

面向融合出版技术的主题演化研究^{*}

曾江峰¹ 黄泳潼¹ 陈静怡¹ 冯昌扬^{1,2}

(1. 华中师范大学信息管理学院, 武汉 430079; 2. 富媒体数字出版内容组织与知识服务重点实验室, 北京 100038)

摘要: 掌握融合出版领域前沿技术的应用现状与趋势, 为融合出版领域的技术应用提供参考。选择2006—2022年融合出版领域的专利、新闻、公众号、学术论文数据, 采用joiNLP算法提取关键短语, 利用BERTopic模型进行主题提取, 并基于all-MiniLM-L6-v2模型将主题向量化作为余弦相似度计算的依据, 以分析2006—2022年融合出版技术主题与应用主题的主题演化与内容演化趋势, 最终构建演化主路径。以多源数据为研究对象, 揭示融合出版领域热点技术可归纳为区块链、人工智能、虚拟现实、增强现实, 并指出未来融合出版将在元宇宙世界具有更广泛的应用空间。使用主题模型从文本语义层面提取技术主题, 结果具有高准确性和强可解释性, 对探究技术的演化规律与趋势具有创新意义。

关键词: 融合出版; 技术主题演化; BERTopic模型

中图分类号: G230.7; G252.8 **DOI:** 10.3772/j.issn.1673-2286.2023.04.002

引文格式: 曾江峰, 黄泳潼, 陈静怡, 等. 面向融合出版技术的主题演化研究[J]. 数字图书馆论坛, 2023(4): 9-18.

全球性科技革命蓬勃发展, 科技创新能力日益成为经济社会发展的决定性力量, 成为国家综合国力得以形成和提升的重要基础。出版业作为承担科技成果交流与传播、科研应用转化使命的基础行业, 服务于科技创新和学术传播, 在提升国家科技竞争力与文化软实力、维系科技事业传承方面发挥着巨大作用。然而随着人工智能、大数据等技术的发展与多媒介融合, 传统的出版形态难以适应当前愈发激烈的科技竞争需要。海量数据为出版选题提出挑战, 期刊影响力有限造成优质稿源流失^[1]等困境促使传统出版逐渐与新兴数字出版从“相加”走向“相融”, 形成了融合出版的新兴业态。围绕融合出版的定义, 出版界给出了技术^[2]、管理^[3]、产业^[4]等不同层面的理解。综合上述定义, 本文认为融合出版是指融合出版业内容、技术、媒介、平台等资源要素, 强化知识服务属性与出版业态创

新, 最终实现传统出版与数字出版相互促进、彼此融合的出版体系。融合出版作为传统出版转型变革的产物, 延续了出版事业传播信息与生产知识的功能, 不仅是期刊数字化、国际化、开放化转型和高质量发展的必由之路, 同时也是构建国际竞争背景下出版业话语权的必要手段。

为推动融合出版向纵深化方向发展, 2022年4月18日中宣部印发《关于推动出版深度融合发展的实施意见》^[5] (以下简称《实施意见》), 旨在健全传统出版与新兴出版“融为一体、合而为一”的体制机制, 实施出版融合发展工程, 探索融合发展新模式、新业态、新领域。《实施意见》具有很强的政策引领性和实践操作性, 其提出应着眼加强技术探索应用、促进成熟技术应用推广、健全科技创新应用体系, 充分发挥技术对出版融合发展的支撑作用。因此, 分析融合出版领域的技

收稿日期: 2023-03-31

^{*}本研究得到富媒体数字出版内容组织与知识服务重点实验室开放基金项目“面向融合出版的前沿技术主题演化及发展趋势预测研究”、教育部人文社会科学研究青年基金项目“情境大数据驱动的社交媒体虚假信息识别模型与治理策略研究”(编号: 21YJC870002)、中央高校基本科研业务费资助项目“信息交互行为与隐私保护研究”(编号: CCNU22QN017)、武汉市知识创新专项项目曙光计划项目“多源知识驱动的社交媒体虚假新闻检测研究”(编号: 2022010801020287)资助。

术主题演化路径,对推动我国融合出版创新发展、为文化强国提供支撑具有重要的战略意义。

1 文献回顾

目前以新兴技术为支撑构建深度融合发展路径引起了众多学者的关注:袁小群等^[6]探索了数字孪生技术与出版相融合的可能性,提出重塑出版业虚实共存的发展格局;延宏等^[7]基于元宇宙语境指出“VR+”将是推动我国传统出版与新兴业态深度融合的创新模式;张海生等^[8]探讨了人工智能与出版融合发展的基本理论与现实问题,提出人工智能与出版融合发展是出版业未来的发展方向。这些研究多聚焦于某一技术在深度融合出版的应用及趋势,缺少对融合出版领域技术应用与演化的整体把握。深入挖掘应用于融合出版的新技术的演化规律,对推动融合出版转型升级具有重要的实践意义。

近年来融合出版领域关于技术的已有研究中,领域技术现状研究多以论文集为数据源进行关键词词频分析、共现分析:孔薇^[9]使用关键词共现知识图谱来识别人工智能技术环境下学术期刊的热点主题和维度特征;刘静等^[10]基于关键词共现分析,揭示我国科技期刊融合出版研究的时空分布、研究热点及演变;周红利等^[11]利用高频关键词和聚类分析对出版融合相关文献进行分析。然而,这一类分析方法仍存在一定不足,例如:单将词频作为主题判定标准,依据并不充足^[12];依赖词间共现关系,未充分考虑关键词之间的语义关联关系;随着出版学科相关的数字资源和网络资源的持续性增长,仅以期刊文献为数据源容易造成信息缺失、知识偏向性等问题^[13],难以全面地反映技术焦点;基于先前经验划分时间阶段,缺乏实证依据,无法科学呈现主题之间的演化关系。

