DOI:10.13530/j.cnki.jlis.2023024

科技文献问答式智能检索总体设计与关键技术探析

陈博立　鲜国建　赵瑞雪　黄永文　李　娇　曹雨晴　孙　坦

摘　要　科技文献是人类记录、学习、传承知识的重要手段。大数据时代,传统的基于字符匹配的关键词检索方式无法承载用户检索需求中丰富的语义信息,也无法满足高效、精准、智能检索海量科技文献的要求。本文分析科技文献问答式智能检索的场景需求,提出设计问答式智能检索应当遵循通用性、模块化、可移植三项基本原则,设计总体技术方案,包括掌握问答语言特征、构建知识图谱、研究问答式智能检索交互技术三大步骤。在此基础上,从科技文献知识图谱构建、检索意图理解与识别、检索意图形式化转换、人机自然语言对话管理、检索结果呈现与交互等方面论述实现科技文献问答式智能检索需要重点突破的关键技术,并提出可行的技术迭代方案。本文提出的科技文献问答式智能检索,结合了科技文献知识图谱和自然语言处理等人工智能技术,更加智能精准地理解用户多维复杂的文献检索需求,为用户提供具有高度相关性的检索结果,较传统科技文献检索在用户输入交互、智能检索效率等方面具有优势。图6。表2。参考文献37。

关键词　科技文献检索　智能问答　知识图谱　语义检索　任务型对话

分类号　G254.9

Overall Design and Key Technology of Q&A Style Intelligent Retrieval for Scientific and Technical Literature

CHEN Boli, XIAN Guojian, ZHAO Ruixue, HUANG Yongwen, LI Jiao, CAO Yuqing & SUN Tan

ABSTRACT

Scientific and technical literature is the most important means for human beings to record, learn and inherit knowledge. In the era of big data, the traditional keyword search method based on string matching cannot carry the rich semantic information in users' complex search requirements, which leads to challenges in achieving efficient and accurate tracking and discovery from massive scientific and technical literature.

Therefore, based on the research progress of scientific and technical literature retrieval services and natural language Q&A, this paper puts forward the concept of Q&A style intelligent retrieval for scientific and technical literature. Firstly, we analyze the users' demand of Q&A style intelligent retrieval scenes for scientific and technical literature, including literature, author, fund, subject, institution, journal, time, etc. And it is proposed that generality, modularity, and portability are the three basic principles to be followed in the design of Q&A intelligent retrieval. Then, we give a three-step technical solution for mastering the language characteristics of question and answering, designing

通信作者:孙坦,Email:suntan@caas.cn,ORCID:0000-0002-8257-5064(Correspondence should be addressed to SUN Tan, Email:suntan@caas.cn, ORCID:0000-0002-8257-5064)

knowledge graph, and researching the interactive technology of Q&A intelligent retrieval. After this, we discuss the critical technologies that need breakthroughs, including the construction of knowledge graph for scientific and technical literature, understanding and recognition of retrieval intent, formal conversion of retrieval intent, human-machine natural language dialogue management, presentation and interaction of retrieval results, and a feasible technical iteration scheme is proposed, as well as details that need to be paid attention to in practice. This paper shows a prototype verification system constructed by our team according to the above path, achieving successful parsing of Chinese questions such as "please help me find the literature on white-backed plant hoppers, published in 2017 and included in the Chinese core journal of Peking University, supported by the general project of the Natural Science Foundation of China in which Academician WAN Jianmin participated", etc. At last, we make an outlook for its further optimization and improvement.

The Q&A style intelligent retrieval for scientific and technical literature proposed in this paper aims to combine scientific and technical literature knowledge graph and natural language processing, and other artificial intelligence technologies to more intelligently and accurately understand users' multi-dimensional complex literature retrieval demands, and to provide users with highly relevant retrieval results. Compared with traditional scientific and technical literature retrieval, it has advantages in user input interaction and intelligent retrieval efficiency. It can provide a reference for constructing a next-generation high-end exchange platform for scientific and technical literature and information. 6 figs. 2 tabs. 37 refs.

KEY WORDS

Scientific and technical literature retrieval. Intelligent question and answer. Knowledge graph. Semantic search. Task-oriented dialogue.

0 引言

科技文献是人类记录知识的重要手段,是人类进行科学交流、传播知识的重要工具[1]。科技文献知识的共享交流能够促进知识学习、创意交流、合作沟通,推动科学发展进步。据国际科学、技术和医学出版商协会(STM)的最新统计报告,2020 年全球科技文献发文量达到 470 万条,且发文量增长率逐年上升[2]。然而发文量增长的背后,引申出的是对科研工作者知识获取效率的担忧。据 Elsevier2019 年发布的统计数据,平均每位科研工作者每周需要花费 4 小时查找文献,而阅读文献则需要 5 小时以上[3]。可见在科技文献数量不断增长的当下,文献检索效率仍需进一步提升。

当前人们查找和获取科技文献资源主要通过检索系统来实现。现行的文献检索系统为实现检索结果与用户需求的最佳匹配,以关键词检索为主要手段,并已开始尝试利用语义技术,通过文本挖掘技术从非结构化文本中发现细粒度知识单元,使用知识组织体系实现文献资源语义丰富化[4]。这主要可使检索引擎"了解"自身拥有的资源,但尚不能完全"读懂"用户的真实需求,用户的真实需求通常蕴藏于表达需求的自然语言之中。当前主流检索系统忽略了用户需求的语义性,仍要求用户首先学会各类检索方法,花费大量时间构建、修改以关键词为主体的检索式,检索系统再基于关键词分词进行匹配检索,此后用户再花费较多时间对结果进行甄别筛选。以上过程耗费了用户大量时间与精力,降低了知识获取的效率。究其原因,主要

在于以关键词检索式作为概念的组合,无法准确表达用户文献检索需求的完整语义[5],造成检索系统难以匹配,无法得到用户真正所需的文献,出现了文献大数据时代新的知识饥渴。

针对上述不足,本文提出了知识图谱与人工智能双轮驱动的科技文献问答式智能检索这一研究议题,通过使用自然语言处理等人工智能技术,捕捉用户在检索语言中传达的多维度复杂检索需求,并基于科技文献知识图谱进行高效检索,为用户提供高度相关的检索结果。

1 研究与实践综述

科技文献检索系统以科技文献作为索引对象,根据学科覆盖范围可分为综合性和专业性两类[6]。综合性科技文献检索系统主要有 Web of Science①、CNKI②、Google Scholar、ScienceDirect③、ProQuest④、Semantic Scholar⑤等,专业性检索系统主要有 PubMed⑥、农业专业知识服务系统⑦等。

笔者从使用方式、用户交互等方面对上述科技文献检索系统进行调研,发现尽管检索系统侧重的专业领域不同、资源类型有异,但在用户与服务的人机交互上保持了较高一致性:①在使用方式方面,用户需要高度总结概括检索需求,抽取概念或关键词,在输入框中键入由关键词和运算符构成的检索式并提交检索请求,得到检索结果;②在检索结果筛选与排序方面,用户可根据需求使用分面选项对检索结果进行进一步过滤,排序项包括相关性、发表时间、被引次数、下载量等。

