

科技文献数据化及组织呈现路径研究^{*}

徐 雷 秦翠玉 李 娇

摘 要 文本型科技文献是当前科学知识表达以及科学交流的主要形态。为了促进科学交流,对日益增长的科技文献中的科学知识进行数据化及组织呈现的研究和实践逐渐增多。本文对科技文献数据化及组织呈现方法、应用场景、实现技术进行了系统梳理,包括科技文献的元数据化、科学词汇抽取、领域实体及其关系识别、篇章功能结构识别、科技文献语义组织以及科技文献呈现与智能化应用六个维度,总结目前该研究领域存在的主要问题;在此基础上设计了科技文献数据化及组织呈现的整体框架,阐述了该框架实现的四个核心技术:识别抽取技术、语义组织技术、分析推理技术以及展陈交互技术;最后归纳总结了该领域面临的挑战,如科学知识自动获取、科学数据质量及信任性、科学知识交互体验等。未来需要加强各方合作,以高质量的科学数据为基础,实现科学知识的叙事生产和转化。图4。表3。参考文献69。

关键词 科技文献 科学知识 数据化 组织呈现 科学交流

分类号 G255.51

Datafication, Organization and Manifestation of Scientific Literature

XU Lei, QIN Cuiyu & LI Jiao

ABSTRACT

In order to facilitate scientific communication, there is an increasing number of research practices that digitize and reorganize scientific knowledge from the growing body of scientific literature. This paper summarizes the main approaches of current scientific knowledge acquisition, including all kinds of academic databases and academic search engines, social network platforms of science, and open access academic platforms, etc. With the convenience of scientific knowledge access, we also face new kinds of scientific communication difficulties, such as quite low efficiency of reading comprehension of large scale of literature.

In order to break this new scientific communication dilemma, datafication, organization and manifestation of scientific literature as the mainstream practices are carried out widely. This paper makes a survey about this research field systematically from aspects of related concepts, application scenarios and implementation technologies. It includes six dimensions, namely meta-datamation of scientific literature,

^{*} 本文系教育部规划基金项目“科学出版物语义组织模式及其实现路径研究”(编号:20A10486066)和武汉大学自主科研项目“图情档领域科学本体构建与应用研究”的研究成果之一。(This article is an outcome of the project “Research on Semantic Organization Mode and Realization Path of Scientific Publications” (No. 20A10486066) supported by the Planning Fund of the Ministry of Education of China and the project “Research on the Construction and Application of Scientific Ontology in LIS” supported by Wuhan University.)

通信作者:徐雷, Email: xlei@whu.edu.cn, ORCID: 0000-0003-4874-3404 (Correspondence should be addressed to XU Lei, Email: xlei@whu.edu.cn, ORCID: 0000-0003-4874-3404)

extraction of scientific vocabularies, recognition of scientific entities and their relations, recognition of discourse function structure, semantic organization of scientific literature, presentation and intelligent application of scientific literature. Recognition and extraction of scientific vocabularies, named entities and their relations, discourse structure such as knowledge units and summary of scientific literature, are the mainstream practices in this research field. A large number of data models and scientific knowledge graphs have been constructed based on scientific literature content, and studies on distributed representation of scientific literature have also been increased gradually. However, the degree of automatic indexing of literature by using current data models needs to be strengthened, and the granularity of scientific knowledge graphs needs to be deepened. In the aspect of intelligent applications about scientific literature, higher intelligent form of literature service platforms has appeared. More and more attentions are paid to this research direction, but it is still not enough to overturn the current form of scientific publications which are mainly based on text narration. There is no unified datafication framework for scientific literature processing, and cognitive differences, different granularity and data organization norms for datafication practice. And there are two directions of datafication and presentation of scientific literature do not fully merge and form a complete closed loop.

Aiming at the main problems existing in this research field, this paper designs a framework of the datafication, organization and manifestation of scientific literature, which realizes the closed-loop from machine-oriented datafication to human-oriented knowledge service, including different granularity of datafication, main forms of data organization and main scenarios. Key technologies supporting this framework, such as recognition and extraction technology, semantic organization technology, analysis and reasoning technology, display and interaction technology of scientific literature content are also elaborated in details.

This paper concludes with a summary of future research and practice directions and potential challenges in this research field, such as automated scientific knowledge acquisition, scientific data quality and trustworthy scientific communication, and interactive experience of scientific knowledge. In the future, it is still necessary to strengthen cooperation among various parties to realize narrative production and transformation of scientific knowledge based on high-quality scientific data. 4 figs. 3 tabs. 69 refs.

KEY WORDS

Scientific literature. Scientific knowledge. Datafication. Organization and manifestation. Scientific communication.

0 引言

科学(Science)一词源于拉丁文“Scientia”,即知识。科学交流意味着知识交流,科技文献(Scientific literature)是当前科学知识的主要载体,承载着知识交流与传播的功能。从15世纪印刷机的发明到20世纪古腾堡电子书计划的实

施,科学知识的形态由印刷出版物演变为数字出版物,科学交流的形式也早从面向少数精英的小众传播发展到面向普通群体的大众传播。数字化的科技文献成为科学知识的主要载体,并呈现出极高的生产效率和传播速度,各类学术数据库、学术搜索引擎以及学术社交平台的涌现进一步提高了科学交流的效率,科学交流形式呈现出高度的数字化、网络化和开放化。

然而,科研人员正面临着一种科学交流的困境:科技文献获取的便利性并不总是意味着科学知识获取的高效性。科学知识被淹没在大量的科技文献之中,如果不借助辅助工具,经过文献检索、排序、过滤等操作,科研人员无法全面把握学科领域的研究发展态势。波士顿大学国际高等教育中心教授 Altbach 和中心主任 de Wit 认为,我们当前面临着大量的学术论文未被阅读,以及大量的同行评议精力被浪费的危机^[1]。为了打破这一困境,目前已有许多有益的实践与探索,诸如通过期刊的影响因子进行期刊分级、抽取科技论文中的知识元、识别科技文献的修辞结构、对科技论文进行文本摘要、构建结构化的科学知识图谱、设计新型出版物类型等,不同的科学实践无疑都直接或间接地为提升科学交流的效率提供了方法与途径。本文将涉及科技文献元数据抽取、知识元识别、内容结构化等实践的过程统称为科技文献的数据化,将数据化状态的科技文献或知识进行再组织再呈现、提供智能化的科学知识服务及应用的过程统称为科技文献的组织呈现。一般而言,数据化是定量分析以及组织呈现的前提,但两者的界限并不绝对。本文首先分析当前科学知识获取途径及科学交流困境,然后从科技文献数据化的不同处理粒度及其语义组织的不同层面,总结归纳实践过程中的应用场景、主要技术方法以及面临的问题和挑战,在此过程中提炼出科技文献数据化及组织呈现的宏观框架与实现路径。

1 当前科学知识获取途径及科学交流困境

1.1 科学知识获取途径

科研人员是科学交流的主体,科学知识如科学实验步骤、科学方法是科学交流的具体对象,而科技文献是科学交流的媒介,具有表征与传播科学知识的作用。除了传统印刷型科技文献外,目前科学知识的获取以各类网络学术搜索引擎以及学术数据库为主^[2],如谷歌学术、微软学术等常见学术搜索引擎及 WoS、Scopus、

PubMed、CNKI 等学术数据库。一般情况下,学术搜索引擎可以在互联网上开放访问,学术数据库则需要机构购买使用权限才能使用。这些学术搜索引擎和数据库由互联网企业、出版商或数据库商经营,具有不同的科技文献类型和覆盖度,并采用不同的收费策略向用户提供科技文献检索服务,包括基本的按文献元数据检索、按不同维度进行结果排序,支持模糊或精确的文献检索与过滤,以及具有各自数据库特色的检索结果可视分析、智能文献推荐、学术资讯订阅等。除了平台资源体量差异外,这些平台的服务区别主要体现在底层资源的数据化程度,这在一定程度上决定了平台的智能化程度。通过这些学术搜索引擎和数据库,科研人员可以获得原始科技文献的题录、关键词、引文等元数据信息以及全文信息,并结合相关文献计量指标分析研究主题的整体发展状况,从而获取宏观的学科知识。这是目前主流的科技文献获取方式,也是图情领域文献计量研究的主要数据来源。