为了客观全面地揭示并认识技术在融合出版领域的应用现状,从数据中准确识别潜在的热点主题、追踪其演化路径并将其可视化,成为出版业界的研究重点,而基于主题模型的主题演化提供了有效的研究方案。基于主题模型的分析方法用于对文本中潜在的语义关系和主题信息进行挖掘,其在揭示研究主题分布、关联结构、演化规律等方面表现出明显优势。在中国知网以“SU=融合出版 AND 主题模型”为检索式进行检索,未得到结果,说明在融合出版领域,对于该主题演化分析应用仍是空白。

为弥补现有研究成果的不足,本文基于新闻报道、期刊论文、学位论文、会议论文、专利等多源异构数据,并进行数据清洗、去除停用词、词条统一化等预处理,基于Python提供的中文文本预处理工具包joiNLP对文档关键词进行提取。选取全文中具有代表性的内容,利用BERTopic模型中的all-MiniLM-L6-v2词嵌入模型提取词向量,并利用统一面域逼近与投影(Uniform Manifold Approximation and Projection, UMAP)算法与基于层次密度的噪声应用空间聚类(Hierarchical Density-Based Spatial Clustering of Applications with Noise, HDBSCAN)算法对主题进行聚类识别,再对不同时间窗口的文档进行主题词提取,计算余弦值以测度主题间演化关系,最终以桑基(sankey)图展现形式对主题内容演化和强度演化路径进行可视分析,以期多维度揭示技术演化规律,为深化融合出版实践提供参考。

2 研究方法

2.1 基于BERTopic模型的主题提取

主题提取是主题演化的基础和前提。目前潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)模型是最为常用的主题模型。然而这一传统模型认为词间相互独立,容易造成词间语义的缺失,难以抽取确切的关键技术主题^[14]。近年来,在用于主题演化的模型中,BERTopic模型表现出优于LDA模型的性能。BERTopic模型是利用all-MiniLM-L6-v2词嵌入模型、基于C-TF-IDF算法聚类的主题建模技术^[15],其采用的词嵌入模型考虑了词间语义关联,能够基于语义层面进行主题提取以弥补LDA模型不足,其主题提取的多样性与一致性指标表现优于其他主题模型^[16]。因此,基于BERTopic模型进行主题提取,其算法具体流程如下。

(1) 嵌入词摘要。采用BERTopic中的all-MiniLM-L6-v2模型进行词嵌入。all-MiniLM-L6-v2模型是Wang等^[17]基于深度学习算法sentence-transformers提出的词向量模型,其是由不同模型压缩得到的预训练模型,具有容量小、速度快等优点,能够捕捉词语之间的语义关系^[18]。

(2) 降维与聚类。BERTopic使用UMAP算法与HDBSCAN算法^[19]进行向量降维与主题聚类,这两个算法是BERTopic在没有事前指定主题数量情况下自动

提取最佳主题数的基础。相较于主成分分析(PCA)、t分布随机邻居嵌入(t-SNE)等常用的降维算法,UMAP能够应用于跨语言场景,并保持数据集局部和全局结构的低维性^[20]。HDBSCAN算法能够减少嵌入向量数量并在单个主题的语义空间中进行高度相似主题的迭代合并,有助于减少最终主题数量。

(3) 创建基于C-TF-IDF算法的主题表示。C-TF-IDF算法改进了基于类的词频-逆文档频率(TF-IDF)算法,测量的是单词在主题集的重要性而不是单个文档的重要性,有助于将主题数量减少为用户指定的数量,并使得BERTopic能够支持多种主题建模,如引导式主题建模、动态主题建模或基于类的主题建模^[21]。

2.2 基于融合出版领域的技术主题演化分析

主题演化是随着时间推移而呈现出的主题强度及关注内容的变化^[22],目前可分为基于主题内容和基于主

题强度的演化分析。其中,基于主题内容的演化分析描绘随着时间推移的主题内容变化的脉络与趋势^[23],主要通过事先界定演化状态(如主题的衰亡、新生、融合、分化等趋势)进行描述,一般依据主题提取的结果并计算不同相邻时间窗口的主题间相似度确定。而基于主题强度的演化分析又称为基于主题热度或主题关注度的演化分析,能够展现主题在某个特定时间窗口被关注的程度,主题强度是判断某研究主题是否为科研热点的量化指标^[22]。综上,基于主题演化状态界定与演化指标构建,进行主题演化分析。

2.2.1 主题演化状态界定

参照李慧等^[24]提出的主题演化状态界定的条件,界定融合出版技术主题的演化状态为生长、延续、衰减、新生、融合、分化与消亡,并基于计算主题内容与主题强度的指标进行演化分析,具体判定如表1所示。

表1 主题演化状态判定条件

主题演化类型	描述
生长	主题受关注度逐年上升
延续	主题受关注度趋于稳定
衰减	主题受关注度逐年下降
新生	新出现的主题,与先前时间窗口内的主题仅存在较低相关性或者不存在相关性
融合	融合主题与多个前置主题都有一定的相关性,是多个主题共同融合的结果
分化	从现有主题演化而来的新主题,与当前主题存在较高的相关性但并不十分相似
消亡	后续时间窗口生成的主题与现有主题均不存在相关性,或者相关性极低

2.2.2 主题演化指标构建

为描述相邻时间窗口下文档主题的演化状态和受关注度以构建演化路径,通过计算相邻时间窗口的主题相似度与主题强度,提取出上述的主题演化状态。

(1) 主题强度。具有较高强度的主题对应的受关注度较高,使用文档支持率^[25]度量主题强度。若某个时间窗口内某主题的文档支持率超过设定的阈值,即认为该主题属于热点主题。

(2) 主题相似度。主题相似度揭示主题的关联紧密程度,使用夹角余弦值表征主题相似度。余弦值越接近1,证明两个主题向量越相似。为去除弱关联和无效关联主题,设定主题相似度阈值为 ϵ :当两个主题的主