为了更好地使科技文献检索结果与用户需求相匹配,检索系统需要对基于关键词匹配的传统信息检索进行语义扩展,具体有两种实现方式:对关键词检索式进行语义扩展和对知识资源进行语义组织[7]。对关键词检索式进行语义扩展主要利用知识组织体系和自然语言处理技术对检索式中的关键词进行概念消歧与概念扩增,如 GO2PUB 使用基因本体对检索关键词进行语义扩展[8],CO-Search 使用 Siamese-Bert 对检索式进行自然语言处理[9]。也有相关研究侧重于对知识资源进行语义组织,如 Thalia 使用每日更新的 PubMed 关键概念词表配合条件随机场模型识别文献中的各类实体,通过识别结果实现语义检索[10];秦春秀提出了知识元本体模型,帮助检索系统实现细粒度文献检索[11]。知识图谱作为一种语义性、关联性更强的知识组织方式,正受到越来越多学者和机构的青睐[12],如 Springer Nature 推出 SciGraph 科研图谱服务,用户可获取相关实体的可视化图谱[13]。但是,这两种方式仍然难以全面准确揭示用户检索需求中蕴含的丰富语义。基于关键词检索式的检索意图实现方法,主要以检索式形态特征与查询词作为特征提取依据[14],检索式作为概念的简单组合,只能捕捉简单的语义关系,难以精准匹配用户需求与文献资源。改变这一困境的契机,恰恰在于自然语言处理技术的进步以及知识图谱的快速规模化和成熟应用。

自然语言处理技术可以帮助检索系统将用户问句转换为知识图谱中语义相似的概念、属性、关系,建立起正确的语义结构,进而得到更加贴合用户意图的结果[15];自然语言问答省略了用户学习使用的过程,可上手即用,降低了检索门槛。因此,使用自然语言检索的问答系统成为提升检索准确性、相关性、易用性的新方案。如戚越创新性地实现了以学术搜索任务为导向的对话式搜索[16]。此外,学者也对基于知识图谱的智能问答系统进行了探索。Soares

① https://www.webofscience.com/wos/alldb/basic-search
② https://www.cnki.net
③ https://www.sciencedirect.com
④ https://www.proquest.com/index
⑤ https://www.semanticscholar.org/me/research
⑥ https://pubmed.ncbi.nlm.nih.gov
⑦ http://agri.ckcest.cn/index.html

梳理了问答系统的概念、分类及技术等[17];Brodes 使用记忆网络(MemNNs)开发了基于知识库嵌入的简单问句问答系统[18];Dong 等引入多列卷积神经网络,并结合低维表示知识库嵌入开发了简单问句问答系统[19];Hu 等提出图数据驱动的 RDF 问答方法,提升了检索质量与处理效率[15]。相关研究表明利用知识图谱开发支持人类自然语言的检索产品,能够降低开发难度,提升检索效率,改善用户检索体验,具有较高的使用价值与市场前景[20]。此外,有学者对在图书馆场景下使用支持自然语言的服务机器人开展了应用研究。如清华大学图书馆的“小图”以 AIML 作为知识描述语言,可满足用户的聊天、查书以及结果朗读等需求[21];国家图书馆数字图书馆体验区的“小图”支持基于馆藏书目信息的简单语音交互[22];南京大学图书馆开发的“图宝”机器人提供馆藏图书查询功能,在自然语言对话方面更侧重于闲聊[23,24]。

综上,现有科技文献语义智能检索研究集中在基于关键词的语义扩展和知识资源的语义化组织方面,未能充分理解用户检索意图、进行高效精准的知识检索。自然语言作为思维的承载载体,蕴含了用户最真实的意图。因此,笔者对科技文献问答式智能检索的需求场景和技术路线进行分析与设计,建立支持用户使用自然语言进行问答式智能检索的技术系统,进一步拓展科技文献检索交互方式,提高检索效率。如果再集成业已成熟的语音交互服务并打造移动应用,将能创新科技文献智能检索服务场景,并给用户带来全新体验。

2 需求场景分析与技术路线设计

设计知识图谱与人工智能双轮驱动的科技文献问答式智能检索,目的在于完整准确地捕捉用户自然语言中所蕴含的检索意图,使检索系统进一步“智能理解”用户的检索需求,实现基于认知智能的文献检索。笔者在调研现有科技文献检索系统的基础上,通过访谈和咨询等方式,对用户检索文献时使用自然语言可能表达的需求进行收集与整理,形成多场景检索需求,包括对检索内容特征的需求和对检索结果处理的需求。

2.1 检索需求场景分析

科技文献检索内容特征主要由待检索文献及相关实体、实体数据属性和关系属性组合构成,反映了用户期望结果的特征。正确理解检索内容特征需求并对其进行正确解析,是保证智能检索结果准确性的基本前提。实体、实体数据属性、关系属性的具体组合可称为“检索场景”;通过罗列场景,可以达成对自然语言检索问句涉及的实体、属性组合的界定,并以此为依据拓展同一场景下的多种提问形式,以实现尽可能全面的需求分析。

以问句“2018 年至今,万建民院士受自科基金支持的发表在 SCI 收录且影响因子大于 3 的期刊上的关于水稻基因育种的文献有哪些?”为例,该问句中包含研究主题、作者姓名、基金类型、发表年份、期刊收录级别、期刊影响因子等需求要素,通过“支持”“发表”等显式谓词或隐含关系,指示实体与数据属性之间的从属、并列等关系。图 1 展示了科技文献检索内容特征的组合场景。科技文献问答式智能检索支持此类自然语言问句,对具有歧义的内容以对话沟通的方式与用户进行交流确认,以准确、高效的方式为用户提供检索答案。自然语言具有复杂性,场景罗列无法涵盖检索内容特征的所有情况,此为本研究的重点与难点,其解析策略将在下文进行详细叙述。

科技文献问答式智能检索为了提供更加人性化的操作方式,给予用户更好的检索体验,还需要分析用户处理检索结果的需求。智能检索应当支持用户获得检索结果后,通过自然语言指示系统对结果进行操作处理,帮助用户进一步了解、学习和利用文献。用户的相关需求如表 1 所示。