除上述主流获取方式外,第二类科学知识的获取渠道是伴随着社交网络而兴起的,诸如 ResearchGate、Academic.edu、Mendeley、科学网等学术社交网站和学者个人的学术博客,以及目前也被用于分享科研成果或科学活动信息的大众类社交平台,如 Weibo、Twitter 等。这些社交类学术平台具有学者学术信息展示、学术成果推荐与分享以及科学问答与评论等社交功能,进一步将科学交流活动扩展到网络空间,形成了科学研究的网络群体,提高了科学成果的曝光度,提供了发现潜在合作者的网络渠道,提升了科研人员的影响力及科学交流的时空便利性。同时,传统的学术数据库也逐渐出现了社交元素,比如增加了科技文献的社交分享按钮,学术社交平台也因此成为了替代计量学研究的主要数据来源。

此外,还有一类科学知识获取途径是伴随开放科学、开放获取(OA)等运动而出现的开放学术平台,如 DOAJ、PLoS 等开放获取期刊,

arXiv 等预印本数据库,开放评审数据、开放引文数据等各类开放科学数据集以及语义出版实践。开放科学运动主要倡导科学研究成果的开放与科学数据的共享,明显体现出学术界的科学知识数据化意识。

1.2 科学交流困境

毋庸置疑,相对于早期的科学交流形态,当前的科学交流模式极大扩展了科学交流的广度和深度。然而以文本为主要媒介载体的科技文献除了在数字化层面上有所变化外,在内容形式上并没有本质改变,学术数据库中的科学知识、社交网站分享的学术链接仍以 PDF/HTML 等粗粒度的文本类型为主。文本型的知识传播方式,符合人类自然语言叙事习惯,同时借助图表、音视频等多媒体资源,可以进行精美的布局排版,以更好地调动读者的感知系统,进而加强对论文内容的理解。然而,海量的科技文献使这种人工文本阅读理解模式变得极其低效,就当前的科学交流方式而言,具备智能化人机交互的科学知识产品和服务仍没有大规模实现,科研人员面临新的科学交流困境。

同时,出现了一系列正在进行的打破科学交流困境的有益尝试。比如目前大部分学术数据库都提供了科技文献原文的章节结构、参考文献、图片索引、作者及机构等可链接的目录和半结构化的文献内容,以便读者在阅读、下载过程中快速跳转到相应的科学实体,如 Elsevier 基

于 HTML5 保留 PDF 的排版布局优势,但增加了图表的标注与链接,图像可以导出带有信息的图片,表格可以导出 Excel 结构化数据;CNKI 数据库提供了知识元、概念库、方法库等检索途径;WoS 数据库提供了通过化学式结构检索科技文献的方式以及结构化摘要,同时论文或评审专家推荐功能也逐渐成为标配。新的出版物组织形态如纳米出版物、可执行论文等都是对科学知识进行数据化的有益实践,科学知识在若干学科领域已经呈现出数据化、智能化的形态。

2 科技文献数据化及组织呈现应用实践

在科学交流中,人们倾向于将承载科学知识的科技文献如论文、著作等出版物作为科学知识本身,并将其视为科学知识网络,而实际上人们在获取科技文献后还需经过阅读理解、参考查阅、甚至实验复原等若干环节才能真正地掌握科学知识内涵,然后才能进行科学知识的传播与创新。这也表明“科学知识”这一术语在使用过程中会表现为不同粒度,实际的科学交流活动以数据化的科学知识为主,如理论、案例、观点、科学假设、研究方法、实验过程、研究发现等知识元。本文从科技文献数据化与组织呈现的不同粒度视角来梳理目前主要的科学实践及应用场景,如图 1 所示。

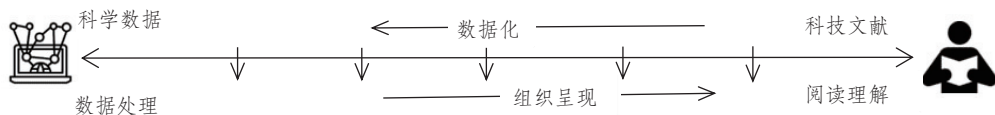


图 1 科技文献数据化及组织呈现循环

图 1 右边是目前主流的科学知识呈现形态,以文本型数字出版物为主,科研人员通过阅读理解活动来获得具体的科学知识;左边是数据化的科技文献或科学知识,可以用数据库、关联数据、知识图谱的形式存储,计算机通过数据

分析处理来理解并呈现科学知识。从右至左的箭头表示科技文献的数据化过程,包括元数据标引、科学词汇抽取、领域实体及其关系识别、篇章修辞结构分割等不同的处理层次和粒度,由左到右的箭头表示对数据化了的科技文献进

行组织呈现的过程,包括常见的科学数据的建设、科技文献的检索推荐与可视化、科技文献自动化摘要、新型科学出版物和科技服务平台的设计等不同场景。向左数据化的过程中进行科学知识挖掘便于计算机索引与分析,向右组织

呈现的过程能满足用户的科学知识理解需求且符合叙事习惯,向下的箭头表示可以在不同的科技文献处理粒度上进行科学知识的组织呈现。下文将从六个维度对目前科技文献数据化及组织呈现应用实践进行分析,如表1所示。

表1 科技文献数据化及组织呈现分析维度

类 型	内 涵
科技文献的元数据化	以科技文献题录、引文、学科主题等文献元数据以及作者、机构、会址等外部科学实体为主,进行数据抽取与组织,以支持文献计量、文献检索等应用。
科学词汇抽取	以科技文献中的关键词、核心术语、科学概念、固定搭配等词汇为主,进行词汇抽取与组织,以支持学科主题识别等应用。
领域实体及其关系识别	以科技文献中的领域实体为主,如算法、疾病、蛋白质等,识别这些实体类型及相互之间的语义关系,以支持科学知识挖掘等应用。
文献篇章功能结构识别	识别句子、段落、章节中包含的修辞结构、论证结构、图表信息,以及其中蕴含的研究方法、学术观点、研究结论等知识元,以支持篇章分析、知识发现等应用。
科技文献语义组织	对上述文献数据化的结果进行语义组织,构建文献数据模型、科学知识图谱等。
科技文献呈现与智能化应用	开发文献检索与推荐应用、对文献数据化结果进行可视化、设计科研辅助功能等智能应用。

2.1 科技文献的元数据化

科技文献的元数据用来描述科技文献的属性信息,包括科技文献的标题、作者、作者机构、发表期刊、年份、国别等题录信息,以及关键词、摘要、主题、参考文献等附属在科技文献正文中的信息,科技文献的元数据化就是对上述信息的识别与抽取的过程。科技文献元数据是目前学术数据库主要的结构化数据类型,是学术数据库提供文献分面检索的主要字段,以及进行宏观文献计量分析的主要数据来源,在诸多方面都具有重要的使用价值,如学术合作模式挖掘、科学评价、科学主题演化分析等应用,这些基于科技文献元数据的分析是目前科研人员获取科技文献以及宏观科学知识的主要来源,也是图情学科研究的主流方向之一。在科技文献越来越容易获取的时代,传统的文献计量等研究已经开始从文献元数据拓展到文献内容层面,如文献正文中的核心词汇、科学概念、科学方法等科研实体,出现了很多基于文献全文的

文献计量方法^[3-5]。

2.2 科学词汇抽取

科技文献中的词汇,尤其是关键词、科学概念、领域术语等词汇,可以反映学科领域信息,常被应用于揭示科技文献的研究主题,开展这类研究的关键环节在于识别并抽取出科技文献中的核心词汇,较常使用自然语言处理技术、深度学习等机器学习技术。科技文献中核心词汇的识别抽取具有广泛的应用场景,是常见的基于全文本的文献计量分析方法之一,识别抽取得到的科学词汇可作为构建学科领域术语词表的参考来源和学科主题差异及演变分析对象。通过词云、共词分析、科学知识图谱等技术构建学科领域词汇网络来表征科学知识,可以达到理解论文主题、发掘学科领域趋势的目的,同时经过索引的科学词汇可以提高科学信息检索与文献内容管理的效率,支持科技文献分面搜索、科学趋势分析和科学知识库构建等任务。