题相似度大于 ϵ 时,则认为其存在关联关系。

2.3 基于主题演化路径的可视化

目前主题演化路径主要通过桑基图^[26]进行可视化,使用Python提供的第三方库pyecharts绘制主题演化路径桑基图。论文整体技术路径如图1所示。

3 实验过程

3.1 数据集构建

通过融合多源数据,能弥补单一数据类型导致的

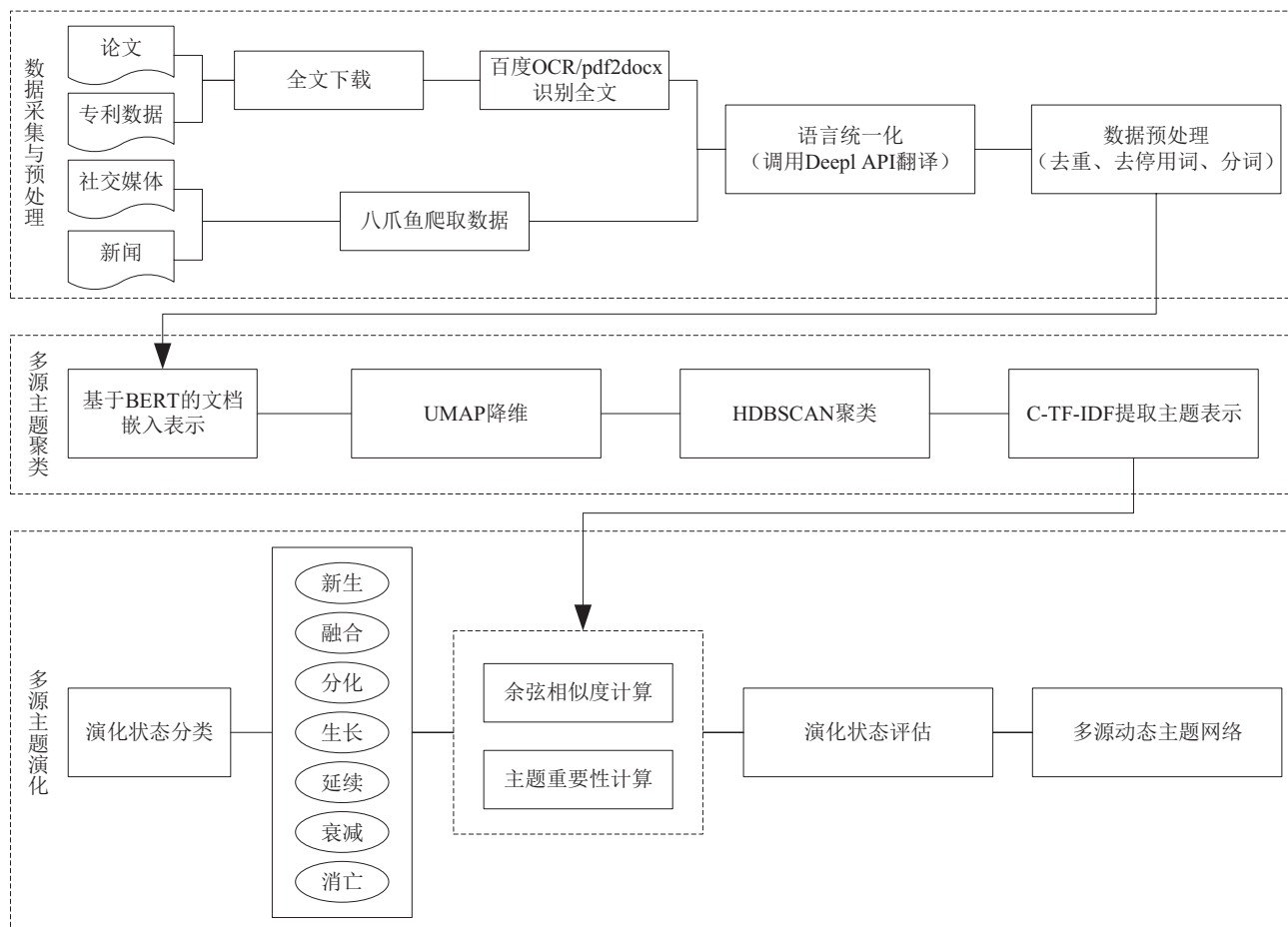


图1 技术路径

信息关联性不足的问题,有效挖掘技术领域复杂主题的关系与演变历程^[24],为突破现有融合出版领域研究的局限,故面向多源数据构建主题演化路径。考虑到融合出版作为出版新业态,有关其技术应用的文献数量少,而媒体融合作为融合出版的概念雏形^[27],应属于融合出版范围,故选择添加“媒体融合”作为拓展检索词。通过人工筛选,最终将属于技术应用的数据纳入分析范围,构建的数据集主要由以下4个部分组成。

(1) 新闻报道。新闻可反映主流媒体揭示的融合出版技术关注焦点。考虑权威性,以央视新闻、人民日报、光明网、中国新闻出版广电网为新闻数据的来源,经过人工筛选,最终确定122条有效数据。

(2) 微信公众号文章。微信作为新媒体平台,活跃账户数量达12亿个,已成为人们日常生活的一部分。考虑到其强大的传播影响力,以微信“搜一搜”为检索接口,最终获得72条有效数据。

(3) 论文。期刊论文能反映最新的基础研究情况;

学位论文专业性强,较为深入全面;会议论文内容新颖,前沿性和针对性强。以中国知网和Web of Science为论文数据的来源,通过人工筛选,纳入570篇中文文献与30篇英文文献。

(4) 专利。专利是反映行业技术创新的重要科技成果,以国家知识产权局网站、incoPat专利数据库与Derwent专利数据库为国内外专利数据的来源,通过查询《国际专利分类表(2022版)》,选定与出版业务相关的国际专利分类号G06F17/00、G06F40/00等进行检索,经人工筛选,最终得到128条有效记录。