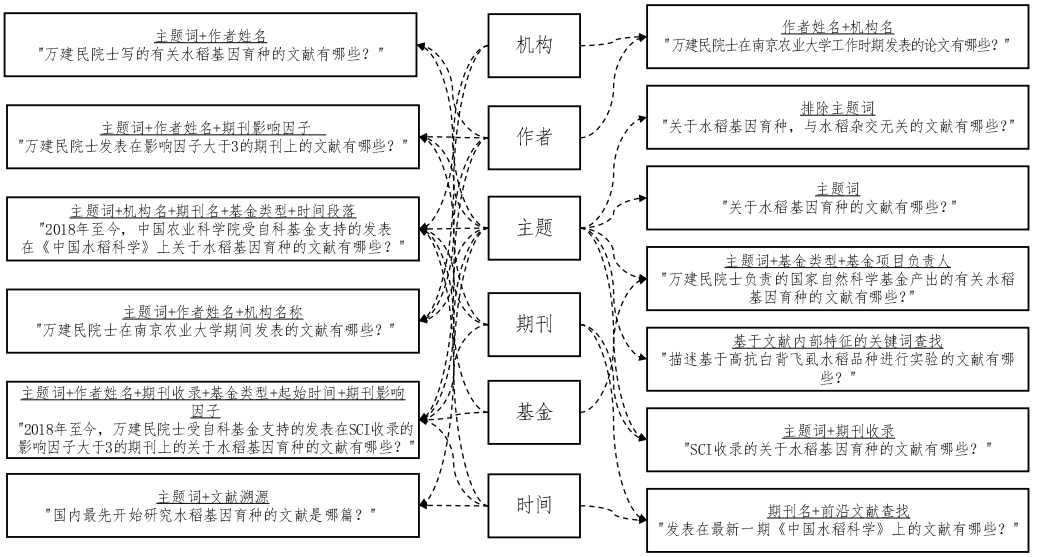

图 1　科技文献检索内容特征的组合场景

表 1　问答式智能检索结果处理指示需求示例

处理对象	需　求	语句示例
检索结果文献	查看文献	打开第 3 篇
	朗读	朗读这一篇文献
	阅览文献	向下/上翻页
	文献标记	把从“水稻育种需要”到“进行大田实验”的段落加入笔记本
检索结果列表	结果选择	选择第 1、2、X 篇
	结果传递	发送检索结果中选定文献的 PDF 格式原文到邮箱
	检索结果追踪	创建文献跟踪
	检索结果导出	把检索得到的文献导出为 EndNote 格式的参考文献，发给我
检索意图修改	意图添加	主题词加入“QTL”
	意图删除	不要有关“水稻白叶枯病”的内容

2.2　遵循的基本原则

科技文献问答式智能检索的设计要充分结合用户检索需求和工程化实施的可行性，需要遵循以下三方面的原则。

(1)通用性原则。为支持用户在不同环境下通过自然语言进行科技文献检索，需要系统全面地梳理、提炼科技文献检索较为通用的需求及核心要素。无论用户以何种形式、在何种平台上检索科技文献，其核心都聚焦在文献研究主题、作者、期刊、发表时间、发文机构、基金项目、影响因子、期刊收录级别等方面。因此，要将上述要素归纳建立科技文献本体模型，并将其作为科技文献知识图谱构建、检索意图识别和形式化转换等诸多环节的基准，进行业务指导和规范。

(2)模块化原则。科技文献问答式智能检索涉及检索意图理解、检索意图形式化转换、人机自然语言对话管理、检索结果呈现与交互等多个环节。为适应用户需求的变化和实现技术

方案的快速迭代优化,在坚持通用性原则的基础上,还需要对整体技术架构进行分层和模块化设计,以提高系统响应能力,提升服务的可维护性与可拓展性。

(3)可移植原则。不同文献检索系统在数据资源存储格式、索引技术结构等方面不尽相同。为实现科技文献问答式智能检索服务的快速移植与灵活部署,需要最大程度减少对原有文献检索系统的升级改造。在遵循通用性原则的基础上,对自然语言问句进行智能理解识别,根据不同的本地集成环境,进行通用检索意图表示向不同技术体系的形式化转换,包括面向不同数据库类型及结构的映射和形式化检索语言的自动适配生成,如针对 Neo4j 的 Cypher、RDF 数据库的 SPARQL、关系型数据库的 SQL 和支持 Solr/ElasticSearch 的 Lucene 检索式等。

2.3　总体技术方案

在明晰用户自然语言中多场景检索需求与检索设计的基本原则后,本文提出科技文献问答式智能检索设计的总体思路及技术路线,如图 2 所示。

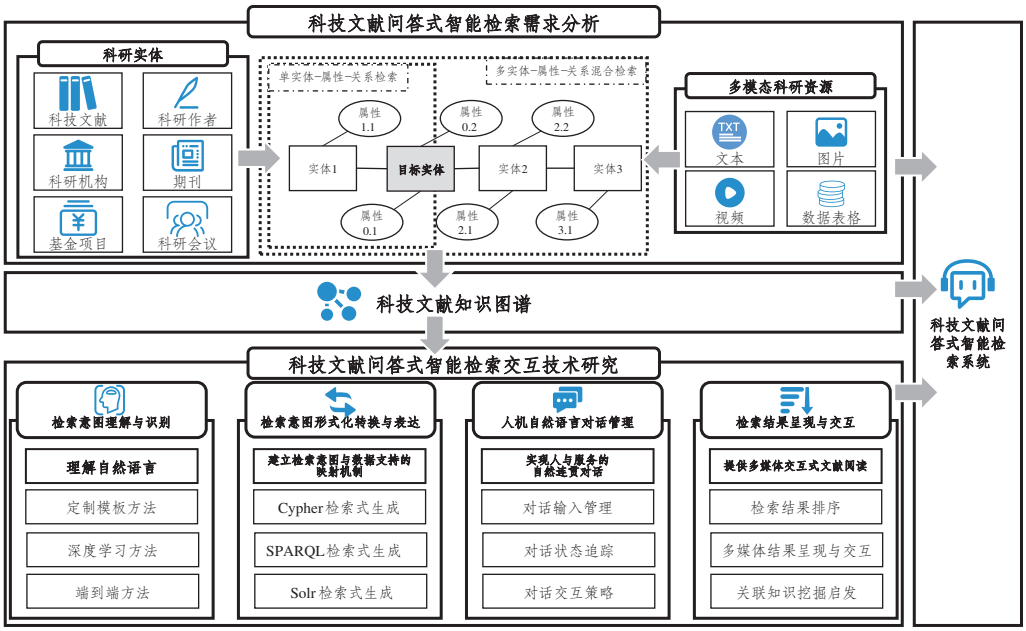

图 2　科技文献问答式智能检索技术路线

科技文献问答式智能检索的实现,大致需要完成三个步骤。

第一步,解析自然语言。需要掌握语言特征,针对不同类别的实体概念设计不同的解析器。通过需求分析,调查厘清检索过程中涉及的实体、实体数据属性、关系属性之间的组合和语言特征,为科技文献知识图谱的构建和交互技术研究奠定基础。

第二步,参考科技文献库、基金项目库、专家学者库、机构库、期刊规范库等,融合现有叙词表、元数据标准、关联数据,对多模态科研资源进行特征梳理,设计科技文献知识图谱模式层。根据模式层定义,可对现有结构化数据进一步规范处理,转换形成科技文献知识图谱数据层,也可只对现有检索系统底层数据结构进行适应性扩展改造。