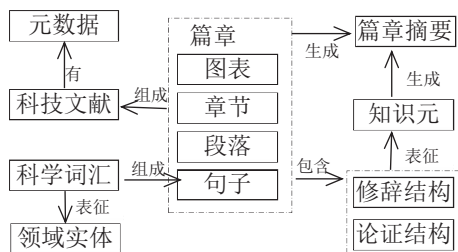
2.3 领域实体及其关系识别

关键词更多的是词汇的语言表现,而词汇所表征的语义实体如组织机构、药物、生物蛋白、计算机程序、算法等具有语义类型,相对于词汇具有更强的领域特性,是更容易被科研人员理解的知识对象,这些语义实体被统称为命名实体^[6],伴随命名实体识别的任务是实体之间的语义关系抽取。实体关系的识别与抽取在科学研究领域具有广泛的应用场景,是学科领域深度知识发现与利用的重要途径。比如,在生物医学领域,通过对药物与药物相互作用、药物与疾病关系、蛋白质与蛋白质相互作用、基因和蛋白质关系^[7]、疾病生物过程描述^[8]、医学概念关系的提取,可让用户在海量的医学文献中发现新的医学知识^[9]、设计新的治疗方案^[10],还可以通过医疗技术实体的识别来发现医学技术发展的趋势^[11];在材料科学领域,基于 Bi-LSTM-CRF 深度学习模型可识别不同类型的命名实体,发现新的分子路径,并辅助材料设计^[12];在计算机领域,可识别科技论文中的算法实体^[13]及算法之间的演化关系^[14];在社会科学领域,通过抽取科学实体可验证媒体报道中的科学新闻是否正确^[15],等等。

2.4 文献篇章功能结构识别

科技论文通过不同的语句、段落、章节、图表等篇章结构及其中的修辞论证关系来叙述科学知识。文献的语段篇章之间具有不同的逻辑功能和语义关系,可用于反映科研人员的工作思路及研究流程,对其进行识别与分析有助于研究人员阅读文献与把握文献逻辑结构。科技文献的篇章组织形式蕴含着文献的修辞结构和论证结构,实践中两者可置于更宽泛的篇章分析任务之下,都可以用来表征科技文献中的知识元,并被应用于科技文献的篇章摘要任务,这些要素之间的关系如图2所示。文献的研究背景、研究对象、研究方法、结果、结论(IMRaD)等元素就是常见的修辞类型^[16],除此之外还有目的、证据、条件等细粒度的修辞元素和关系类

型。论证结构侧重于对逻辑推理和科学论证过程的描述,可包括断言、保证、反驳、支持等论证要素及其论证关系^[17]。相对于语言文字而言,公式、图表、图片甚至经过修饰的章节标题等图形元素是一类特殊的篇章类型,具有直观、高效、更容易被理解的特性,也常被用于科技文献的功能结构表达。鉴于这些元素之间关系复杂且基本都符合篇章粒度,本文将科技文献的知识元识别、图表元素抽取以及篇章摘要等任务统一视为篇章功能结构识别任务,以便于归纳分析。



科技文献中的概念、定义、方法、数据、结论等要素及其中的逻辑关系是科学知识的核心组成部分,也称为知识元。知识元具有不同的类型和粒度,目前有大量关于知识元的研究实践。比如,从科技文献中抽取论文亮点、创新点、研究方法、研究问题、科学定义、知识基因、未来的创新话题,以及从论文标题及摘要数据中抽取学术理论,从科技文献中抽取常见的叙事语言表达模式,等等。除了抽取单一类型的知识元外,多类型知识元的抽取也较多,比如从医学文献中抽取研究目的、研究类别以及疾病来源等信息,从学术文献中抽取研究背景、方法、数据、结果等知识元,从论文标题中抽取词汇对应的修辞功能如方法和研究问题,从科技文献摘要中抽取提议、断言等论证实体以及支持、反对等论证关系,等等。知识元作为科学交流的核心要素,是智能化科学知识服务的关键,具有广阔的应用场景。比如,通过对创新点的识别与分析可以获得期刊载文用词特点等统计规律^[18];通过 PubMed 的论文摘要和结论部分,使用主题

模型和神经网络方法分析论文的结论和之前发表的论文结论是否存在冲突,以辅助论文的评审工作^[19];通过对句子的语法、语义、上下文的编码,抽取论文关键句并辅助生成展示 PPT^[20];从同行评议的内容和结构中挖掘出学术观点,以评估同行评审的效率与质量^[21];通过规则标注人文学科论文中的科学发现、提议、方法等语篇元素,发现基于标注语料的分面检索比基于关键词的文献检索更有效^[22]。

章节图表等形式元素可作为科技文献功能结构的一种可视呈现形态,两者互为表里,如:通过识别科技文献中的图形要素来获取相应的篇章功能元素^[23],通过论文章节标题来识别章节的功能结构^[24],通过抽取生物医学论文中的图表及图题来提供关于分子交互的证据^[25],通过抽取综述类文献中的图表来构建学科领域知识图谱^[26]。纯粹的科技文献图形元素抽取也很常见,比如抽取论文的章节标题、识别论文中的表格和段落块以及论文的逻辑布局结构等信息。此外也出现了大规模的学术文档解析工具,如 CCS 可以识别科技文献的标题、作者、摘要、图标等元素,实现 PDF 文档转换与内容自动重建的功能。

篇章摘要是对篇章内容的概括总结,可以有效帮助用户迅速理解科技文献,Elsevier 甚至要求作者提供论文的图形摘要以方便读者快速阅读^[27]。通常意义的篇章摘要是抽取或生成科技文献的关键句或主题句,这些句子可能对应着科技文献的知识元,如学术观点、科学结论

等,可视为科技文献摘要的一种简化形式。基于科技文献生成篇章摘要是一种典型的科技知识服务,不仅在科学交流领域具有强烈的需求,而且在大众消费、新闻等领域也同样有广泛应用场景。篇章摘要技术既涉及对篇章关键句、逻辑论证结构的识别等数据化过程,其生成的摘要句子又能体现用户的阅读叙事习惯。当前基于科技文献的篇章摘要技术出现了很多较高智能形态的应用场景,比如能生成比传统论文摘要更简短的 TLDR 摘要,进一步满足了读者的阅读理解需求;同时也出现了结合科技文献文本摘要和图片标题进而抽取相关图片生成论文的可视摘要,以及科技论文自动综述等应用。

2.5 科技文献语义组织

元数据、科学词汇、领域实体、篇章结构及其相互之间的关系组合成完整的科技文献内容,对这些要素及其蕴含的科学知识进行语义组织是提升科学交流效率和提供智能化科技服务的基石。本文重点关注科技文献的数据模型、科学知识图谱 (Scientific Knowledge Graph) 构建以及文献的分布式语义表征。科技文献的数据模型是为组织以及揭示科技论文的元数据、词汇、知识元、篇章结构等要素及要素间语义关系而构建的语义组织框架,即对科技文献数据化的结果进行显式组织,进而形成科学知识图谱数据集。文献的分布式语义表征是一种文本隐式表示技术,它和科学知识图谱之间的关系如图 3 所示。

