

智能文献计量学视角下的知识发现： 人工智能 + 信息计量学*

张 巍**

(悉尼科技大学澳大利亚人工智能研究所, 悉尼 NSW 2007)

摘 要:如何高效准确地发现知识并提炼情报是信息管理与图情领域的重要课题。大数据时代人工智能技术的飞速发展为此课题带来新视野。本文由读《学科信息学与学科知识发现》所悟所感出发,探讨应用于知识发现的文献计量学方法。着眼于文献计量学方法在大规模数据处理、复杂关系发现以及模型可适性、复用性与鲁棒性等问题上的缺陷,本文以智能文献计量学为核心,剖析其与人工智能技术的交互、整合与再创新,并介绍其在生物医学领域知识发现的最新应用。此外,本文还介绍了智能文献计量学在复杂实践中的缺陷,浅谈其发展前景。

关键词:智能文献计量学;知识发现;人工智能;信息计量学;文献计量学

DOI:10.16507/j.issn.1006-6055.2021.01.014

从数据(Data)到信息(Information),再到知识(Knowledge)与情报(Intelligence),这条看似简单的链条,囊括了大数据框架下“数据驱动”的要义,也贯穿了数据科学的始终。以机器学习为代表的人工智能技术的飞速发展,为数据科学打开了新视野,带来了新方法。如何高效准确地发现知识、提炼情报,成为信息管理与图情领域的重要课题。《学科信息学与学科知识发现》一书由理论出发,介绍学科信息学体系;以数据为着眼点,剖析并归纳科技数据特征;立足于技术,探究学科知识发现的工具与方法;回归应用,挖掘新兴学科技术情报与创新机会。本文为读《学科信息学与学科知识发现》所悟所感,探讨当前应用于知识发现的文献计量学方法,以智能文献计量学为核心,强调其与人工智能技术的交互、整合与再创新,并剖析这一新兴技术趋势的应用优势及缺陷,

浅谈其发展前景。

1 文献计量学

现代文献计量学源自 Derek Price 博士对科学活动模式的观测。早期的文献计量学强调用于分析图书的数学与统计学方法,而伴随着 20 世纪 90 年代以来信息技术的迅速发展,与数据形式及规模的爆炸式增长,文献计量学的分析对象从图书扩展到包含学术论文、专利、技术报告以及项目申请书等在内的广义的科技文献数据。同时,科学活动的载体不再局限于科技文献,广泛的科技数据、大数据时代所兴起的社交媒体数据以及蕴含科学活动模式的一切信息资源纷纷进入文献计量学的研究范畴,至此文献计量学、科学计量学(Scientometrics)与信息计量学(Informetrics)相互融合,分析方法逐步趋同。

* 澳大利亚研究理事会早期职业研究员奖(ARC DECRA)项目“跟踪与预测技术变化的智能文献计量学(Intelligent Bibliometrics for Tracking and Predicting Technological Change)”(DE190100994)

** E-mail: Yi. Zhang@uts.edu.au

以引文分析与合著者分析为核心的文献计量学在引入以自然语言处理技术为代表的文本挖掘方法与信息可视化技术后,适用范围大幅提升,被广泛应用于科学学、科技管理与科技创新政策等众多领域,如剖析特定技术领域与多学科交互、发现科技主题及其潜在关系、追踪技术演化路径等。显然,诸如上述应用场景,文献计量学的核心任务正是面向科学技术的知识发现。然而,在大数据背景下,文献计量学方法在大规模数据处理(如百万级及以上的文献处理与全文本分析)、复杂关系发现(如继承关系、因果关系等)以及模型的可适性、复用性与鲁棒性等问题上相对乏力。如何在大数据框架下提升文献计量学方法发现知识、提炼情报的能力成为学界的热门话题之一。

2 智能文献计量学

毋庸置疑,人工智能技术强大的适应性学习与数据分析能力为文献计量学的长足发展带来新契机。智能文献计量学(Intelligent Bibliometrics)应运而生,它强调构建将文献计量指标(如,引用信息、主题词信息、作者信息等)与人工智能方法相结合的计算模型,并使其适应广泛的科学、技术、创新与政策分析场景。当前,智能文献计量学方法的尝试与创新集中于思考如何从人工智能技术中“取长补短”,例如,基于词嵌入(Word Embedding)与核(Kernel)方法的科技主题萃取利用 Word2Vec 模型生成的词向量矩阵替代传统共现矩阵,提升了特征提取能力并助力于后续主题聚类算法;引入流数据分析(Streaming Data Analysis)的继承关系发现有效追踪了特定时间区间内科技主题诞生、演化、消亡与再生等形态变化,较之传统的相似度计算,关系刻画的准确性、主题演化的还原性以及计算模型的适应性均大幅

提升;主谓宾分析(Subject-Action-Object Analysis)与网络分析的创造性结合为依托网络拓扑结构预测节点间潜在关系(即,“问题—解决方案”关系)提供了可能,有效降低了传统主谓宾分析中对专家知识的高度依赖。