第三步,研究科技文献问答式智能检索交互技术。针对理解用户输入的自然语言、实现自然语言与形式化查询语句的转换、处理用户输入的查询与指令、精准化呈现检索结果等四

个重点问题,分别通过检索意图理解、检索意图形式化转换、人机自然语言对话管理、检索结果呈现与交互等关键技术来实现。

3 关键技术探析与原型验证

在明确科技文献问答式智能检索的总体设计与实现思路后,本节重点关注实现科技文献问答式智能检索的关键技术,就其目标与任务进行阐述,并探讨原型系统的设计与构建。

3.1 科技文献知识图谱构建

知识图谱为知识组织提供了一种更为精确直观的方式,具有高度的可扩展性。本文采用自底向上构建知识图谱的技术路线[25],通过构建科技文献知识图谱,为自然语言语义解析及知识查询提供支持。本体为知识图谱提供了定义和推理知识图谱节点间语义的方法[26],并为检索意图表示提供了依据。笔者认为,知识图谱对问答式智能检索的支持包含两个层面:①知识图谱本体模式层可作为"通用科技文献检索意图表示模式",为检索意图语义解析、形式化转换以及底层数据库改造提供支持;②构建以科技文献为枢纽的知识图谱数据层,有助于支撑文献问答式检索。通过对科技文献及相关实体进行调研分析,对主流科技文献检索系统、用户需求等进行整理归纳,形成了科技文献知识图谱本体概要模型(见图3)。

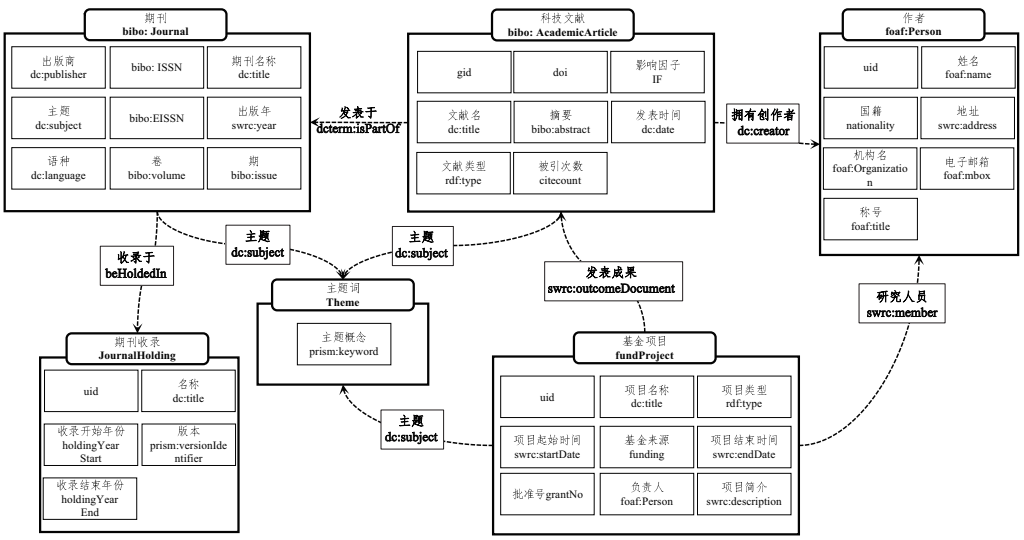

图3 科技文献知识图谱本体概要模型

该本体模型中包含科技文献、作者、期刊、期刊收录、主题词、基金项目等实体类型,关系属性包含参与、发表、收录、撰写等。将解析用户问句后得到的实体、数据属性、关系属性等信息匹配填入该本体模型,可推理得到准确完整的用户检索意图。

此外,用户检索科技文献时所提出的检索问题往往包含目标文献主题词。主题词涉及学科领域,存在着领域知识组织间的异构问题。若将学科领域知识整合入同一知识图谱,容易造成领域间知识表示粒度不一致、一词多义、构建成本过高等问题。但是,仅以文献外部特征作为本体模型,无法建立图谱内知识的广泛关联,导致无法实现精准有效的文献检索。针对该问题,本文构建的科技文献知识图谱本体主要侧重描述文献外部特征,同时充分利用知识图谱灵活的数据结构,以"主题词"实体作为与领域知识组织连接的桥梁。对领域资源的组

织,一方面需要参考现有的叙词表、本体及知识图谱,另一方面需要对文献文本进行深入挖掘,抽取知识作为知识图谱的有力补充。以检索“请帮我查询自科基金面上项目中,发表于 2017 年,被北大核心收录的,万建民院士参与撰写的有关白背飞虱的文献”为例[27],图 4 展示了基于知识图谱本体层的科技文献检索意图表示与领域知识图谱结合的实例。

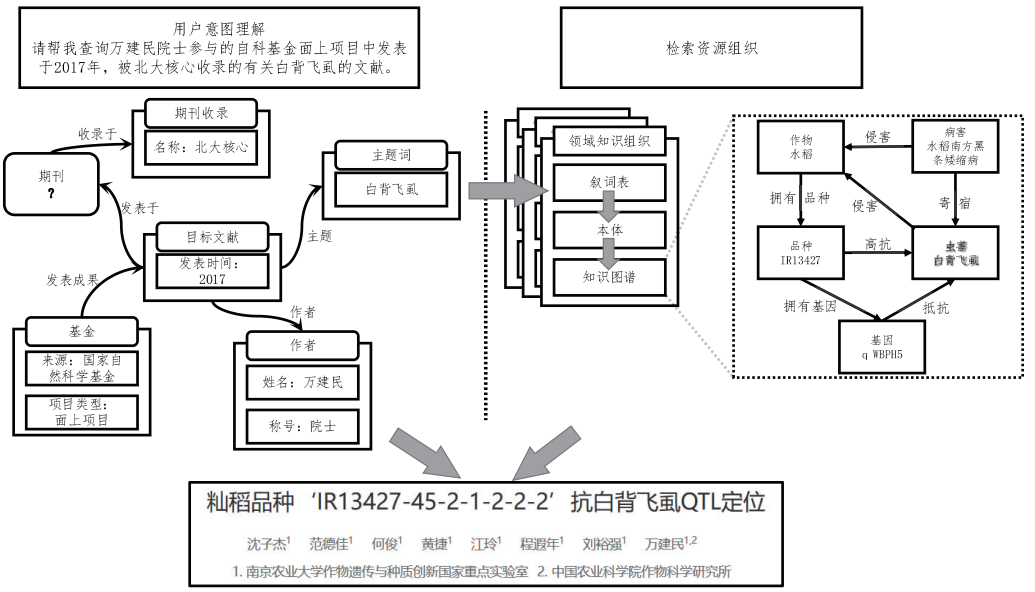

图 4　科技文献检索意图表示与领域知识图谱结合的实例

完成本体设计后,需要基于本体对数据进行映射、转换等处理。科技文献资源不仅包括图书、期刊文献、会议文献、学位论文等文献数据,还包括作者、期刊、基金项目、会议等多类型科研实体数据。在正式填充知识图谱数据层前,需对人名、机构、期刊等进行实体消歧、知识对齐等工作。当前,在小规模实证研究中通常选用 Neo4j 作为知识图谱管理系统。相较于 RDF 模型,Neo4j 的属性图结构支持更加丰富灵活的语义表示[26]。但在工程化实施过程中,考虑到性能和改造代价,还可以选择对现有底层数据库系统进行适应性改造。此外,还可使用知识抽取工具对文献全文进行知识抽取,以进一步丰富文献特征,例如使用主题词提取工具抽取主题词,使用三元组提取工具配合已建立的知识实体词典抽取知识元等。