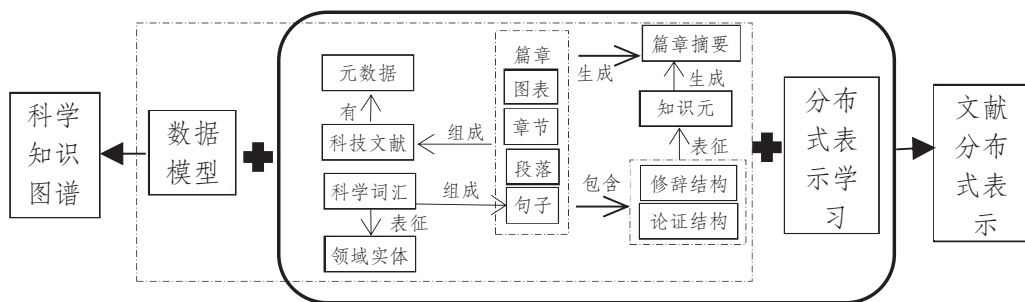


图3 科技文献语义组织方式

目前,关于科技文献的数据模型很多,且具有不同的表示粒度、维度、功能及组织形态,表2列举了部分较为流行的科技文献数据模型。科技文献数据模型的构建是当前知识组织的热门领域之一,未列入表2的研究实践还有:文献细粒度标注本体^[28],包含科学方法和评价结果的科学发现本体^[29],包含论文结构以及学术活动等要素的学术文献标注框架及本体模型^[30],将数据转换为可操作的知识单元的数字对象模

型、研究模型与模型变量本体、知识对象参考本体^[31],研究方法以及理论等要素的知识元本体^[32],引入动态视角的细粒度科学知识表示模型^[33],科技文献的研究设计指纹框架^[34],融合科技文献外部元数据和内部结构特征的知识元模型^[35],以及涉及具体学科的知识本体,如计算机领域的概念本体CSO、数学公式本体、科学事件本体等。

表2 常见科技文献数据模型

数据模型	内容简介
JATS 科技文献标签集	一种国际通用的期刊数据交换文档格式,包含存档交换、期刊出版、文章创作三个标签集。
SPAR 系列本体	涉及出版物引文类型、修辞结构、科学贡献、论文发表流程、科学论证过程等多个主题的语义出版数据模型。
纳米出版物模型	包含声明、出版信息、溯源三个部分,突出细粒度的科学发现的组织及知识溯源。
微型出版物模型	主要对论文中的声明、证据等论证要素进行建模。
Research Objects	主要对科学研究流程及各类科学资源进行建模,并应用在诸多学科领域。
SWAN 篇章本体	包含科学问题、研究陈述、研究人员等科学实体的数据模型。
SALT	包含科技文献图形元素、修辞结构、题录信息等维度的文献标注框架。
SciAnnotDoc	包含科学文献领域概念、元数据、篇章要素以及引用关系四个部分。

经过数据模型填充等语义组织操作后形成的语义化科技文献数据集,称为科学知识图谱。目前科学知识图谱以文献元数据类型为主,内容一般涉及会议、期刊、学术论文、科研项目、引

文信息、作者信息等科技文献的元数据及外部相关实体内容,表3列举了部分科学知识图谱类型。

表3 常见科学知识图谱类型

类 型	科学知识图谱
文献元数据	LOD DBLP、PARDA、Microsoft Academic Graph、Arnetminer SciKG、AceKG 等
开放评议、开放引文数据	PeerRead、OpenCitation 等
科学实体、陈述、知识元等	论文分类陈述数据集、语义网领域和人工智能领域的科学知识图谱(含任务、方法、数据等知识元)、STEM-ECR(研究过程、方法、材料、数据等知识元)、新冠科学知识图谱(含科技文献、专家、疾病、药物等实体)等

与知识图谱显式的符号表示相比,分布式表示是一种隐式表示学习技术,也常用来对科学文本进行表征,它将字符、实体、网络等对象

转换为低维的稠密向量,从而把所有对象表示在同一个语义空间中。根据被表示对象的粒度和类型,分布式表示可应用于词汇、句子、篇章

等粒度级别,目前已有很多相关的分布式表示学习算法模型,分布式表示在命名实体识别、文本分类、序列标注、依存句法等自然语言处理任务中都有出色的表现。科技文献的分布式表示可用于文本相似度检测以及论文推荐、学者推荐、审稿人及学术会议推荐、作者合作及论文影响力预测、文章摘要生成等知识服务场景中。

2.6 科技文献呈现与智能化应用

科技文献与科学知识的呈现形态直接影响用户的阅读体验和知识理解效率。当前的科技文献以印刷文本以及 PDF/HTML 等格式的数字出版物为主要形式,数字化文献可借助特定的软件工具或设备进行呈现,科技文献的数据化与语义组织则为智能化科学知识服务奠定了基础。经过不同程度的数据化及语义组织的科技文献,可提供不同智能水平的科学知识服务,如智能化的文献检索与推荐、知识可视化、论文写作智能辅助等。

文献检索是获取科学知识的前提,除了主流的学术数据库和学术搜索引擎外,也出现了一些科技论文智能检索系统或智能检索功能。Semantic Scholar^① 学术搜索引擎基于深度学习技术,从论文中挑选出最重要的关键词和短语用于判断文章所论述的主题,亦可从论文中抽取图表呈现在检索结果中,也可以辨别一篇文章引用的参考文献是否具有重要的参考价值,并用于评价论文的学术影响力;AMiner^② 通过多个维度呈现学者的学术画像,可用于学者推荐及论文写作,以及快速生成论文内容描述视频。这些平台都是以数据化的科学知识为支撑,类似的系统还包括结合科学知识图谱和文本分布式表示的学术论文检索系统^[36]、结合本体技术的 PubMed 文献搜索引擎 ANDDigest^[37]、基于论文表格数据开发的学术问答系统^[38],以及科技文献 AI 引擎 SciAEngine^[39] 等。与科技

文献相对应,科学视频 (Scientific Video) 作为一种新的形式被提出^[40,41],这种视频是一种带标注的“研究视频”,作为科学知识新的呈现载体,目前被应用于科学实验、艺术设计等诸多研究领域。

2.7 科技文献数据化及组织呈现应用实践存在的问题

“数据作为一项基础设施”的观念形成已久,数据化为智能化的科学知识服务提供了可能,通过系统梳理近期科技文献数据化及组织呈现应用实践的研究,发现还存在以下问题。①科学研究中对科学知识的数据化表达与组织渐成主流,但还不足以颠覆目前以文本叙事为主的科技论文基本形态,且学科实践主体间的差异较大。生命科学、计算机科学、化学等“硬学科”在知识元识别、科学知识图谱构建等方面仍占有较大的体量,但伴随着数字人文、计算传播等新文科的兴起,这种状况会有所改变。②基于全文本的科学计量研究逐渐受到重视,深度学习等新技术越来越多地被应用在科技文献数据化过程中,科技文献数据模型的构建研究层出不穷,这些实践俨然是热门研究方向。但科技文献的数据化实践还没有形成统一的数据化框架,数据化形态仍存在不同的认识差异、粒度和数据组织规范,数据化实践的算法也存在较多的领域局限性。③当前面向机器的科技文献的数据化及组织研究较多,面向人的智能化科学知识服务应用也逐渐受到重视,但这两个方向并没有充分地融合并形成完整的闭环,更多的研究实践往往偏重一端且具有技术、学科等领域限制。常见的情形是,开发出各类自动化的科学知识抽取算法、科学知识图谱等成果后就戛然而止,比如链接开放数据云 (LOD) 中的出版物数据集一直没有有效利用起来^[42],没有形成持续有效的支撑后续科学服务场景的闭环。

① <https://www.semanticscholar.org>

② <https://www.aminer.cn>

3 科技文献数据化及组织呈现框架与核心技术

3.1 整体框架

鉴于当前的相关应用实践存在前文所述的

问题,本文依据现有实践经验,从科技文献数据化及组织呈现两个方面归纳出该过程涉及的核心环节,设计科技文献数据化及组织呈现整体框架,在宏观的科学交流视野下明确当前应用实践的定位,并在框架的核心技术上提供相应的实践参照,如图4所示。