3.2 数据采集与预处理

对于新闻报道和微信公众号文章数据,均采用八爪鱼爬虫采集器进行全文爬取。对于论文和专利数据,则分别下载全文pdf文件,使用百度OCR(Optical Character Recognition)接口或者Python第三方库

pdf2docx进行文字提取。使用Deep API的翻译接口, 统一将数据翻译为简体中文, 并结合jieba库等对每一篇文档进行去除特殊符号、分词、停用词等工作。采用全文爬取方式构建数据集, 导致无关内容较多, 直接使用数据集会影响模型识别效果, 故对部分无关内容较多的文档, 采用Python提供的第三方库joiNLP提取文本关键短语, 最终形成文本摘要。

3.3 主题识别与聚类

采用all-MiniLM-L6-v2的词向量模型对预处理的文本进行嵌入表示, 使用CountVectorizer函数对该数据的文本特征进行处理, 计算每篇文档中各主题出现的次数, 并对文档词频进行限制。在不设置聚类主题数

量的情况下, 根据BERTopic模型计算出的主题间相似度矩阵与UMAP、HDBSCAN算法合并相似度比较高的主题。由于HDBSCAN算法会产生异常值, 即部分文档不属于任何一个主题, 对于这部分文档, 计算它属于每个主题的概率, 将其归并到概率最高的主题中。

在主题数量上, 首先不调整参数, 使模型自动生成相应的主题, 共得到16个主题。随后, 使用UMAP算法, 将主题向量降到二维, 使用plotly库画出主题距离图, 如图2(a)所示。观察发现, 存在部分主题重叠的情况, 说明聚类效果还没有达到最优。依据主题相似度发现类似主题, 对这些主题进行合并(Topic#0和Topic#1、Topic#10和Topic#12), 最终提取得到15个主题。如图2(b)所示, 各个主题没有重叠, 在主题距离图上有明显的距离, 说明主题聚类效果良好。

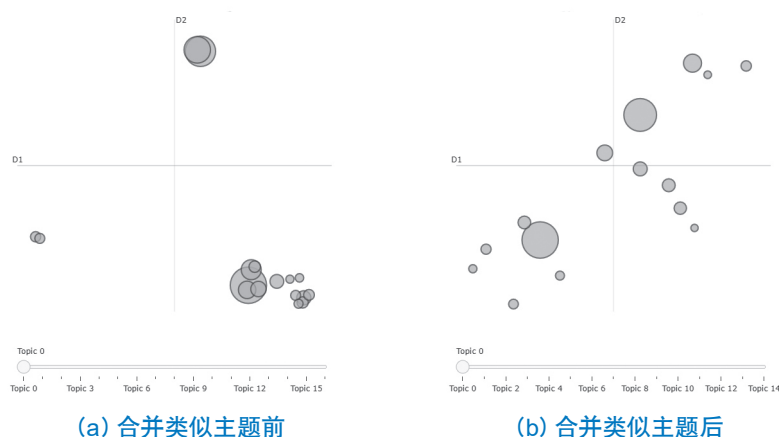


图2 主题距离图

经过主题聚类, 共提取15个主题, 分别重命名为区块链应用(Topic#0)、学术出版(Topic#1)、媒体融合出版(Topic#2)、人工智能在教育出版的应用(Topic#3)、互联网选题策划(Topic#4)、机器学习在教育出版的应用(Topic#5)、沉浸式数字技术应用(Topic#6)、期刊编辑(Topic#7)、主题出版(Topic#8)、虚拟现实意识形态出版应用(Topic#9)、增强现实在有声出版的应用(Topic#10)、虚拟现实在少儿科普出版的应用(Topic#11)、新闻出版(Topic#12)、精品出版(Topic#13)、红色党史教育出版(Topic#14)。

3.4 主题强度演化分析

结合不同时间窗口主题文档支持率绘制主题

强度变化图谱, 最终结果如图3所示。可以看出, 主题整体呈现逐年生长、延续的趋势, 无衰减主题, 这表明近年来融合出版得到广泛关注与实践, 传统出版与新兴出版融合协同进程不断提速, 区块链、增强现实等技术与学术出版、教育出版等领域不断融合。部分主题强度在2017—2020年呈现较大幅度波动, 如区块链应用(Topic#0)、学术出版(Topic#1)、媒体融合出版(Topic#2)、增强现实在有声出版的应用(Topic#10), 政策与技术发展为上述主题强度波动提供解释。

2016年工业和信息化部发布的《中国区块链技术和应用发展白皮书》^[28]将区块链技术上升至国家科技战略水平, 基于区块链的比特币市值超25亿美元, 进一步加快了区块链技术的成熟与应用落地。在国家战略

政策和良好市场前景的双重推动下,区块链技术成为业界关注的科技热点,出版领域也积极将区块链技术广泛应用于新闻溯源、版权保护、选题策划、知识服务、智库建设等场景,因此引起了区块链应用(Topic#0)这一主题强度的变化。

2018年中央全面深化改革委员会第五次会议通过了《关于加强和改进出版工作的意见》^[29],明确指出要以建设世界一流科技期刊为目标,做精做强一批基础和传统优势领域期刊。在数字技术的驱动下,学术出版业逐渐向以学术出版为基础、以学术服务为目的、以学术资源整合开发为核心的出版模式转变,这些进展也使得学术出版(Topic#1)这一主题有更高的关注度。

2019年,科技部等部门印发《关于促进文化和科技深度融合的指导意见》^[30],提出要推动媒体融合向纵深发展,促进文化和科技深度融合。2021年底,国家新闻出版署发布《2020年新闻出版产业分析报告》^[31],从顶层设计上指明了出版行业的发展方向,要求出版行业融合多种媒介,创新新闻出版的发展模式。由此,出版单位积极开拓线上业务,推动出版业务数据化,推出电子书、有声书等产品。因此,媒体融合出版(Topic#2)这一主题强度在2018—2022年有较大的起伏。

2018年,第42次《中国互联网络发展状况统计报告》^[32]显示,国内有声阅读用户规模已达2.32亿人,巨大的用户存量表明有声市场有巨大的发展潜力。2020年,腾讯音乐加入有声市场,同时微信推出了“微信听书”功能。随着增强现实技术的发展,增强现实立体有声阅读、增强现实扫读等服务相继出现,并且被运用在

