所长，同时可以较为友好地进行人机交互；二是对专家智慧进行有效管理，让专家的知识可存储、可重用、可计算。

3　探索与实践

近年来，我们秉持理技融合的发展理念，在面向典型场景的科技情报人机协同应用方面进行了持续探索，取得了阶段成果。

3.1　情报线索碎片化还原

针对科技情报研究观点生成和信息佐证问题，开展情报线索碎片化还原方法与技术研究。 情报线索碎片化还原主要有两类数据产出：一是信息流，即从海量数据中萃取关键信息；二是报文流，基于对关键信息的碎片化，辅助形成情报研究假设，并基于假设开展碎片化信息循证。

整个流程充分体现融合数据智能与专家智慧。 在融合数据智能方面，应用大数据、

人工智能等技术对海量信息进行采集处理和内容挖掘，涉及信息清洗、自动摘要、自动翻译、自动分类、信息抽取、观点挖掘、情感分析等一系列技术研发与应用。同时，还要建立碎片化信息语义标签体系，如基本情况、性能参数、组装配置、里程碑节点、专家评价等。在融合专家智慧方面，需要专家定义研究问题、明确信息需求、梳理问题假设、整编情报产品等。

3.2 基本情况体系化积累

针对科技情报相关知识的汇聚、更新、共享和应用问题，开展基本情况体系化积累方法与技术研究，形成初具规模的领域知识库＋知识图谱。

重点围绕解决知识覆盖率、准确性和鲜活度这三个问题，提出了一种较为可行的基本情况体系化积累模式。其中，态势扫描监测是驱动，责任主体是动态编辑部，要求细致、及时捕捉发展变化，成果形式是各类要闻；基本情况研究是基础，责任主体是情报研究专家，要求深入、准确地形成基本情况，成果形式是研究报告、百科等；知识协同构建是关键，责任主体是知识工人，要求知识内容全面、鲜活，成果形式是知识库＋知识图谱。在此过程中，应用深度学习、自然语言处理等技术开发关键模型算法，为人机协同构建领域知识提供高效支撑。

3.3 前沿技术扫描与预测

针对前沿技术的跟踪与研判问题，开展前沿技术扫描与预测相关方法及技术研究。其中，围绕技术预测这个重难点问题，目前业界通行做法是依靠专家调查，但是受制于专家的研究背景和认知局限，实施过程和效果还有较大的提升空间。我们通过不断摸索，提出并实现了一种融合专家智慧与数据智能的技术预测模式。具体来说，选取发展里程碑、新兴度、突变性、研发力量、应用功效等关键指标，研究提出数据支撑的指标计算方法，并据此对大数据资源体系进行适应性加工组织，开发算法工具，实现对指标的高效计算。在此基础上通过数据驱动形成初始技术清单，由专家对初始技术清单进行调整，并参考相关指标对技术的性度进行评价。此种模式综合了专家智慧与数据智能的优势，各取所长，有效提升了技术预测的效率和质量。

4 结语

构建数据密集型科技情报范式是体系性工程，需要加大学术界与科技情报从业机构的合作，共同凝练关键问题、研究方法途径、开发工具系统，通过不断的实践探索，切实提升科技情报工作整体水平，为我国科技创新提供有效支撑。

参考文献

[1] The White House. National security strategy of the United States of America[EB/OL].[2020-11-30].https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905-2.pdf.

[2] Scopus Guide[EB/OL].[2020-11-29].https://guides.lib.rpi.edu/scopus_guide.

[3] Statistical reports on Medline/PubMed baseline data[EB/OL].[2020-11-30].https://www.nlm.nih.gov/bsd/licensee/baselinestats.html.

[4] 赵丹群.文献计量范式下的科学知识图谱研究：新进展与新挑战[J].情报学进展,2020,13:354-380.

[5] Nasar Z, Jaffry S W, Malik M K. Information extraction from scientific articles: A survey[J].Scientometrics,2018(117):1931-1990.

[6] Ammar W, Groeneveld D, Bhagavatula C,et al. Construction of the literature graph in semantic scholar[C]//Proceedings of NAACL-HLT,2018. Association for Computational Linguistics,2018:84-91.

[7] 陈美华，王延飞. 科技管理决策中的地平线扫描方法应用评析[J]. 情报理论与实践，2017(12):63-68.

[8] McKeown K, Daume III H, Chaturvedi S, et al. Predicting the impact of scientific concepts using full-text features[J]. Journal of the Association for Information Science and Technology, 2016, 67(11): 2684-2696.

[9] Babko-malaya O, Hunter D B, Seidel A C, et al. A method for detection and characterization of technical emergence and associated methods[P]. U.S:Patent Application 15/035,555, 2016-10-06.

[10] Dewey M. Finding patterns of emergence—foresight and understanding from scientific exposition(FUSE)[EB/OL].[2020-11-28].https://www.cendi.gov/presentations/01_09_14_FUSE.pdf, 2014.

（收稿日期：2021-01-11）