最后,需要对知识图谱建设成果进行验证。本文主要考虑如下指标。

(1)可获取性。可获取性首先表现在用户已知存在的实体信息可被作为检索条件进行检索;其次,用户必须可检索到最新的文献数据[28]。可获取性直接关系到检索体验,例如,当用户期望获得“万建民发表于 2021 年的最新文献”时,若得到“查无此人”或“结果未收录”的反馈,则会影响用户对检索系统的评价。因此,科技文献知识图谱的建设需要尽可能多地收集数据,并做到定期动态更新图谱数据。

(2)准确性。科技文献知识图谱的准确性有两层含义:①图谱的语义结构准确,即本体设计符合语义,数据映射、标注符合本体设计;②图谱内数据能正确、客观地反应实际情况[29]。在科技文献知识图谱中,数据准确性最容易受到人名、机构等实体多值、歧义的干扰。例如,同一作者的多篇文献,由于作者的名字书写格式以及机构、职称等信息不同,在知识图谱中可

能无法正确地指向同一作者,导致图谱数据准确性下降。因此,实体消歧是知识图谱构建中的关键问题之一。

(3)可信性。科技文献知识图谱的可信性取决于数据来源的可靠性和数据的准确性。因此在选择数据来源时,应当选择权威、可信的数据库,设计全面的数据统一策略并严格执行。在进行实体消歧、知识对齐时,需要以准确度为优先标准。

综上,为了维持科技文献知识图谱的可获取性,需要保持定期、高频率的动态知识更新。知识更新有两个主要任务:①持续对现有数据源进行采集、收割,向知识图谱中补充新增数据;②寻找并使用新的数据源。新数据源与已有数据的对齐、融合需要重点解决三个问题:①新数据结构与已有数据结构的融合、对齐;②数据的对齐、融合、消除冗余;③冲突数据的消歧和单值属性在多数据源中多值的取舍[30]。

3.2 检索意图理解与识别

检索意图是指用户在科技文献检索表达式中传递的特征描述与需求,包含了信息需求、检索目标、检索动机等。不同于关键词检索式仅包含简单的字面信息,自然语言问句承载了完整的语义信息,可以通过解析获取更精准丰富的检索意图。因此,检索意图理解与识别任务的目标是基于科技文献检索业务逻辑,将自然语言问句转换为具体的科技文献检索意图表示。检索意图理解与识别任务包含两个子任务:实体识别和关系抽取。

在科技文献问答式智能检索中,实体识别需要识别的内容包含主题词、文献名称、作者、基金、机构、期刊名称、期刊收录情况等,可通过结合已经建立的实体名称词典与现有的开源软件(如 jieba①、LTP[31])的方式,保证识别具有较高的成功率,并将标识用户问句中关键词所属的实体和属性,形成一系列键值对组合。然而,主题词、文献名称、期刊名称等实体均存在一词多义的情况,例如“水稻”既可能是文献主题词,是名为《水稻》的期刊,也有可能是文献名称的一部分。针对这种情况,一般需要根据业务逻辑,结合上下文情境进行判断。如文献主题词“水稻”所在上下文中,可能出现“有关”或“关于”等词汇,而期刊名《水稻》所在上下文中存在“发表于”“刊登”“期刊”等内容。可使用正则匹配确定实体类型,也可采用依存分析的方法,通过生成语义结构树,对实体间语义关系进行分析,以此提高实体类型判别的精准度[15]。对于无法判断所属实体的新词或专业名词,如其有可能是文献关键词,也可能是基金项目名称或者其他内容时,需要将其在知识图谱中进行全局搜索,并将命中的键值对作为识别结果。

在科技文献检索场景下,因关系类型相对较少且固定,可使用基于规则的方法进行关系匹配,即根据知识图谱本体中对关系、属性的定义为实体识别获得的键值对建立关系。对于存在歧义的情况,例如作者实例存在“撰写”文献与“负责”基金项目两种关系时,可通过正则匹配触发词或依存分析的方法进行消歧。此外,针对整体识别失败的检索问句,应设置默认处理机制,如对文本进行分词,并使用分词结果在预先设置好的检索范围中进行全局检索。

已有研究提出了基于 JSON 格式,通过 ABNF 来描述语法规则的学术搜索领域特定查询语言 SSL,作为学术检索意图的统一描述[16]。这种做法能够统一学术检索系统在开发时所面对的功能需求,形成可复用的框架产出,其思想与本文提出的检索意图结构化表示存在相似之处,在开发过程中简单易实现。检索意图理解与识别的最终结果将生成检索意图结构化表示,供下游任务使用。检索意图表示中包含了检索目标与信息需求,如图 5 所示。

① https://github.com/fxsjy/jieba

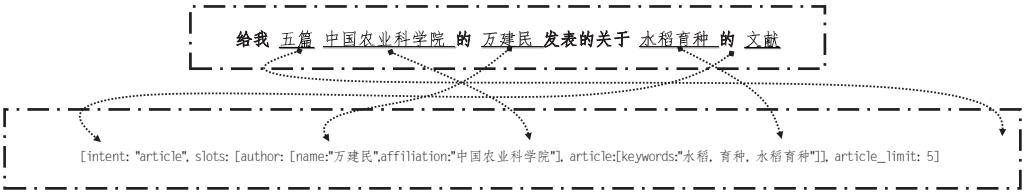

图 5　检索意图结构化表示实例

3.3　检索意图形式化转换

检索意图形式化转换是指将经过自然语言处理后的用户检索意图识别结果,转换为与文献检索系统底层数据库相匹配的形式化查询语句。转换过程由检索意图形式化转换任务执行,该过程需确保:①转换结果符合检索意图表示;②转换结果表达语义符合本地数据组织。

目前存在多种实现检索意图形式化转换的手段,可使转换结果符合检索意图表示的语义。常用的是使用手工模板的方式,其优点在于能够精确高效地对特定类型的问题实现从检索意图表示到数据检索式的转换。该方式的实现流程是根据场景设计中出现的需求类型和通用科技文献检索意图表示模式编写查询模板,将查询模板中设置的变量槽与检索意图表示内容相互匹配,生成可执行的图数据库或其他存储体系的查询语句,而后交由后台检索系统执行查询。以图 6 中的检索意图为例,首先确定检索对象为科技文献实体,然后对检索意图表示中的各项进行逐条解析,生成 Cypher 检索式进行查询,最后根据意图解析结果,对查询结果进行排序并返回相关度最高的五篇文献。