图书馆、展览馆中^[33]。巨大的市场潜力与技术的发展使得增强现实在有声出版的应用(Topic#10)有更高的关注度。

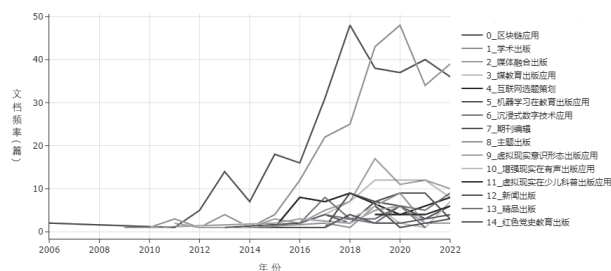


图3 不同时间窗口主题支持文档数量

3.5 主题内容演化分析与演化路径

2006—2016年融合出版领域正处于萌芽期,该领域成果产出主要集中于2017—2022年。为避免时间窗口划分的主观经验性,根据BERTopic模型提取的主题在各年份的分布结果,将2006—2016年为1个时间段,以2年为单位将2017—2022年分段,共得到4个时间段。将特定时间窗口所有主题相似度的平均值设置为阈值 ϵ ,若主题相似度大于该阈值,则认为主题间存在演化关系,最终结合余弦值计算结果构建演化路径,如图4所示,其中Y0~Y3表示时间窗口,T0~T14表示主题。

根据演化路径,可观察到2006—2022年无衰亡主题,大部分主题在4个时间窗口内均呈现出新生、融合、分化的演化状态,主要包括区块链应用(Topic#0)、学术出版(Topic#1)、媒体融合出版(Topic#2)、互联

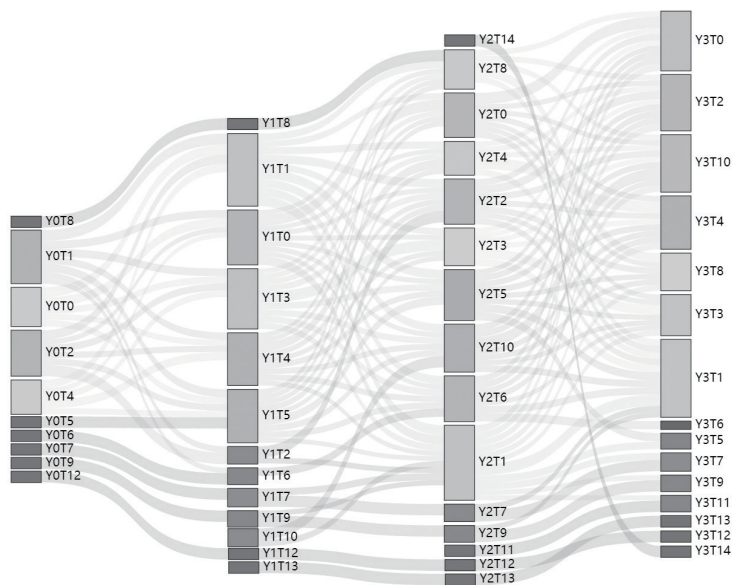


图4 多源主题演化路径

网选题策划 (Topic#4)、机器学习在教育出版的应用 (Topic#5)、沉浸式数字技术应用 (Topic#6)、期刊编辑 (Topic#7)、主题出版 (Topic#8)、虚拟现实意识形态出版应用 (Topic#9)、新闻出版 (Topic#12)。上述主题是融合出版领域的重点, 在近年来政策与科技发展的共同推动下成为出版转型的助推力。

在2017—2018年与2019—2020年主题精品出版 (Topic#13), 2019—2020年红色党史教育出版 (Topic#14)、虚拟现实在少儿科普出版的应用 (Topic#11) 在演化过程中突然出现并延续, 属于新生主题。结合国家政策, 认为国家新闻出版署对于主题出版和精品出版的指导工作为该演化提供解释。自2016年原国家新闻出版广电总局启动“中国文艺原创精品出版工程”以来, 各机构围绕体现党和国家意志与社会需要的主题, 融合虚拟现实 (Virtual Reality, VR) 技术、全媒体平台等元素, 挖掘中国共产党与传统文化深厚底蕴, 产出一系列创新出版精品, 例如AR/VR互动内容运营商“博乐信息”推出的“党建书屋”应用, 融合增强现实、人机交互等技术为用户营造新型党建阅读体验。这表明技术正赋予主题出版新的发展机遇, 使得传统主题出版与以技术为驱动的新型主题出版迈向更深层次的融合阶段。同时, 2017年原国家新闻出版广电总局发布主题出版物, 强调少儿题材出版的重要性, 突显我国对于少儿出版的重视, 这也从侧面验证了本次主题演化结果与实际情况吻合。

在2021—2022年主题分化与融合的趋势显著提高, 如学术出版 (Topic#1) 与区块链应用 (Topic#0)、媒体融合出版 (Topic#2)、人工智能在教育出版的应用 (Topic#3)、互联网选题策划 (Topic#4)、机器学习在教育出版的应用 (Topic#5)、沉浸式数字技术应用 (Topic#6)、期刊编辑 (Topic#7)、虚拟现实意识形态出版应用 (Topic#9) 多个主题融合。学术出版行业在近两年向跨技术、跨平台方向发展, 单一的技术应用已难以满足其数字化与智慧化的转型需求, 融合成为新的出版趋势, 如学术出版领域利用人工智能技术辅助选题策划^[34]、依托融媒体实现出版传播^[35]、使用虚拟现实技术提供交互式阅读体验^[36]等。