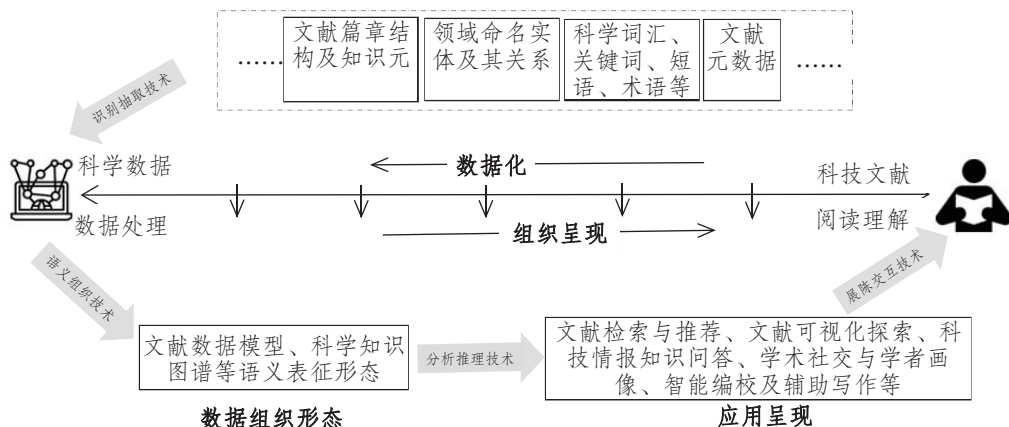


图4 科技文献数据化及组织呈现整体框架

在图4中,科技文献数据化过程有不同的粒度,从文献元数据到篇章结构,并对应着不同的数据组织形态,支持不同的科学服务应用场景。一般而言,数据化的维度越全面、粒度越细、关联度越高,在此基础上提供科学知识的重组应用和服务就越灵活,对提升用户阅读理解等知识获取活动的效果和体验的支持程度就越高,相应地投入其中的研发成本也较高;反之,用户接触传统印刷型文献的阅读体验就越多,相应的用户阅读理解成本就越高。该框架实现了从面向机器的原始科技文献到细粒度的科学知识,再到面向人的支撑智能化的科学知识服务的闭环,通过科技文献的识别抽取技术、语义组织技术、分析推理技术以及展陈交互技术实现了从科学数据到用户阅读需求的映射。

3.2 核心技术

(1) 识别抽取技术。这是科学知识组织的前提,具体可使用文本挖掘、机器学习、深度学习、自然语言处理等技术,从科技文献中识别抽

取不同粒度的知识,比如元数据、科学词汇、领域实体、研究问题、研究方法等知识元以及图表、公式和文摘等篇章结构。按照被抽取的科学知识的粒度划分,目前包括以下四种识别抽取技术。

①科技文献元数据抽取技术,主要涉及元数据抽取规则的制定、作者及机构名等命名实体的识别与消歧^[43]、文献计量指标的设计以及科技文献元数据框架的设计与组织^[44]等技术。在科技文献元数据组织标准方面,除了传统的都柏林核心元数据之外,各学术数据库一般都有自己的科技文献元数据组织标准,这些标准大同小异,可通过 Crossref 等平台并借助数字对象标识符(DOI)等技术实现科技文献的广泛映射互联与访问,在提升科技文献访问存取便利性的同时,也为自动化的科技文献元数据抽取与宏观的科学计量统计研究提供了便利。

②科学词汇识别抽取技术。传统自然语言处理技术如分词、词性标注等,通过构造词汇的语法句法等语言表现特征来进行科学词汇的初

步识别,进而再结合词汇的上下文信息、语义结构信息等统计特征进行关键词、科学术语或主题词的筛选识别^[45]。从具体技术上看,科技文献中关键词的抽取,大致可以分为有监督与无监督的方法^[46],由于无监督方法不需要高昂的人工语料标注成本,所以在实际使用中具有较强的适应性,比如经典的 TF-IDF 指标、TextRank 等算法常用于科学文本中关键词的抽取。深度学习是目前科技文献关键词抽取技术的主流,这类技术能够自动学习词汇的抽象语言特征,目前已经在关键词提取方面达到了较高的水平,如融合双向长短时记忆神经网络和条件随机场技术的 Bi-LSTM-CRF 模型、基于双向编码器的 BERT 预训练语言模型、学习排序模型、主动学习技术、基于词嵌入的随机游走方法、结合 PageRank 及词嵌入技术的方法、结合句法依存及语义依存图结构信息的深度学习方法、序列标注等方法,都已被应用在科技文献关键词的抽取过程中,并取得了很好的效果。

③实体及其关系识别抽取技术。从科技文献中抽取实体及其关系,目前主要采用自然语言处理、机器学习等方法,传统的实体识别与关系抽取一般采用流水线(pipeline)的方式,即先进行实体识别,然后对识别的实体进行分类、关系抽取^[25]。但由于这种方式容易忽略任务之间的相关性,因而出现了使用联合模型来同时进行实体与关系抽取的研究^[47]。同时,基于 GRU 的多任务学习框架、MAT-LSTM 模型、BERT-CRF 模型、基于知识图谱及句子嵌入特征的模型等深度学习方法,以及结合自然语言处理技术的 TechMiner 算法等,也被用于科技文献实体识别与关系抽取任务。

④篇章要素识别抽取技术。自然语言处理技术常被用于科技文献篇章要素的识别,如使用依存句法分析抽取科技文献中常见的叙事语言表达模式、科学问题和对应的解决方案等。基于规则的技术、序列标注模型、深度学习技术也大量地被应用在科技文献的修辞类型、论证要素以及章节图表的抽取上,修辞结构理论同

样也常被应用于篇章分析的过程中。在科技文献自动摘要技术方面,主要分为基于文摘生成的摘要技术、基于引文的摘要技术以及面向特定语篇类型的摘要技术^[48]。基于文摘生成的摘要技术通过筛选论文不同章节的主题句来生成全文摘要^[49],如自动生成引文的描述文本的摘要技术、从多个科技文献中生成文摘的多文档摘要技术。基于引文的摘要技术一般先收集引用该论文的文献集合,然后抽取这些文献中的引用句,并通过文本聚类、过滤等技术来生成候选摘要,比如可使用这类技术进行自动文献综述^[50];基于引文的摘要技术还可以融合其他信息进行文献摘要,如结合引文及文献内容影响力、多目标聚类技术、博弈论理论的科技文献摘要方法等。面向特定语篇类型的摘要技术则针对科技文献的不同章节,如论文中的图表、研究进展等部分,设计特定的摘要算法。

(2)语义组织技术,包括元数据、关联数据、本体、知识图谱等组织技术^[51]以及基于深度学习的分布式语义表征技术,主要被用于设计科学知识数据模型、构建科学知识图谱等场景,并满足网络环境下多源异构、多粒度的科学知识的组织与知识融合需求。科学知识数据模型以本体构建技术最为常见,除了常规的基于专家领域分析的本体构建方法外,具体过程还涉及本体匹配、本体学习、知识图谱补全等技术。科学知识图谱的构建一般会用到科技文献数据化过程中的技术,比如:使用基于规则的方法从科技文献以及其他多源异构数据中构建造血干细胞癌症治疗方案的科学知识图谱^[52];使用 LSTM 神经网络模型抽取论文元数据及学术实体及关系来构建科学知识图谱^[53];将科学实体、关系及指代消解任务视为分类任务,使用 Bi-LSTM 神经网络模型来构建军事科学知识图谱^[54];使用循环神经网络序列标注模型,从生命科学论文摘要中获取科学事实及其成立的条件来构建科学知识图谱^[55]等。在科技文献分布式表征上,较为典型的是 SciBERT 模型^[56],使用该模型在计算机科学和生物医学文献语料上训练了一个

科技文献语义表示预训练模型,常被应用在科技文献处理的相关任务上。除此之外,还有融合引文网络和文本信息的文献表示方法^[57]、结合文献结构及 Doc2Vec 算法的文献表示技术^[58]等。同时,为了提升这些语义科学数据集的可发现性、可溯源性、可交互性和可重复性等特性,当前科学数据的开发与组织倡导遵循 FAIR 原则^[59],与之相关的技术包括数字对象标识符(DOI)、持久访问技术、数据溯源技术等。