事实上,与《学科信息学与学科知识发现》一书所系统介绍的生物医学文本挖掘方法以及案例相印证,面向生物医学文献的知识发现正成为智能文献计量学的重要应用之一。一方面,智能文献计量学有效继承了传统文献计量学分析科技文献的思路与方法,保留了基于多维度文献计量指标分析的能力。例如,以共现分析(如:共引、主题词共现等)为核心的主题分析,发现特定领域技术分布与关联等。另一方面,智能文献计量学与人工智能技术的交互与生物信息学(Bioinformatics)方法具有天然亲和力,为跨学科创新提供便利。例如,2019 年末以来新冠肺炎疫情的暴发迅速破坏了传统冠状病毒研究的既定知识体系,我们运用前述提到的数据流分析与网络分析方法,以复杂系统的视角追踪并检测知识系统分裂与再适应的过程,同时,与“技术挖掘(Tech Mining)”体系的深度结合绘制自 20 世纪 40 年代以来 80 余年的冠状病毒主题演化图谱,为冠状病毒科研人员以及公共卫生政策制定者提供决策支持。此外,相关的应用还包括:绘制基因相关的心脑血管疾病主题演化图谱、基于链路预测方法发现房颤相关的重要基因等。

在 2021 年的 iConference 上,以“人工智能 + 信息计量学”为主题的研讨会(<https://ai-informetrics.github.io/>)将智能文献计量学的理论基础、方法框架与应用范畴推向了一个更为广阔的多学科平台。显然,在构建人工智能视角下的知识发现方法体系这一共同目标上,智能文献计

量学与《学科信息学与学科知识发现》多学科融合与交互的主旨思想是相得益彰的。

3 前景

无论是学科知识发现,还是更为广泛的科学、技术、创新与政策分析,问题是复杂的——如何识别内在的层级与耦合,如何考量外部的影响与作用,如何发现随着时间演化的动力系统等。在当前的技术条件下,即便是人工智能,亦不能实现完全替代人类的真正“智能”。因此,思考以智能文献计量方法为代表的定量方法与以专家知识为代表的定性方法的有效沟通与深度交互仍然是亟待

思考的重要问题。在这一方向上,《学科信息学与学科知识发现》亦为我们提供了很好的范例。无论是干细胞领域新兴技术主题的突破性预测,还是医学领域前沿的主题识别,恰当的专家知识宛若画龙点睛,从深度与广度上延伸了知识发现的能力与效用。

在 21 世纪的第三个十年悄然开启的今天,人工智能的“潘多拉魔盒”已然打开,如何理解与驾驭人工智能,并将其应用于新时代的知识发现,也许还有很长的路要走。所谓,“如将不尽,与古为新”,借古人言,与诸君共勉。



美国成为全球生命科学创新领导者的关键原因

美国科技智库信息技术与创新基金会 (Information Technology and Innovation Foundation, ITIF) 副会长 Stephen Ezell 于 2021 年 6 月 12 日在 Alpha-1 基金会年会上作题为《确保美国生物制药持续竞争力》(Ensuring Continued U. S. Biopharmaceutical Competitiveness) 的报告,分析了美国成为全球生命科学创新领导者的关键原因,对社会各界批评美国生命科学产业的各种观点进行了反驳,并给出了确保美国生物制药领域全球领导地位的具体政策建议。报告认为,美国在新药开发等生命科学创新方面保持全球领先水平,其保持生命科学创新领导力的关键原因包括以下 5 个方面:

- (1)对生物医学研究领域强有力的公共/私人投资:美国企业和政府在生物制药研发方面的投资处于全球领先地位。
- (2)促进生命科学创新的积极激励措施:1981 年,成为第一个引入研究与实验 (R&E) 税收抵免制度的国家,有 40 多个州至今仍提供 R&E 税收抵免;1983 年,实行了罕见病药物税收抵免制度,1983—2018 年,为临床试验费用提供 50% 的信贷(现在为 25%),且已批准了 850 多种可用于治疗 250 余种罕见疾病的药物。
- (3)有效的监管/药物审批制度:1992 年《处方药使用者付费法案》(FDUFA)批准实施之后,美国的新药批准进程明显加快。
- (4)生命科学创新需要的强有力的知识产权保护制度:《拜杜法案》(Bayh-Dole Act,为解决和促进科研成果商业化而设立)允许联邦资助的研究者获得其所完成发明的专利权;要求对成果的可专利化和权利资格有明确的标准。
- (5)合理的利润对生物制药创新至关重要:美国国会预算办公室(CBO)指出,未来十年由于对药物价格的控制将会使新药数量减少 3% ~5%。

冯志刚、张志强(中国科学院成都文献情报中心)编译自
<https://www2.itif.org/2021-life-sciences-alpha-1-foundation.pdf>