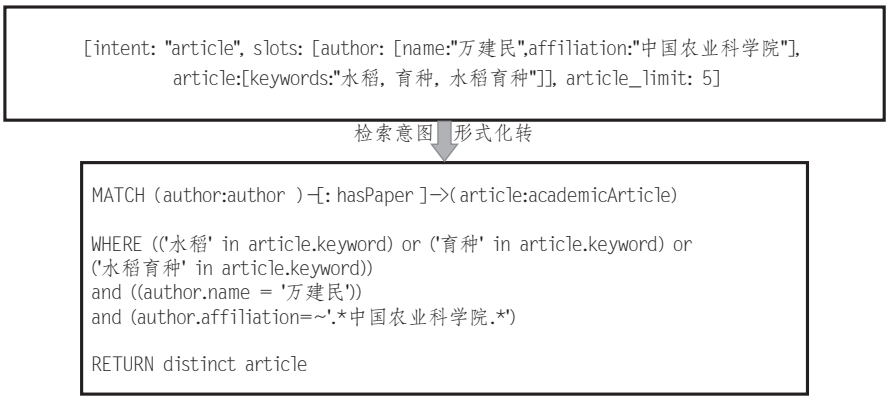

图 6　检索意图表示经形式化转换生成 Cypher 可执行检索式

检索意图形式化转换也可基于知识图谱模式层生成自动化查询语句[32]。首先扫描知识图谱,收集知识图谱模式层实体类型、谓词表述,将其与检索意图表示进行映射匹配,并自动构造符合逻辑的可执行检索式。相较于手工模板方法,基于知识图谱模式层的自动化生成方式耗费人工较少,易于拓展服务应用,适合在领域知识图谱中使用。此外,当收集到足够多的用户真实查询数据时,在检索意图理解与识别阶段和检索意图形式化转换与表达阶段,可以通过深度学习的方式完成联合建模,由用户自然语言问句直接转换为知识图谱查询语句,但是需要一定的标注成本。用户真实查询数据一方面可通过问卷调查直接从潜在用户处收集语料并进行人工标注获得,另一方面可为用户提供手工模板服务或自动化生成服务,从用户使用数据中获得初步标注的

数据,通过人工校验后得到完整的训练数据。受限于语料不足,手工模板方法或基于知识图谱模式层的自动化生成方法是科技文献问答式智能检索启动阶段的选择方案。

针对已将科技文献转换为知识图谱的情形,随着知识图谱的扩展,知识表述存在前后不一致的可能性,进而导致为意图表达匹配知识表述具有不确定性。因此,检索意图形式化转换阶段还需要建立检索意图与知识图谱结构之间的映射机制,以提高知识图谱的一致性和检索意图匹配的准确性。建立映射机制的主要实现方式为:根据知识图谱的实际情况,定期手工维护,并针对性地编写映射程序;同时,应编写通用的映射语法,并提供可视化界面,由用户手工制定映射规则,实现自动化映射。为了减少工作量,可以由系统定期扫描知识图谱模式层,检测近似、冲突表达并呈现给管理人员以进一步设置。

为实现在已运行的文献检索系统中进行问答式检索,可考虑对现有底层数据库或索引的知识组织结构进行改造。首先需要在数据准备阶段将数据重构至符合科技文献知识图谱本体约定的要求,而后根据各个数据库、索引工具所用检索语言的特性,分别编写由通用科技文献检索意图表示模式向不同技术体系实现形式化检索语句转换的程序。数据库、索引工具包括使用 Cypher 语言的 Neo4j,使用 SPARQL 语言的 Virtuoso、GraphDB,基于 Lucene 规范的 Solr、ElasticSearch,使用 SQL 语言的 MySQL 等。

3.4 人机自然语言对话管理

问答式智能检索的关键特征——问答,需要通过检索系统与用户对话交流来实现。对话交流能够帮助消除一词多义、识别错误等情况,迭代理解用户意图、提高解析成功率,增强检索准确度,为提升用户检索体验打下良好基础。人机自然语言对话管理模块主要分为三个子任务:对话输入管理、对话状态追踪和对话交互策略[33]。

对话输入管理用于对用户语音或文字输入内容做出响应。如果是语音交互,需要先调用语音识别模块,将用户语音输入内容转换为文字,供意图理解识别模块统一处理。该子任务需对用户的查询需求及指令需求进行判断,并选择相应的执行程序。同时应按照用户输入的先后顺序,按照先进先出的原则调用理解与查询模块,回应用户请求。

对话状态追踪是保证查询质量的基石,需对意图解析表填充的实体、属性、关系基于知识图谱实现逐步查询以确认其存在性,并根据意图解析表填充质量,判断是否应当保存对话上下文,并进行多轮对话以补充查询信息。当查询失败时,将选取相似度较高的结果作为备选实体留待最终结果查询;当在知识图谱中查询到多个同名实体时,将开启与用户的确认交流,请求用户选择希望查询的实体。同时,由于手工模板设计存在无法支持用户多样化提问的可能性,对话状态追踪也需要检测检索意图表示与问题模板间的匹配度,在无法完全匹配时应与用户进行交互,告知用户当前情况与实际检索内容。

对话交互策略部分负责通过交互提升用户体验,主要包括以下三个功能:①判断服务的开启与结束,为用户送上问候语;②检索过程中告知用户当前程序进展,减少用户等待焦虑;③检索完成后提示用户可进行的下一步操作。此外,对话交互策略还负责针对用户指令需求调用响应程序,实现对检索结果的朗读播报、确认用户电子邮箱以执行结果发送等功能。

3.5 检索结果呈现与交互

检索结果呈现与交互是影响用户体验最直接的模块或环节。针对科技文献问答式智能检索而言,检索结果呈现与交互任务,除了集成上述几部分的技术外,还需要解决多模态内容展示、实体排序、文本排序等问题。

首先是多模态检索结果的展示区位排列问题。多模态资源包含文献、数据集、图片、视频等。检索系统需根据问句意图在页面主视觉区优先展示用户期望模态的相关资源,在主视觉区外侧(如左、右侧边栏)推荐相关多模态内容。

其次是实体排序问题。检索系统需要将实体、实体属性、关联实体信息集中于程序自动生成的虚拟文档中,按照文本排序的思路对实体进行排序[34]。其中,不同属性、不同关系中的关联实体所占权重,需要根据实际情况进行调整,以实现排序效果的最优化。

最后是文本排序问题。检索系统可使用BM25及TF-IDF等基于字面匹配方式工作的技术对文献/虚拟文档与问句相关性进行匹配;还可以考虑结合双塔模型、BERT等深度学习方法,学习问句、文档的稠密向量表示,并通过计算向量距离获得匹配度,按匹配度对检索结果排序。

此外,本任务还需要研究用户与检索结果的交互方式,基于多模态资源个性化响应满足用户对检索结果的获取、阅读或观看需求。为此,检索系统需要合理设置页面布局与相关交互按钮,接受对话交互模块调用,以响应用户诸如朗读文献、将检索结果发送至电子邮箱等需求。