4 技术主题识别与结果验证

基于BERTopic模型的主题提取结果, 可得出2006—2022年区块链、人工智能、虚拟现实、增强现实

与融合上述技术的产物元宇宙成为融合出版领域的热点技术。为验证主题演化结果的有效性与可靠性, 采用文献回顾与出版实践分析法阐释演化结果中出现的主题热点。

4.1 区块链在融合出版中的应用

区块链是用块链结构记录数据, 按照数据产生的先后顺序存储, 并以共识机制进行区块间的验证、以密码学保证链上数据不可篡改的分布式数字账本。根据区块链应用 (Topic#0) 的下位主题词“加密”“数字出版”, 发现区块链广泛应用于数字出版的加密保护。利用区块链的非对称加密技术, 能够追溯版权和隐私信息, 降低作品著作权受侵害的风险。同时根据区块链应用 (Topic#0) 在时间窗口内的内容演化路径, 可发现随着元宇宙元年的到来, 相关主管部门、出版商、高校和科研机构积极探索应用场景, 区块链在作者开放研究者与贡献者身份识别码 (ORCID) 认证和论文防篡改^[37] (Y1T1-Y3T1)、跨媒体内容生产^[38] (Y1T2-Y3T2)、图书著作信息透明与精准选题策划^[39] (Y1T4-Y3T4)、教材版权保护^[40] (Y1T5-Y3T5)、数字藏品版权保护^[41] (Y2T8-Y3T8) 和有声出版物增值^[42] (Y2T10-Y3T10) 方面均得到应用。虽然目前出版机构尚未在大范围内应用区块链技术, 但区块链及其出版构想已在出版界引起广泛讨论, 未来势必拓展到更深层次的应用场景, 从而推动数字出版领域的深度融合。

4.2 人工智能在融合出版中的应用

人工智能作为海量数据时代知识发现和价值挖掘的前沿工具, 在融合出版领域发挥着不可替代的作用, 并被广泛应用于出版转型实践。根据人工智能在教育出版的应用 (Topic#3) 和机器学习在教育出版的应用 (Topic#5) 的下位主题词“学术出版”“教育出版”, 可识别出人工智能技术在学术出版和教育出版领域, 在选题策划、在线教育资源、教材出版等方面发挥作用, 例如: 根据当前学科热点及焦点问题, 结合读者需求, 制定具有前瞻性、引导性、实用性的选题方向、选题形式及内容; 利用机器学习和数据挖掘技术提供在线教育资源的智能个性化推荐、开展高等教材教辅的内容生产, 深度挖掘用户个性化需求以开展精准知识服务。同时从虚拟现实意识形态出版应用 (Topic#9) 下位词

“人工智能”可以得到,人工智能作为虚拟现实的重要支撑技术,正呈现出与虚拟现实技术融合发展的趋势。人工智能强大的数据挖掘功能和虚拟现实沉浸式场景化的技术特点共同赋能意识形态出版,打造兼具知识型、娱乐型的多重产品功能生态。种种实践均说明人工智能技术正赋予出版业态新活力,在提高出版效率的同时,利用自然语言处理等技术满足用户深度研读需求、提供个性化服务,使得现有出版模式逐渐向智慧出版模式发展。

4.3 虚拟现实在融合出版中的应用

虚拟现实技术是结合仿真技术与计算机图形学、人机接口技术、传感技术、多媒体技术来提供虚拟环境的技术。根据虚拟现实意识形态出版应用(Topic#9)和虚拟现实在少儿科普出版的应用(Topic#11)的下位主题词“童书出版”“新闻出版”,近年来虚拟现实技术被广泛应用于少儿科普出版、新闻出版等领域。利用强互动性与平台延展性等技术特性,虚拟现实技术能够构建不同的阅读场景,为读者提供一个复制真实或构想的沉浸式体验,并改变传统出版原有的纸质叙事模式,使得出版内容以更丰富多元的形态呈现,如山东教育出版社的《大开眼界:恐龙世界大冒险》虚拟现实系列图书,通过打造全景视频,为少儿读者展现了虚拟立体的恐龙世界,消弭了现实与虚拟环境边界,实现读者与出版内容的实时交互。同时,在2017—2022年,虚拟现实技术应用趋势凸显,在实现科技期刊出版增值^[36](Y2T6-Y3T1)、优化教育出版物质量^[43](Y2T6-Y3T3)、打造交互式图书产品^[44](Y2T6-Y3T4)、数字阅读推广^[45](Y2T6-Y3T8)方面发挥重要作用。虚拟现实技术创新了出版者与读者原有的交互模式,未来必然在出版工作中带来创造性的发展机遇。

4.4 增强现实在融合出版中的应用

增强现实是将虚拟信息与真实世界巧妙融合的技术,广泛运用了多媒体、传感等技术手段,将计算机生成的文字、图像、视频等虚拟信息模拟仿真并应用到真实世界。增强现实在有声出版的应用(Topic#10)的下位主题词“有声出版”“盲文出版”显示,增强现实技术赋予相关出版行业转型新的活力。利用独特的叙事模式与3D交互技术,增强现实技术构建了真实、立体

的阅读空间,并大大增强用户的沉浸式体验感,被应用于出版与其他元素的融合出版实践,如索尼公司推出的《魔法书》利用虚拟现实、增强现实技术,使得少儿读者可以通过语音输入方式与书里的虚拟人物互动,为读者带来新的阅读体验。此外,从演化路径来看,在2021—2022年增强现实技术应用趋势凸显,使得学术出版(Y1T10-Y3T1)、媒体融合出版(Y2T10-Y3T2)、教育出版(Y2T10-Y3T3)、主题出版(Y2T10-Y3T8)等行业的数字内容产品形态日渐多样,如《纽约客》纸质期刊增强现实封面、《纽约时报》打造的“NYT VR”新闻客户端等,同时增强现实技术在辅助图书出版物的选题策划(Y2T10-Y3T4)方面发挥重要作用。增强现实技术以其独特的技术优势促进传统出版物形态向全媒体形态转变,为出版全流程实现对声音、媒介、空间等元素融合提供有力支撑,未来在出版中将发挥更深层次的作用。