(3) 分析推理技术。无论科技文献数据化为哪种粒度,都需要通过进一步的数据分析以及结合领域背景知识进行推断定论,才能产生可验证、可解释且有创新价值的科学发现,并支持智能化科学知识服务。分析推理技术基本可以被用在科技文献数据化的各个阶段,也正因为如此,分析推理技术常常依附在数据化技术之中,比如通常在识别抽取任务之后就马上进行一定的分析推理任务,再比如句法分析等自然语言处理技术本身就带有分析推理的特性。分析推理技术是衔接数据化与展陈交互过程的桥梁,可以是常规的统计分析、文献计量、文本分析、网络分析等技术,被用于学科领域主题和科技趋势分析与预测等宏观科学知识的发掘;也可以是本体推理、语义查询等技术,被用于科学知识图谱的网络结构分析推断。此外,还有前文所述的篇章分析技术,可用于分析科技文献逻辑结构、观点情感,以及摘要式阅读与领域报告自动生成等场景。

(4) 展陈交互技术。展陈交互技术是一种将用户的感知转换为认知的手段,被用于科学数据产品化与服务化的过程中。这类技术包括文献检索与推荐技术、可视化技术、数据叙事技术以及 VR/AR 等人机交互技术,直接对接用户需求体验场景,支撑科技文献检索与动态交互、领域专家推荐、学术社交与学者画像、科技情报知识问答、智能编校及辅助写作等应用。其中,可视化技术的应用较为普遍,比如:ReVealD 使用关联数据和可视化技术帮助非专业人员实时探索生物医学领域的数据;VIStory 可视化系统

通过抽取、探索论文图表等图形元素,以叙事可视化方式展示论文章节长度、引文位置等信息,以发现论文的写作、用词、引用规律;ConfVisExplorer 被设计为一个用于学术会议主题可视化及计算会议成熟度的智能平台,用于科技主题会议信息的抽取和整理。在科技论文阅读评审及写作辅助方面,有基于深度学习技术的在论文写作中实时辅助搜寻相关文献的编辑器^[60]、可以在论文中执行实验程序的“可执行论文”平台^[14]、基于纳米出版物模型的科技论文评审工具^[61]、支持数学公式阅读理解的工具^[62]等。此外还有一些支持科技论文语义化写作的工具,如支持作者以及读者等多方的去中心化的论文写作工具 Dokieli、支持语义标引的 HTML 论文写作平台 DOME0 和 RASH,等等。

纵观所有的技术实践,针对科技文献识别抽取技术的研究最多,其中关键词抽取以及领域命名实体识别与关系抽取是目前科学研究和实践的热门领域之一,SemEval 会议常年都设置有学术文献关键词、学术术语以及命名实体识别与语义关系抽取的任务^[63,64],这也反映出科技文献的数据化技术相对于科技文献的展陈交互要更为基础,更受科研人员的关注。这类基础性技术是构筑以科学数据为核心的科学交流基础设施的核心,也是科技文献深度语义组织以及基于文献的知识发现与科学创新的主要技术手段。在完成识别抽取后,语义组织技术被用于识别抽取任务结果的结构化组织,分析推理技术可在此基础上进一步发掘新知识,所有阶段的结果都可借助展陈交互技术进行展示。在上述所有技术中,深度学习等人工智能技术在各项任务及应用场景中表现出较好的性能,目前已开始在科技文献数据化及组织呈现的各个阶段中广泛应用。实践中深度学习技术常结合传统特征工程、语法等领域规则、领域知识图谱等信息来提升相关任务的性能,以更好地支撑后续任务或应用,这表明科学知识自动发现、科技文献数据化的结果本身也可以反过来促进相关任务技术性能的提升。融合领域规则、知

识图谱等显性化科学知识的人工智能技术是当前及未来科技文献智能处理的重要发展方向。

4 科技文献数据化及组织呈现面临的挑战

综合当前的研究和实践来看,科技文献的数据化及组织呈现主要面临三方面的挑战。

(1) 科学知识自动获取技术的挑战。从科技文献中获取结构化科学知识是科技文献数据化的核心过程,当前科学知识获取应用文本分析、深度学习等技术,呈现出高度的智能化特征,但科学知识的自动获取技术仍然存在以下问题。①不同学科对科技文献的数据化有不同的认知及需求,目前自动获取技术难以被应用在所有学科。计算机、生物医学等学科应用较多,比如纳米出版物主要来自生命科学领域,且是从现有的数据库中进行抽取与转换。相关实践研究所使用的方法和技术具有一定的领域局限和移植困难,比如文学与生物医学领域的文献数据化差异就很大,完全进行算法移植可能并不有效。②技术本身具有一定的精度,不能完全无指导地获取科学知识,一般都需要人工校对、标注后才能形成科学知识的黄金数据集,完全自动化的科学知识获取仍不现实。如有研究通过随机抽取科技文献中的句子,使用目前可以免费获取的开放信息抽取工具抽取句子中的三元组,和人工标注的结果进行对比,发现这些工具在科技文献中的应用效果并不好,并且发现存在缺少结果的报告、缺少方法的细节描述、对使用的数据解释模糊、缺少对各类技术的交叉验证等问题^[65]。因此,SciencePod^①等科学内容营销平台主要依赖专家团队对科技文献进行专业解读并形成专业的科学文摘,以白皮书、专家访谈、信息图等形式,向客户提供定制化的科学知识服务,而不是完全依赖自动化算法。此外,在科学知识的组织方面,尽管已有很多论

文标引数据模型,但因其维度、建模语言、抽象程度具有不同的复杂性,在实践应用中仍主要以手工标注为主。因此,科学知识的自动获取仍然是该领域的一个重要挑战,为了提升算法的通用性和效率,需要处理好机器与人工、效果与效率之间的关系,兼顾好计算成本和人工标引成本。

(2) 科学数据质量及信任性的挑战。科技文献数据化是一个数据增值的过程,但数据化本身可能会带来信息丢失、事实扭曲等质量问题。科学数据的质量是影响用户科学知识理解体验的最关键因素,不准确的数据化科学知识会严重影响科研人员的科学交流体验,比如不能复现科学结果等。数据化的科学知识的质量问题客观上主要是由科学知识获取技术本身引起的,比如抽取的知识元不完整、科学数据在融合过程中存在误配失配等问题。同时,科学知识图谱在实际使用中也可能存在链接映射不完整、查询端不可用等问题。另外,原始的科技文献可通过同行评审来保证科学发现的质量,而目前数据化过程则缺少这个质量保障环节,因此科研人员主观上对经过数据化处理的科技文献产生的科学知识会形成不信任的质量认知^[66]。因此,虽然当前学术数据库、学术搜索引擎流行,但智能文献分析等科学服务只能作为学术交流辅助功能。针对科学数据的信任性问题,Kuhn 和 Dumontier 认为,与其面对科技文献发表后再进行识别抽取而带来的质量问题,不如在科技论文创作过程中就自动编辑成细粒度的语义知识^[67];Moreau 和 Groth 通过设计数据溯源框架来记录数据处理的每一个操作步骤,以保证科学数据更新的可追溯^[68];Ramachandran 和 Kantarcioglu 使用区块链以及永久标识技术来保证科学数据处理过程的不可篡改^[69];也有使用容器技术、Jupyter 技术实现科学结果的快速复现。当前,这一领域仍处于初步探索阶段,科学数据的质量评定仍然是科学数据应用前需要重

① <https://www.sciencepod.net>

视的环节,如何解决科学知识生产以及转换过程中的信任性问题仍然极具挑战性,科学数据出版任重道远。

(3) 科学知识交互体验的挑战。科学知识的展陈交互方式是影响用户阅读理解体验的最后一个环节,前文提到了文献的检索与推荐、科学知识问答、可视化探索等交互场景,以及自动文摘、可视化摘要、科学视频等展陈方式,这些场景及方式通过语义标引等方法降低了读者理解科学知识的难度,并提供了多种动态交互机制。然而就当前的实践研究来看,科技文献的数据化与科学知识的交互体验过程仍没有完全对接,在科学知识的呈现与感知界面设计以及提升科研人员对科学知识的认知理解效率等方面仍面临巨大挑战。通过前文总结发现,大部分的研究主要以算法模型的开发为主,而数据化的结果比如抽取的知识元以及构建的科学知识图谱,一般是作为验证算法模型有效性的副产品,科学知识成了静态的语料库,科学数据、智能算法等开发完毕后有相当大的比例没有对接到具体的应用场景,数据化的科学知识映射到用户体验环节有所不足,科技文献数据化与科学知识交互体验等研究领域仍有很大的合作空间,未来仍需要多方参与、共同解决科学知识的抽取、组织、溯源、展陈等问题。我们需要更多地关注用户的阅读理解体验,智能感知用户需求,设计适配不同终端设备、不同身份科研人员的科学知识呈现与管理界面,对科学知识进行按需动态重组,深入研究科学数据叙事的核心要素及实践路径,在传统自然语言叙事习惯和数字叙事方式之间做出权衡,设计出新型的