3.6 原型系统实现与验证

在对科技文献问答式智能检索构建过程中涉及的各项关键技术进行设计时,笔者构建了原型系统以验证技术的可行性与系统的实用性。首先根据知识图谱模式层设计,构建了以水稻基因育种为主题、基于Neo4j的文献知识图谱,作为原型系统的数据基础,其中包含期刊论文实体8 140篇,作者实体9 979个,基金项目实体9 907个,相应项目负责人实体9 907个,共计37 933个节点,28 154条关系。在此基础上,建立了问答式文献检索原型系统。由于冷启动阶段尚未收集到足够多的用户提问数据,手工编写并进行标注的数据具有较大的主观性,可能与实际数据分布不符,造成训练数据偏移,影响深度学习模型训练效果[35]。因此,原型系统选择基于模板匹配的方法对用户输入的自然语言问句进行解析。构建流程如下:首先抽取知识图谱内的属性信息,将其作为用户自定义词典,增强了开源自然语言处理工具jieba的实体识别功能,实现了对自然语言问句中的作者、机构、期刊、基金项目等信息检索意图的识别;然后配合正则表达式对实体前后谓词进行识别控制,原型系统能够额外识别出用户对有关文献关键词、基金负责人、期刊收录情况等需求;最后系统将检索意图理解结果通过形式化转换模块转换为Cypher语句。

笔者面向部分研究生和专家开展原型系统的用户体验测试,体验者认为通过自然语言和更智能的问答式方式来表达文献检索需求具有一定创新,并认为本文提出的总体框架及关键技术具有较强的实用性和可行性。在未来具体应用场景和服务模式方面,科技文献问答式智能检索可通过提供意图识别结果、形式化转换结果和文献检索结果三种不同层次和形态的API接口服务,嵌入现有文献检索系统,支持其进行自然语言的问答式检索;也可以基于高质量的知识图谱数据,研发独立的学术搜索引擎、手机APP、小程序等,直接向终端用户提供不限于网页客户端、可在移动设备上使用的问答式检索系统,随时随地为用户提供科技文献检索服务。

4 总结与展望

在当今的大数据智能环境下,用户对科技文献检索提出了新需求。笔者从多场景检索需求、总体实现思路与技术路线、关键技术等方面对科技文献问答式智能检索进行了总体设计与论述,并尝试构建了原型系统,以期为关键技术攻关提供依据和基础。

通过实证分析发现,集成自然语言理解和知识图谱等人工智能技术的科技文献问答式智能检索,可以充分捕捉用户完整的检索需求,并提供维度更丰富、粒度更精细的科技文献服务内容,最终提高科技文献检索服务精准性、提升用户科技文献检索效率。未来通过多模态知识呈现可进一步丰富知识展示维度,帮助用户更好地理解和发现知识内容;通过实现独立学术搜索引擎、手机APP、小程序等移动服务,加入自动语音理解、多轮对话交互等技术,可打破科

技文献服务在使用时间和空间上的桎梏,用户可以跨媒介检索科技文献,获得更加良好的使用体验。对科技文献问答式智能检索与传统科技文献检索进行比较,发现问答式智能检索在用户输入、检索资源、使用条件、歧义修正与意图增补等多个方面具有优势(见表2),这说明科技文献问答式智能检索创新了文献检索方式,具有进一步研究的意义。

表2 科技文献问答式智能检索与传统文献检索的对比

对比维度	科技文献问答式智能检索	传统科技文献检索
用户输入	蕴含完整语义的自然语言问句	关键词或条件组合表达式
检索资源	以知识图谱存储的多模态知识资源	文献元数据或语义增强的文献资源
使用条件	使用语音输入或文本输入,培训学习成本较低	使用文本输入,需具有构建有效关键词检索式的经验
歧义修正,意图增补	支持通过多轮对话的方式与用户沟通,消除歧义;在交流中依靠知识图谱关联关系快速明确用户意图	用户通过筛选条件过滤部分歧义内容;重新构建检索式以添加或过滤意图内容
结果呈现	基于知识图谱的多模态资源推荐,涵盖文本、图像、数据集等	文献文本资源
用户推荐	基于知识图谱的推荐可解决数据稀疏性问题,实现精准语义匹配与多样化推荐[36]	基于内容、关联规则、协同过滤、混合推荐、深度学习的文献推荐[37]
语音服务	依托语音识别、自然语言处理、语音生成等技术实现语音指令理解(如搜索、发送结果、朗读文献等)和语音播报	无

下一步研究工作中,笔者将围绕上述总体设计和实现思路,在已有原型系统基础上,开展大规模知识图谱构建与工程化实践探索,在检索意图智能解析过程中进一步融合多种知识组织体系和深度学习等人工智能方法,为多样化、多场景用户检索意图的准确识别与认知理解提供技术支撑,在多轮对话管理、多模态检索结果呈现与交互方面开展技术研发,并在下一代开放知识服务平台中开展集成应用。

致谢:本文系国家社会科学基金一般项目"融合多种知识组织体系的认知搜索模式研究"(编号:20BTQ014)和国家科技图书文献中心专项"下一代开放知识服务平台关键技术优化集成与系统研发"(编号:2022XM28)的研究成果。

参考文献

[1] 王春林. 科技编辑大辞典[M]. 上海:第二军医大学出版社,2001:69.(WANG C L. Technology editor's dictionary[M]. Shanghai:Second Military Medical University Press,2001:69.)

[2] STM. STM global brief 2021 economics and market size[R/OL]. (2022-08-24)[2023-05-10]. https://www.stm-assoc.org/2022_08_24_STM_White_Report_a4_v15.pdf.

[3] Elsevier Global Communications. Trust in research[EB/OL]. (2019-08-27)[2021-12-12]. https://www.elsevier.com/connect/trust-in-research.

[4] 王颖,吴振新,谢靖. 面向科技文献的语义检索系统研究综述[J]. 现代图书情报技术,2015(5):1-7.(WANG Y,WU Z X,XIE J. Review on sementic retrieval system for scientific literature[J]. New Technology Li-

brary and Information Service,2015(5):1-7.)

[5] 贾彦德. 汉语语义学[M]. 北京:北京大学出版社,1999:17-23.(JIA Y D. Chinese semantics[M]. Beijing: Peking University Press,1999:17-23.)

[6] 曾建勋,丁遒劲. 基于语义的国家科技信息发现服务体系研究[J]. 中国图书馆学报,2017,43(4):51-62.(ZENG J X,DING Q J. Research on the system of national scientific information resources discovery service based on semantic association[J]. Journal of Library Science in China,2017,43(4):51-62.)

[7] 黄敏,赖茂生. 语义检索研究综述[J]. 图书情报工作,2008(6):63-66.(HUANG M,LAI M S. Survey of semantic search[J]. Library and Information Service,2008(6):63-66.)

[8] BETTEMBOURG C,DIOT C,BURGUN A,et al. GO2PUB:querying PubMed with semantic expansion of gene ontology terms[J/OL]. Journal of Biomedical Semantics,2012[2021-12-13]. https://pubmed.ncbi.nlm.nih.gov/22958570/.

[9] ESTEVA A,KALE A,PAULUS R,et al. CO-Search:COVID-19 information retrieval with semantic search,question answering,and abstractive summarization[J/OL]. arXiv,2020[2021-09-09]. https://arxiv.org/abs/2006.09595.

[10] SOTO A J,PRZYBZBA P,ANANIADOU S. Thalia:semantic search engine for biomedical abstracts[J]. Bioinformatics,2019,35(10):1799-1801.

[11] 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.)