4.5 元宇宙在融合出版的应用

元宇宙是融合区块链、人工智能、虚拟现实、增强现实等技术的数字虚拟世界。根据沉浸式数字技术应用(Topic#6)的下位主题词及其演化路径可观察到,作为近两年融合出版的热点,元宇宙融合多种技术,正在重构出版行业的生态链,展现出更具有包容性的出版新业态:一是为出版产业提供更具前瞻性的技术发展思路,如人工智能技术优化学术出版的投审稿流程;二是依托技术融合实现多模态的出版物,如借助虚拟现实与增强现实技术推动内容生产模式的交互化、融合化^[6],产出如有声图书、融图文声像于一体的出版物等;三是创新用户服务方式,如借助多融合技术的泛在化特点提供数字阅读场景等。目前已有出版机构(如武汉理工数字传播工程有限公司)投入元宇宙出版实践,而未来融合出版必然在元宇宙世界具有更广阔的应用空间。

5 结语

本文选择专利、论文等多源数据,基于BERTopic模型进行2006—2022年融合出版领域的技术主题演化研究,发现在科技高速发展的时代背景下,区块链、人工智能、虚拟现实等热点技术已在融合出版领域表现出巨大的发展空间与潜力,其促进了出版资源的整合,

提高出版服务的响应速度, 精准服务用户, 提升了出版传播话语权, 正在重构传统出版与数字出版的融合态势。面临全新传播语境, 技术必然助力出版行业实现从传统出版到数字出版、智慧出版的模式转变, 推进出版流程各要素在该领域的深度融合, 为塑造国家科技竞争力和文化软实力提供支撑。本文是对融合出版领域的技术主题演化的初步探索, 仍然存在一些方面的不足与局限: 文档数据量较少, 对于主题提取效果产生一定影响; 依赖主题相似度计算, 忽略了语义。后续研究应扩大数据来源范围, 使用能够揭示语义关联的模型, 以弥补相似度计算方法造成的语义缺失。

参考文献

- [1] 何春娥, 吴浩, 朱晓华. 中文科技期刊应服务国家重大战略需求: 以《自然资源学报》为例[J]. 编辑学报, 2022, 34(1): 26-30.
- [2] 曹继东. 基于数字化技术和互联网思维的“融合出版”[J]. 科技与出版, 2014(9): 15-18.
- [3] 全国科学技术名词审定委员会. 全国科技名词委组织召开融合出版概念及定义审定会议[EB/OL]. [2022-11-12]. http://www.cnterm.cn/xwdt/tpxw/202201/t20220114_678525.html.
- [4] 李文. 论出版融合发展的前提、本质与路径[J]. 出版科学, 2021, 29(3): 58-63.
- [5] 国家新闻出版署. 中共中央宣传部印发《关于推动出版深度融合发展的实施意见》的通知[EB/OL]. [2022-11-12]. <https://www.nppa.gov.cn/nppa/contents/279/103878.shtml>.
- [6] 袁小群, 关灵爽. 虚实共生: 数字孪生技术赋能出版及融合发展研究[J]. 出版发行研究, 2022(2): 29-34.
- [7] 延宏, 王华. “VR+”: “元宇宙”视域下出版业的融合发展模式——以青岛出版集团为例[J]. 出版广角, 2021(22): 80-82.
- [8] 张海生, 吴朝平. 人工智能与出版融合发展: 内在机理、现实问题与路径选择[J]. 中国科技期刊研究, 2019, 30(3): 225-231.
- [9] 孔薇. 人工智能环境下学术期刊的融合出版: 热点主题、维度特征和发展路径[J]. 中国编辑, 2021(4): 39-44.
- [10] 刘静, 翟亚丽, 朱琳, 等. 我国科技期刊融合出版研究现状: 基于CiteSpace的知识图谱分析[J]. 中国科技期刊研究, 2019, 30(5): 487-493.
- [11] 周红利, 冷怀明. 我国出版融合发展的研究热点与态势: 基于CiteSpace的可视化分析[J]. 今传媒, 2022, 30(4): 134-137.
- [12] 谭春辉, 熊梦媛. 基于LDA模型的国内外数据挖掘研究热点主题演化对比分析[J]. 情报科学, 2021, 39(4): 174-185.
- [13] 刘怀兰, 刘盛, 周源, 等. 基于多源文本挖掘的技术演化路径识别[J]. 情报理论与实践, 2022, 45(11): 178-187.
- [14] 王秀红, 高敏. 基于BERT-LDA的关键技术识别方法及其实证研究: 以农业机器人为例[J]. 图书情报工作, 2021, 65(22): 114-125.
- [15] WANG Y F, HUANG H Y, RUDIN C, et al. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization[J/OL]. CoRR, 2020. <https://arxiv.org/abs/2012.04456>.
- [16] EGGER R Y J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts[J/OL]. Front Sociol: 1-16 [2022-12-11]. <https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full>.
- [17] WANG W, BAO H, HUANG S, et al. MiniLMv2: multi-head self-attention relation distillation for compressing pretrained transformers[J/OL]. CoRR, 2020. <https://arxiv.org/abs/2012.15828>.
- [18] ALHAJ F, AL-HAJ A, SHARIEH A, et al. Improving Arabic cognitive distortion classification in twitter using BERTopic[J]. Computación y Sistemas, 2020, 24(4): 1403-1414.
- [19] CAMPELLO R J G B, MOULAVI D, SANDER J O R. Density-based clustering based on hierarchical density estimates[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 160-172.
- [20] MCINNES L, HEALY J, MELVILLE J. UMAP: uniform manifold approximation and projection for dimension reduction[J]. The Journal of Open Source Software, 2018, 3(29): 861.
- [21] ABUZAYED A, AL-KHALIFA H. BERT for Arabic topic modeling: an experimental study on BERTopic technique[J]. Procedia Computer Science, 2021, 189: 191-194.
- [22] 梁爽, 刘小平. 基于文本挖掘的科技文献主题演化研究进展[J]. 图书情报工作, 2022, 66(13): 138-149.
- [23] 隗玲, 许海云, 胡正银, 等. 学科主题演化路径的多模式识别与预测: 一个情报学学科主题演化案例[J]. 图书情报工作, 2016, 60(13): 71-81.
- [24] 李慧, 胡吉霞, 佟志颖. 面向多源数据的学科主题挖掘与演化分析[J]. 数据分析与知识发现, 2022, 6(7): 44-55.
- [25] 徐佳俊, 杨飏, 姚天昉, 等. 基于LDA模型的论坛热点话题识别和追踪[J]. 中文信息学报, 2016, 30(1): 43-49.
- [26] 陈稳, 陈伟. 科学与技术对比视角下的前沿主题识别与演化分析[J]. 情报杂志, 2022, 41(1): 67-73, 163.
- [27] 任娇蕊, 王珏, 初景利. 学术期刊融合出版的实践与思考: 以《图

