图书馆智能知识服务的未来

孙　坦

当前,知识关联分析、前沿态势分析、技术演化分析及技术预见是图书馆为科技创新和技术转移提供情报服务的主要模式,运用的主要技术方法是文献计量。但是,辛丑年春节前后,两个机器人"吵架"的视频不仅带火了江西省图书馆,也揭示了人工智能技术在机器自主互操作领域的应用前景,更激发我们思考图书馆智能知识服务的未来——如何通过机器人利用图书馆海量知识帮助人类更高效地传播、学习和发现知识,促进技术创新。

开放科学的发展和人工智能技术的进步是支撑我们期待智能知识服务未来的两个主要因素。首先,在持续进步的大科学装置的支撑下,人类拥有越来越强大的从基因组到宇宙粒子等各个领域更加细粒化的智能信息感知和数据获取能力,这为数据密集型科学发现提供了海量的大数据,也为跨学科领域的知识交叉融合奠定了基础。当前呈现出来的生物技术、信息技术和材料科学的高度交叉融合即是很好的例证。同时,网络化、数据化和人工智能技术的迅猛发展为不同机构、不同领域、不同地域的科学家开展协同研究构建了一个良好的虚拟研究环境。多源大数据和虚拟研究环境共同孕育和持续推动着开放科学的不断发展。

其次,人类日益活跃的科学研究和技术创新活动持续产出大量的科技文献,系统记述了科学发现和技术创新的全部过程和知识,构成了比原始科学数据更加优质、高效的文本大数据,这更有助于人们进行知识传播、知识发现和技术创新。但是,文本大数据的非结构化、弱语义表示和知识计算难度制约了科技文献的高效率、深层次利用能力。深度机器学习和文本挖掘等人工智能技术的突破性进展让这些障碍迎刃而解。目前,文本挖掘是支撑图书馆智能知识处理的核心关键技术。2017 年 8 月,Allahyari 等人[1]系统总结了文本挖掘技术的技术体系与进展。2019 年 7 月,Brett Drury 和 Mathieu Roche[2]又系统总结了文本挖掘技术在农业领域的应用。2021 年 2 月,孙坦等[3]对文本挖掘技术在知识服务中的应用进行了述评与展望。上述研究总结与分析表明,依托海量的科技文献资源,应用文本挖掘等人工智能技术,图书馆知识分析服务的智能化前景充满光明。

当前,图书馆智能知识服务的相关研究多聚焦于知识图谱(Knowledge Graph)、语步识别(Moves Recognition / Rhetorical Move Detection)、研究侧写(Research Profile)、自动分类与聚类以及重大技术突破识别(Breakthrough Discoveries Identification)等方面。采纳的主要技术路线是以科技文献为基础语料和对象,运用 BERT 等机器学习算法工具,融合经过个性化改进设计的学习策略,开展科技文献中各类实体的识别、抽取、聚类分类以及关联计算等。其中,语义关系是核心瓶颈问题,相关的研究普遍存在两个方面的不足:①将人工智能和自然语言理解领域的技术方法与图书情报领域的知识组织方法进行融合,但是忽略了原技术方法应用于大数据文本目标下的非监督性、时间成本等指标,只是以小文本量实验数据指标与原有技术方法进行比较;②局限于自然语言处理和人工智能技术

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)

领域对文本语义关系的理解和限定,缺乏面向科学研究需求的关键性语义关系的深入挖掘与应用。笔者研究团队的实践表明,基础知识库和训练语料库的构建策略、语义模型的构建方法、语义关系处理策略是突破现有研究局限的可能路径。