[12] 常娥,夏婧. 多种知识组织方法比较[J]. 图书馆论坛,2016,36(8):1-6.(CHANG E,XIA J. Comparative study of knowledge organization[J]. Library Tribune,2016,36(8):1-6.)

[13] Spinger Nature. SN SciGraph:a linked open data platform for the scholarly domain[EB/OL].[2021-08-23]. https://www.springernature.com/cn/researchers/scigraph.

[14] 陆伟,周红霞,张晓娟. 查询意图研究综述[J]. 中国图书馆学报,2013,39(1):100-111.(LU W,ZHOU H X,ZHANG X J. Review of research on query intent[J]. Journal of Library Science in China,2013,39(1):100-111.)

[15] HU S,ZOU L,YU J X,et al. Answering natural language questions by subgraph matching over knowledge graphs[J]. IEEE Transactions on Knowledge and Data Engineering,2018,30(5):824-837.

[16] 戚越. 面向自动问答的学术搜索通用查询语言设计与实现[D]. 武汉:武汉大学,2020.(QI Y. Design and implementation of a general query language for academic search based on automatic question answering[D]. Wuhan:Wuhan University,2020.)

[17] SOARES M A C,PARREIRAS F S. A literature review on question answering techniques,paradigms and systems[J]. Journal of King Saud University:Computer and Information Sciences,2020,32(6):635-646.

[18] BORDES A,USUNIER N,CHOPRA S,et al. Large-scale simple question answering with memory networks[J/OL]. arXiv,2015[2021-06-21]. https://arxiv.org/abs/1506.02075.

[19] DONG L,WEI F R,ZHOU M,et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Association for Computational Linguistics,2015:260-269.

[20] PUDARUTH S,BOODHOO K,GOOLBUDUN L. An intelligent question answering system for ICT[C]//2016 International Conference on Electrical,Electronics,and Optimization Techniques(ICEEOT). IEEE,2016:2895-2899.

[21] 姚飞,张成昱,陈武. 清华智能聊天机器人“小图”的移动应用[J]. 现代图书情报技术,2014(7):120-126.(YAO F,ZHANG C Y,CHEN W. The mobile application of “Xiaotu”:the smart talking robot of Tsinghua University Library[J]. New Technology of Library and Information Service,2014(7):120-126.)

[22] 申悦. 人工智能机器人在图书馆的设计与实现——以国家图书馆数字图书馆体验区为例[J]. 图书馆,2020(6):37-41.(SHEN Y. The design and realization of intelligent robot in library:a case study of digital library experience area in National Library of China[J]. Library,2020(6):37-41.)

[23] 樊慧丽,邵波. 智能机器人图书盘点创新实践与思考——以南京大学图书馆为例[J]. 图书馆,2018(9):96-100.(FAN H L,SHAO B. Reflection and innovative practice of book inventory with intelligent robot:a case

study of Nanjing University Library[J]. Library,2018(9):96-100.)

[24] 杨子帅,王颖纯,刘燕权. 图书馆服务中人工智能技术应用的调查研究[J]. 图书馆,2019(10):34-40.(YANG Z S,WANG Y C,LIU Y Q. Investigation and research on the application of artificial intelligence technology in library service[J]. Library,2019(10):34-40.)

[25] FENSEL D,ŞIMŞEK U,ANGELE K,et al. How to build a knowledge graph[M]//Knowledge graphs:methodology,tools and selected use cases. Cham:Springer International Publishing,2020:11-68.

[26] HOGAN A,BLOMQVIST E,COCHEZ M,et al. Knowledge graphs[J/OL]. arXiv,2020[2021-12-27]. https://ui. adsabs. harvard. edu/abs/2020arXiv200302320H.

[27] 沈子杰,范德佳,何俊,等. 籼稻品种'IR13427-45-2-1-2-2-2'抗白背飞虱QTL定位[J]. 南京农业大学学报,2017,40(6):957-962.(SHEN Z J,FAN D J,HE J,et al. Mapping of the quantitative trait locus conferring white-backed planthopper resistance in indica cultivar 'IR13427-45-2-1-2-2-2'[J]. Journal of Nanjing Agricultural University,2017,40(6):957-962.)

[28] WANG R Y,STRONG D M. Beyond accuracy:what data quality means to data consumers[J]. Journal of Management Information Systems,1996,12(4):5-33.

[29] FÄRBER M,BARTSCHERER F,MENNE C,et al. Linked data quality of DBpedia,Freebase,OpenCyc,Wikidata,and YAGO[J]. Semantic Web,2018,9:77-129.

[30] BLEIHOLDER J,NAUMANN F. Data fusion[J]. ACM Computing Surveys,2009,41(1):1-41.

[31] CHE W X,LI Z H,LIU T. LTP:a Chinese language technology platform[C]//Coling 2010:Demonstrations. Coling 2010 Organizing Committee,2010:13-16.

[32] GORSHKOV S,KONDRATIEV C,SHEBALOV R. Ontology-based question answering over corporate structured data[C]//2021 International Symposium on Knowledge,Ontology,and Theory(KNOTH). IEEE,2021:70-75.

[33] 赵阳洋,王振宇,王佩,等. 任务型对话系统研究综述[J]. 计算机学报,2020,43(10):1862-1896.(ZHAO Y Y,WANG Z Y,WANG P,et al. A survey on task-oriented dialogue systems[J]. Chinese Journal of Computers,2020,43(10):1862-1896.)

[34] BAST H,BUCHHOLD B,HAUSSMANN E. Semantic search on text and knowledge bases[J]. Foundations and Trends© in Information Retrieval,2016,10(2-3):119-271.

[35] GU Y,KASE S,VANNI M,et al. Beyond I. I. D.:three levels of generalization for question answering on knowledge bases[C]//Proceedings of the Web Conference 2021. New York:Association for Computing Mahinery,2021:3477-3488.

[36] 阳德青,夏西,叶琳,等. 知识驱动的推荐系统:现状与展望[J]. 信息安全学报,2021,6(5):35-51.(YANG D Q,XIA X,YE L,et al. Knowledge-enhanced recommender systems:a survey and prospect[J]. Journal of Cyber Security,2021,6(5):35-51.)

[37] 屈冰洋,王亚民. 基于深度学习的科技信息文献推荐模型研究[J]. 情报理论与实践,2021,44(11):160-165.(QU B Y,WANG Y M. Research on science and technology information papers recommendation model based on deep learning[J]. Information Studies:Theory & Application,2021,44(11):160-165.)

陈博立 中国农业科学院农业信息研究所硕士研究生。北京 100081。

鲜国建 中国农业科学院农业信息研究所研究员,博士生导师。北京 100081。

赵瑞雪 中国农业科学院农业信息研究所研究员,博士生导师。北京 100081。

黄永文 中国农业科学院农业信息研究所研究员。北京 100081。

李 娇 中国农业科学院农业信息研究所助理馆员。北京 100081。

曹雨晴 中国农业科学院农业信息研究所硕士研究生。北京 100081。

孙 坦 中国农业科学院副院长,农业农村部农业大数据重点实验室主任,研究馆员,博士生导师。北京 100081。

(收稿日期:2022-08-12;修回日期:2023-04-25)