科学出版物及科学知识交互模式,探索新型的科学成果发表与传播机制。

5 结语

科学交流的形式早已完成了从传统印刷出版到数字化的转变,当前正处于数字化向数据化转变的时期。科研人员早已意识到科学知识数据化的重要性,为实现科技文献的文本叙事形态向数据叙事形态的转变进行了大量的研究与实践。本文从科技文献粒度化的视角,分析了科技文献在数据化及组织呈现过程中的应用场景、主要技术以及面临的挑战,归纳设计了科技文献数据化及组织呈现的路径框架。纵观当前实践,科学交流从一维线性、固定的文本叙事为主的交流形态发展到多维动态的数据叙事形态,未来新型的科学交流仍有巨大的突破空间。科学知识获取的自动化、智能化、通用化仍需要加强,科技文献数据化过程中的质量问题以及信任性问题仍需要持续关注,科研人员的知识需求感知与满足以及科学数据产品和服务的开发仍需要大量的跨界合作实践。在未来持续演化的科学环境中,作者、编辑、评审、读者、出版社、技术公司等多方仍需共同努力,打破文本叙事难以适应科学大数据环境下的科学知识获取与传播困境,让未来的科学交流载体更好地发挥支撑科学发现与科学论证的作用,更好地支持科学创作与评审等科学交流活动,在发挥机器优势的同时重视用户体验,实现高质量的、数据化的科学知识的叙事生产与转变。

参考文献

- [1] Altbach P G, de Wit H. Too much academic research is being published[J]. International Higher Education, 2019 (96): 2-3.
- [2] Kim S, Chung E, Lee J Y. Latest trends in innovative global scholarly journal publication and distribution platforms[J]. Sci Ed, 2018, 5(5): 100-112.
- [3] Ding Y, Zhang G, Chambers T, et al. Content-based citation analysis: the next generation of citation analysis[J].

- Journal of the Association for Information Science & Technology, 2014, 65(9):1820–1833.
- [4] Zhao H, Luo Z, Feng C, et al. A context-based framework for modeling the role and function of on-line resource citations in scientific literature[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019:5206–5215.
- [5] Pan X, Yan E, Wang Q, et al. Assessing the impact of software on science: a bootstrapped learning of software entities in full-text papers[J]. Journal of Informetrics, 2015, 9(4):860–871.
- [6] Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA, 2018:2145–2158.
- [7] Lamurias A, Clarke L A, Couto F M. Extracting microRNA–gene relations from biomedical literature using distant supervision[J]. PloS One, 2017, 12(3):e0171929.
- [8] Valenzuela-Escúrciga M A, Babur Ö, Hahn-Powell G, et al. Large-scale automated machine reading discovers new cancer-driving mechanisms[J]. Database, 2018:98.
- [9] Bugnon L A, Yones C, Raad J, et al. DL4papers: a deep learning approach for the automatic interpretation of scientific articles[J]. Bioinformatics, 2020, 36(11):3499–3506.
- [10] Hong L, Lin J, Li S, et al. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories[J]. Nature Machine Intelligence, 2020, 2(6):347–355.
- [11] Gao H, Gui L, Luo W. Scientific literature based big data analysis for technology insight[J]. Journal of Physics: Conference Series, 2019, 1168(3):032007.
- [12] Hong Z, Tchoua R, Chard K, et al. SciNER: extracting named entities from scientific literature[C]//20th International Conference on Computational Science. Amsterdam, Netherlands, 2020:308–321.
- [13] Wang Y, Zhang C. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing[J]. Journal of Informetrics, 2020, 14(4):101091.
- [14] Zha H, Chen W, Li K, et al. Mining algorithm roadmap in scientific publications[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, AK, USA, 2019:1083–1092.
- [15] Hoque M R U, Bradley D, Kwan C, et al. Searching for evidence of scientific news in scholarly big data[C]//Proceedings of the 10th International Conference on Knowledge Capture. California, USA, 2019:251–254.
- [16] Nair P K R, Nair V D. Scientific writing and communication in agriculture and natural resources[M]. Cham, Switzerland: Springer, 2014:13–25.
- [17] Lippi M, Torroni P. Argumentation mining: state of the art and emerging trends[J]. ACM Transactions on Internet Technology (TOIT), 2016, 16(2):1–25.
- [18] 曹树金, 闫欣阳, 张倩, 等. 中外情报学论文创新性特征研究[J]. 图书情报工作, 2020, 64(1):80–92. (Cao S J, Yan X Y, Zhang Q, et al. Research on characteristics of innovation in Chinese and international academic literature of information science[J]. Library and Information Service, 2020, 64(1):80–92.)
- [19] Pinto J M G, Balke W T. Assessing plausibility of scientific claims to support high-quality content in digital collections[J]. International Journal on Digital Libraries, 2020, 21(1):47–60.
- [20] Sefid A, Wu J, Mitra P, et al. Automatic slide generation for scientific papers[C]//SciKnow@K-CAP, Marina Del Rey, CA, USA, 2019:11–16.
- [21] Hua X, Nikolov M, Badugu N, et al. Argument mining for understanding peer reviews[EB/OL]. [2022-05-07]. <https://arxiv.org/abs/1903.10104>.

- [22] Ribaupierre H, Falquet G. Extracting discourse elements and annotating scientific documents using the SciAnnot-Doc model: a use case in gender documents [J]. *International Journal on Digital Libraries*, 2018, 19(2-3): 271-286.
- [23] 丁培, 叶兰. 科技文献中学术图表标注研究进展 [J]. *现代情报*, 2021, 41(4): 165-177. (Ding P, Ye L. Research progress of academic figures and tables annotation in scientific literature [J]. *Journal of Modern Information*, 2021, 41(4): 165-177.)
- [24] 秦成磊, 章成志. 基于层次注意力网络模型的学术文本结构功能识别 [J]. *数据分析与知识发现*, 2020(11): 30-46. (Qin C L, Zhang C Z. Using hierarchical attention network model to recognize structure functions of academic articles [J]. *Data Analysis and Knowledge Discovery*, 2020(11): 30-46.)
- [25] Burns G, Shi X, Wu Y, et al. Towards evidence extraction: analysis of scientific figures from studies of molecular interactions [M]// *Emerging Topics in Semantic Technologies*. Amsterdam, Netherlands: IOS Press, 2018: 95-102.
- [26] Oelen A, Stocker M, Auer S. Creating a scholarly knowledge graph from survey article tables [M]// *Digital Libraries at Times of Massive Societal Transition*. Cham, Switzerland: Springer, 2020: 373-389.
- [27] Graphical abstracts [EB/OL]. [2022-05-07]. <https://www.elsevier.com/authors/tools-and-resources/graphical-abstract>.
- [28] 马翠婷, 曹树金. 网络学术文档细粒度聚合本体构建研究 [J]. *图书情报工作*, 2019, 63(24): 107-118. (Ma C C, Cao S J. Study on the construction of fine-grained aggregation ontology for academic documents in the Internet environment [J]. *Library and Information Service*, 2019, 63(24): 107-118.)
- [29] Fathalla S, Vahdati S, Auer S, et al. SemSur: a core ontology for the semantic representation of research findings [J]. *Procedia Computer Science*, 2018, 137: 151-162.
- [30] 孙建军, 裴雷, 蒋婷. 面向学科领域的学术文献语义标注框架研究 [J]. *情报学报*, 2018, 37(11): 1077-1086. (Sun J J, Pei L, Jiang T. Research on semantic annotation in academic literature [J]. *Journal of the China Society for Scientific and Technical Information*, 2018, 37(11): 1077-1086.)
- [31] Flynn A J, Friedman C P, Boisvert P, et al. The knowledge object reference ontology (KORO): a formalism to support management and sharing of computable biomedical knowledge for learning health systems [J]. *Learning Health Systems*, 2018, 2: e10054.
- [32] 索传军, 盖双双. 知识元的内涵、结构与描述模型研究 [J]. *中国图书馆学报*, 2018, 44(4): 54-72. (Suo C J, Gai S S. The connotation, structure and description model of knowledge unit [J]. *Journal of Library Science in China*, 2018, 44(4): 54-72.)
- [33] 王忠义, 夏立新, 李玉海. 基于知识内容的数字图书馆跨学科多粒度知识表示模型构建 [J]. *中国图书馆学报*, 2019, 45(6): 50-64. (Wang Z Y, Xia L X, Li Y H. Construction of interdisciplinary multi-granularity knowledge representation model in digital library based on knowledge content [J]. *Journal of Library Science in China*, 2019, 45(6): 50-64.)
- [34] 钱力, 张晓林, 王茜. 基于科技文献的研究设计指纹描述框架研究 [J]. *大学图书馆学报*, 2015, 33(1): 14-20. (Qian L, Zhang X L, Wang Q. Research design fingerprint description framework base on scientific papers [J]. *Journal of Academic Libraries*, 2015, 33(1): 14-20.)
- [35] 秦春秀, 杨智娟, 赵捧未, 等. 面向科技文献知识表示的知识元本体模型 [J]. *图书情报工作*, 2018, 62(3): 94-103. (Qin C X, Yang Z J, Zhao P W, et al. The knowledge element ontology model of scientific literature for knowledge representation [J]. *Library and Information Service*, 2018, 62(3): 94-103.)
- [36] Mai G, Janowicz K, Yan B. Combining text embedding and knowledge graph embedding techniques for academic search engines [C]// *Semdeep/NLIWoD@ISWC*. Monterey, CA, USA, 2018: 77-88.