1 发挥图书馆知识库与语料库资源优势的技术策略

图书馆拥有海量的科技文献,经过数字化、结构化转换和适当的知识组织即可成为优质可信的基础知识库,作为图书馆智能知识服务的大数据仓库。通常情况下,相关研究会截取其中少量的数据集合经过人工标注后作为机器学习的训练语料,并以此为基础开展模型训练和算法改进,优化学习策略。相关研究表明,实体及实体间语义关系在文档中所处段落位置不同,其对文档核心主题的表征力度和角度也不同,实体间语义关系的精准识别与上下文背景也呈现出密切相关,而谓词表征的实体语义关系是最关键的语义关系[4-6]。当前,实体语义关系识别、语义模型构建及知识计算研究中,相关的语义相似度计算包括词向量、句子向量和文档向量三个级别。目前普遍采用的计算方法主要基于词向量级别,但当突破了计算能力和时间成本的限制之后,句子向量和文档向量级别的探索或将具有更大的潜力。笔者的研究实践表明,抽取基础知识库中的部分数据集进行人工标注作为训练语料的方法噪声较大,一种可行的改进方向是增加迭代,在人工标注基础语料的基础上,融合实体和实体关系在文档中所处的段落位置及其上下文,按照其对整篇文档主题和语义关系的表征度分类进行句子级抽取,继而构建句子级预训练语料库。

2 从 A-Box 到 T-Box 构建语义模型

语义模型是开展知识单元语义互操作和知识计算的重要基础,特别是对开展知识关联、推理和自动问答服务具有不可或缺的核心作用。通常情况下,由于采用 A-Box 模式构建语义模型需要具备海量数据的计算处理能力,多数研究更倾向于采用自顶向下的 T-Box 方法构建语义模型。构建通用语义模型似乎相对容易,可以继承或复用已有的模型并进行个性化的改进优化,但是由于更多地表征领域通用知识框架和语义关系,有可能导致其与具体的语料集存在冲突或缺失,从而无法更精准有效地反映目标对象数据集的真实情况。笔者在词向量、句子向量和文档向量三个层次上的实验表明,在高性能计算环境的支持下,采用 A-Box 方法自底向上个性化构建语义模型的质量和可应用性好于 T-Box 模式,特别是在边缘交叉和新兴领域表现更为明显。由此形成了一条值得期待的新技术路线:针对给定预训练语料集,首先通过无监督或弱监督机器学习进行相似度聚类和分类,再分别对各子类语料集人工标注后进行实体及实体关系的识别抽取,构建初步的语义网络或知识图谱,最后根据已有对应领域的通用模型进行校验和优化。

3 从语义模型到抽象语义表示(Abstract Meaning Representation,AMR)

随着人工智能技术的综合发展,基于协同推理和知识计算的智能问答将成为智能知识服务的主要方式。依照发展阶段和应用场景可以将智能问答服务划分为两种模式:人机交互迭代式问答服务和智能系统间的交互问答服务。前者主要是辅助用户快速高效地获取所需的知识,并根据用户需求进行知识分析,在人机交互中完成知识发现与知识获取;后者则主要是应用于复杂智能知识服务系统内的自主语义互操作,例如承担不同知识处理任务的智能系统或设备之间的自主交互对话。现有的知识图谱、语义模型等基本可以满足人机交互迭代的知识处理需求,但是难以满足智能知识系

统间自主语义互操作的要求,因而需要更加严谨表达语义的框架和方法。抽象语义表示方法采用了单根有向无环图模式表达语义依存,且允许增加文档之外的概念节点作为虚节点辅助表达语义依存关系,基本具备比较完整准确表示一个句子语义的能力[7]。因此,面向智能知识服务系统中的机器间会话、自主推理和知识计算的需求,抽象语义表示方法具有值得探索的重要前景,特别是中文语义表达的优势使其在中文知识处理与计算中的应用更加值得期待。

4　聚焦关键语义关系,探索计算因果推理(Computational Causal Inference, CompCI)

相对日常用语而言,科技文献中蕴含着十分丰富的实体、语义类型和语义关系,例如统一医学语言系统的语义网络(UMLs Semantic Network)定义了135种语义类型和54种语义关系,尚不能有效覆盖所有语义类型和关系[8]。从知识关联、推理和计算的需求看,在某种程度上我们可以将很多语义关系抽象为因果关系,如作用关系、影响关系、治疗关系、演化关系等。因果关系是推理、问答等智能知识服务最依赖的关键语义关系,在大数据文本中对因果关系进行识别、抽取、推理和计算是至关重要的。Wong等[9]提出,计算因果推理软件的设计应遵循机器学习和潜在结果框架的设计原则,将因果效应的框架和识别与其估计策略分开,这样可以为各种各样的因果效应开发一个通用和公共框架,然后个性化优化其内部数值引擎。这种软件设计将允许用户通过简单的指定模型的形式和数据的假设来估计因果效应,这些假设可以用来确定效果是否被识别。该研究为我们清晰揭示了计算因果推理在智能知识服务中的广阔应用前景,可以支撑智能知识服务在海量复杂文本中自动梳理出知识发展和演化的脉络,探索创新链的演化全景,甚至可以进行技术预见和发展态势预测,也可以支持不同智能知识服务装备间的自主互操作。