- 书情报工作》为例[J]. 数字图书馆论坛, 2022(10): 35-40.
- [28] 李鸣, 李佳祢, 孙琳. 区块链标准化现状及思路[J]. 中国信息安全, 2018, 101(5): 96-98.
- [29] 中央全面深化改革委员会第五次会议通过《关于加强和改进出版工作的意见》[J]. 出版与印刷, 2018(4): 50.
- [30] 中华人民共和国科学技术部. 科技部等六部门印发《关于促进文化和科技深度融合的指导意见》的通知[EB/OL]. [2023-04-11]. http://www.gov.cn/xinwen/2019-08/27/content_5424912.htm.
- [31] 中国新闻出版研究院. 2020年新闻出版产业分析报告[EB/OL]. [2023-04-11]. <https://www.chinaxwcb.com/info/576485>.
- [32] 中国互联网络信息中心. CNNIC发布第42次《中国互联网络发展状况统计报告》[EB/OL]. [2023-04-11]. http://www.cac.gov.cn/2018-08/20/c_1123296859.htm?ivk_sa=1024320u.
- [33] 赵礼寿, 马丽娜, 杨佚琳. 我国有声读物出版产业发展模式创新研究[J]. 出版广角, 2022(21): 71-75.
- [24] 刘平, 杨志辉. 人工智能构建科技期刊智慧出版模式[J]. 中国科技期刊研究, 2019, 30(5): 462-468.
- [35] 艾岚, 李金霞. 媒体融合视域下学术期刊全媒体出版传播体系的构建[J]. 中国编辑, 2021(1): 62-66.
- [36] 刘星星, 崔金贵. 5G时代科技期刊融合VR出版: 优势、挑战及对策[J]. 中国科技期刊研究, 2021, 32(8): 1026-1031.
- [37] 张小强, 曹馨予. 区块链技术赋能学术期刊深度融合发展的模式与路径研究[J]. 中国科技期刊研究, 2022, 33(8): 999-1012.
- [38] 温凤鸣, 解学芳. 基于区块链赋能的短视频内容产业创新与优化路径[J]. 中国编辑, 2021(9): 81-86.
- [39] 姚顺, 张彬. 区块链技术赋能图书出版业应用场景研究[J]. 中国出版, 2022(2): 51-54.
- [40] 刘九如. 新技术赋能, 数字出版呈现新业态[J]. 出版广角, 2022(9): 44-49.
- [41] 刘玉柱, 李广宇. 数字藏品版权保护问题研究[J]. 出版广角, 2022(11): 47-51.
- [42] 沈悦, 金圣钧. 智媒时代有声出版的作用机理及优化路径[J]. 中国编辑, 2022(11): 86-91.
- [43] 胡玉玺, 程海威. 数字化转型中教育出版的沉浸式场景构建[J]. 出版发行研究, 2021(4): 5-12.
- [44] 王薇. 技术赋能交互式图书出版的逻辑与路径[J]. 出版广角, 2022(22): 79-82.
- [45] 刘宇, 刘颖. 基于虚拟现实出版的儿童数字化阅读推广研究[J]. 出版科学, 2019, 27(6): 98-101.

作者简介

曾江峰, 男, 1988年生, 博士, 讲师, 研究方向: 虚假新闻检测。

黄泳潼, 女, 1999年生, 硕士研究生, 研究方向: 虚假新闻检测。

陈静怡, 女, 2002年生, 本科生, 研究方向: 信息政策分析。

冯昌扬, 女, 1991年生, 博士, 讲师, 通信作者, 研究方向: 信息政策分析, E-mail: cyfeng@ccnu.edu.cn。

Theme Evolution for Technology of Integrated Publishing

ZENG JiangFeng¹ HUANG YongTong¹ CHEN JingYi¹ FENG ChangYang^{1,2}

(1. School of Information Management, Central China Normal University, Wuhan 430079, P. R. China; 2. Rich Media Key Laboratory of Digital Publishing Content Organization and Knowledge Service, Beijing 100038, P. R. China)

Abstract: This paper aims to master the application status and trend of cutting-edge technologies in the field of integrated publishing, and provide references for the technology application in the field of integrated publishing. The data of patents, news, public accounts, and academic papers in the field of integrated publishing from 2006 to 2022 years are selected, the key phrases are extracted by joiNLP algorithm, the topics are extracted by BERTopic model, and the topic vectorization is used as the basis for cosine similarity calculation based on all-MiniLM-L6-v2 model. This paper analyzes the intensity evolution and content evolution trend of the integrated publishing technology theme and application theme from 2006 to 2022, and finally builds the evolution main path. Taking multi-source data as the research object, it reveals that the hot technologies in integrated publishing field can be summarized as block chain, artificial intelligence, virtual reality, and augmented reality, and points out that integrated publishing will have a wider application space in the metaverse world in the future. By using topic model to extract technical topics from the semantic level of text, the results have high accuracy and strong interpretability, which have innovative significance to explore the evolution law and trend of technology.

Keywords: Integrated Publishing; Technology Topic Evolution; BERTopic Model

(责任编辑: 雷雪)