- [37] Ivanisenko T V, Saik O V, Demenkov P S, et al. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature[J]. BMC Bioinformatics, 2020, 21(11): 1–21.
- [38] Jaradeh M Y, Stocker M, Auer S. Question answering on scholarly knowledge graphs[C]//International Conference on Theory and Practice of Digital Libraries. Lyon, France, 2020: 19–32.
- [39] 张智雄, 刘欢, 于改红. 构建基于科技文献知识的人工智能引擎[J]. 农业图书情报学报, 2021, 33(1): 17–31. (Zhang Z X, Liu H, Yu G H. Building an artificial intelligence engine based on scientific and technological literature knowledge[J]. Journal of Library and Information Science in Agriculture, 2021, 33(1): 17–31.)
- [40] Xu S, Yu H, Hemminger B M, et al. Who, what, why? An exploration of JoVE scientific video publications in tweets[J]. Scientometrics, 2018, 117(2): 845–856.
- [41] Lösel G. Tags and tracks and annotations—research video as a new form of publication of embodied knowledge[J]. International Journal of Performance Arts and Digital Media, 2021, 17(1): 31–45.
- [42] Debattista J, Lange C, Auer S, et al. Evaluating the quality of the LOD cloud: an empirical investigation[J]. Semantic Web, 2018, 9(6): 859–901.
- [43] Xu J, Shen S, Li D, et al. A network-embedding based method for author disambiguation[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy, 2018: 1735–1738.
- [44] Lipinski M, Yao K, Breiting C, et al. Evaluation of header metadata extraction approaches and tools for scientific PDF documents[C]//Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. Indianapolis, IN, USA, 2013: 385–386.
- [45] Patel K, Caragea C, Wu J, et al. Keyphrase extraction in scholarly digital library search engines[C]//27th International Conference on Web Services. Honolulu, HI, USA, 2020: 179–196.
- [46] Hasan K S, Ng V. Automatic keyphrase extraction: a survey of the state of the art[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland, USA, 2014: 1262–1273.
- [47] Luo L, Yang Z, Cao M, et al. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature[J]. Journal of Biomedical Informatics, 2020, 103: 103384.
- [48] Altmami N I, Menai M E B. Automatic summarization of scientific articles: a survey[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(4): 1011–1028.
- [49] Erera S, Shmueli-Scheuer M, Feigenblat G, et al. A summarization system for scientific documents[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Hong Kong, China, 2019: 211–216.
- [50] Portenoy J, West J D. Supervised learning for automated literature review[C]//BIRNDL@SIGIR. Paris, France, 2019: 83–91.
- [51] Vivar J O, Segarra J, Villazón-Terrazas B, et al. REDI: towards knowledge graph-powered scholarly information management and research networking[J]. Journal of Information Science, 2022, 48(2): 167–181.
- [52] 胡正银, 刘蕾蕾, 代冰, 等. 基于领域知识图谱的生命医学学科知识发现探析[J]. 数据分析与知识发现, 2020, 4(11): 1–14. (Hu Z Y, Liu L L, Dai B, et al. Discovering subject knowledge in life and medical sciences with knowledge graph[J]. Data Analysis and Knowledge Discovery, 2020, 4(11): 1–14.)
- [53] Ammar W, Groeneveld D, Bhagavatula C, et al. Construction of the literature graph in semantic scholar[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). New Orleans, Louisiana, USA, 2018: 84–91.
- [54] Hu R, Li Z, Li J. Research on entity coreference resolution technology oriented to military knowledge graph[J].

- Journal of Physics: Conference Series, 2020, 1624(5):052003.
- [55] Jiang T, Zhao T, Qin B, et al. The role of “Condition” a novel scientific knowledge graph representation and construction model[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, AK, USA, 2019:1634–1642.
- [56] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019:3615–3620.
- [57] Wang B, Xu T, Wang H, et al. Author contributed representation for scholarly network[C]//Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Tianjin, China, 2020:558–573.
- [58] Lu Y, Zhai Y, Luo J, et al. MLPV: text representation of scientific papers based on structural information and Doc2vec[J]. American Journal of Information Science and Technology, 2019, 3(3):62.
- [59] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3(1):1–9.
- [60] Gökçe O, Prada J, Nikolov N I, et al. Embedding-based scientific literature discovery in a text editor application [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020:320–326.
- [61] Bucur C I, Kuhn T, Ceolin D. A unified nanopublication model for effective and user-friendly access to the elements of scientific publishing[C]//International Conference on Knowledge Engineering and Knowledge Management. Bolzano, Italy, 2020:104–119.
- [62] Head A, Lo K, Kang D, et al. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021:1–18.
- [63] Augenstein I, Das M, Riedel S, et al. SemEval 2017 Task 10: ScienceIE-extracting keyphrases and relations from scientific publications[C]//SemEval@ ACL. Vancouver, Canada, 2017.
- [64] Gábor K, Buscaldi D, Schumann A K, et al. Semeval-2018 task 7: semantic relation extraction and classification in scientific papers[C]//Proceedings of the 12th International Workshop on Semantic Evaluation. New Orleans, Louisiana, USA, 2018:679–688.
- [65] Groth P, Lauruhn M, Scerri A, et al. Open information extraction on scientific text: an evaluation[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA, 2018:3414–3423.
- [66] Coelho R, Braga R, David J M, et al. Blockflow: trust in scientific provenance data[C]//Anais do XIII Brazilian e-Science Workshop. SBC, 2019.
- [67] Kuhn T, Dumontier M. Genuine semantic publishing[J]. Data Science, 2017, 1(1–2):139–154.
- [68] Moreau L, Groth P. Provenance: an introduction to PROV[J]. Synthesis Lectures on the Semantic Web: Theory and Technology, 2013, 3(4):1–129.
- [69] Ramachandran A, Kantarcioglu M. Smartprovenance: a distributed, blockchain based dataprovenance system [C]//Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. Tempe, AZ, USA, 2018:35–42.

徐 雷 武汉大学信息管理学院副教授。湖北 武汉 430072。

秦翠玉 武汉大学语义出版与知识服务重点实验室硕士研究生。湖北 武汉 430072。

李 娇 武汉大学语义出版与知识服务重点实验室硕士研究生。湖北 武汉 430072。

(收稿日期: 2021-07-30)