通过对文本挖掘关键技术发展和技术应用策略的分析,结合自然语言理解技术的发展,我们可以比较清晰地感受到图书馆智能知识服务技术的发展方向,更可以看到图书馆能够充分发挥自身在海量知识资源和用户群体方面的优势。我们应当充分把握人工智能和大数据技术发展的契机,为智能知识服务的发展提供丰富的基础知识库和语料库,围绕面向科学发现、技术创新和技术扩散的知识服务需求,大力开发图书馆智能知识服务技术、系统和云边端一体化的智能装备,为图书馆在新技术时代的服务创新奠定强大基础。

参考文献

[1] Allahyari M, Pouriyeh S, Assefi M, et al. A brief survey of text mining: classification, clustering and extraction techniques[J/OL]. [2021-03-03]. https://arxiv.org/pdf/1707.02919.pdf.

[2] Drury B, Roche M. A survey of the applications of text mining for agriculture[J]. Computers and Electronics in Agriculture, 2019, 163(104864): 1-13.

[3] 孙坦,丁培,黄永文,等. 文本挖掘技术在农业知识服务中的应用述评[J]. 农业图书情报学报,2021,33(1):4-16. (Sun Tan, Ding Pei, Huang Yongwen, et al. Review on the application and development strategies of text mining in agriculture knowledge services[J]. Agricultural Library and Information, 2021, 33(1): 4-16.)

[4] Zhang Tianxiao. Entity-relation search: context pattern driven relation ranking[D]. University of Illinois at Urbana-Champaign, 2016.

[5] Kim H, Sun Y, HockenmaierJ, et al. ETM: entity topic models for mining documents associated with entities[C]// 2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium, 2012: 349-358.

[6] 何琳, 常颖聪. 不同标引策略下的文本主题表达质量比较研究[J]. 图书馆杂志, 2014, 33(5): 29-33. (He Lin, Chang Yingcong. Comparative study of subject presentation with different indexing strategies[J]. Library Journal, 2014, 33(5): 29-33.)

[7] Bonial C, Donatelli L, Abrams M, et al. Dialogue-AMR: abstract meaning representation for dialogue[C]// Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 2020: 684-695.

[8] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology[J]. Nucleic Acids Research, 2004, 32(Database issue): D267-D270.

[9] Wong J C. Computational causal inference[J/OL]. [2021-03-03]. https://arxiv.org/pdf/2007.10979.pdf.

孙 坦 中国农业科学院副院长, 农业信息研究所研究馆员, 中国图书馆学会副理事长。北京 100081。

(收稿日期: 2021-03-03)

在新发展格局中推进公共图书馆的创新和高质量发展

王世伟

2021年3月第十三届全国人大四次会议表决通过的《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》(以下简称《纲要》)中指出: 形成强大国内市场, 构建新发展格局; 并提出了畅通国内大循环、促进国内国际双循环、加快培育完整内需体系的重大举措和发展路线图。这是与时俱进提升我国经济发展水平的战略抉择, 也是塑造我国国际经济合作和竞争新优势的战略抉择, 体现出前瞻性思考、全局性谋划、战略性布局、整体性推进的思路和理念, 这对于谋划"十四五"时期公共图书馆的创新和高质量发展并实现在2035年建成文化强国的发展目标, 同样具有重要的战略指导意义和指引价值。

1 以加快构建国内大循环为主体提升公共图书馆服务水平

加快构建以国内大循环为主体、国内国际双循环相互促进的新发展格局是"十四五"时期我国发展的重要指导思想, 是在危机中育先机、于变局中开新局的科学主动应对之策, 它对于

通信作者: 王世伟, Email: swwang@sass.org.cn, ORCID: 0000-0001-9179-9580(Correspondence should be addressed to WANG Shiwei, Email: swwang@sass.org.cn, ORCID: 0000-0001-9179-9580)