doi:10.3772/j.issn.2095-915x.2023.05.010

面向融合出版前沿主题发现的学术论文未来工作句挖掘研究

开放科学（资源服务）标识码（OSID）

谢林蕾[1] 向熠[1] 章成志[1,2]

1. 南京理工大学经济管理学院信息管理系 南京 210094
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：[目的 / 意义] 近年来，随着传统出版与数字出版的不断融合，形成了融合出版的新兴范式。如何科学准确地把握融合出版领域未来研究趋势具有重要研究意义。学术论文中描述未来研究工作的句子（简称“未来工作句”），不但可以辅助预测未来可能出现的前沿主题，还可为科研工作者、特别是初学者选题提供参考。[方法 / 过程] 对融合出版领域论文中的未来工作句进行人工标注和类别划分，构建未来工作句识别与分类语料库。在此基础上，使用支持向量机、朴素贝叶斯和随机森林三种模型结合 SelectKBest 特征选择方法，来训练未来工作句自动识别模型。[结果 / 结论] LinearSVC 在未来工作句自动识别任务中表现最为出色，其加权 F_1 值达到 92.08%。另外，本文对分类语料库中的未来工作句内容及其类别进行分析，得到融合出版领域未来工作句的类别分布及其变化规律。

关键词：融合出版；未来工作句；机器学习；文本分类；内容分析

中图分类号：G35

Research on Future Working Sentences Mining of Academic Papers for Frontier Topics Discovery of Integrated Publishing

XIE Linlei[1] XIANG Yi[1] ZHANG Chengzhi[1,2]

1. Department of Information Management, School of Economics & Management, Nanjing University of Science and Technology, Nanjing 210094, China
2. Rich Media Key Laboratory of Digital Publishing Content Organization and Knowledge Service, Beijing 100038, China

Abstract: [Objective/Significance] In recent years, the continuous integration of traditional publishing and digital publishing has given birth to an emerging paradigm known as integrated publishing. Understanding the future research trends in this field is of

基金项目 2022 年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发现研究”(zd2022-10/02)。

作者简介 谢林蕾 (2001-)，硕士研究生，研究方向为文本挖掘；向熠 (2001-)，博士研究生，研究方向为学术文本挖掘；章成志 (1977-)，博士，教授，研究方向为信息组织，信息检索、文本挖掘和自然语言处理。

引用格式 谢林蕾，向熠，章成志 . 面向融合出版前沿主题发现的学术论文未来工作句挖掘研究 [J]. 情报工程，2023, 9(5): 123-138.

doi:10.3772/j.issn.2095-915x.2023.05.010

great research significance. The sentences in academic papers that describe future research work, also known as “future working sentences”, can not only predict potential future topics, but also guide researchers, especially beginners, in selecting their topics. [Methods/Processes] This study carries out artificial labeling and classification of future working sentences from papers related to integrated publishing and constructs a corpus for recognizing and classifying future working sentences. Subsequently, three models - Support Vector Machine, Naïve Bayes, and Random Forest – combined with the SelectKBest feature selection method are utilized to train an automated recognition model for future working sentences. [Results/Conclusions] The experimental results indicate that LinearSVC offers the best performance in the task of automatic recognition of future working sentences, achieving a weighted F_1 value of 92.08%. Furthermore, this paper analyzes the content and categories of future working sentences in the taxonomic corpus, revealing the category distribution and trends of such sentences within the field of integrated publishing.
Keywords: Integrated Publishing; Future Work Sentence; Machine Learning; Text Classification; Content Analysis

引言

近年来，随着数字化技术的快速发展，传统出版业愈发受到数字出版的冲击，它们开始主动寻求与数字出版的“融合”[1]。在此背景下，“融合出版”模式应运而生。作为一种新兴的出版模式，融合出版可实现传统出版与数字出版的共赢。借助互联网等技术传播快、成本低、影响力大的优势[2]，融合出版可改善传统出版内容有限、发行渠道单一、出版周期较长且更新较慢的缺点[3-4]，更好地促进知识传播。自 2014 年以来，融合出版开始引起学者的关注。之后，关于融合出版的相关研究文献数量不断增长。根据关键词知网节的统计结果显示，知网文献库中每年发表的中文文献数量从 2017 年前的不足 10 篇已增长到现今的 100 篇以上，2022 年相关文献数量已达 125 篇①。面对融合出版领域科技文献数量爆炸式增长的现象，前沿研究愈发表现出其必要性。然而，如何对研究前沿进行更为精准的定位是科研工作者在科研选题时经常面临的问题。未来研究能够尽早捕捉研究领域未来的新兴与热点话题，帮助科研人员确定研究领域与对象，有针对性地开展研究。

学者通常在学术论文最后提出论文的未来研究工作展望，表明今后进一步的研究工作方向。本文将论文中描述未来研究工作的句子称为“未来工作句”。系统性梳理和归纳特定领域的未来工作句，可以辅助我们更好地预测该领域的未来发展趋势，为科研选题等提供有价值的参考。

本文以融合出版领域为研究对象，从该领域的学术论文中抽取未来工作句并进行分类，构建未来工作句识别与分类语料库，并在此基础上训练机器学习模型从而实现未来工作句的自动识别。另外，本文基于未来工作句分类语料库对不同类别进行分析，深入探究该领域的发展状况。该文研究可为融合出版领域未来发展提供一定的借鉴。

①检索日期为 2023 年 8 月 16 日

1 相关研究概述

与本文相关的研究包括融合出版研究和未来工作句研究，本节从这两个角度对相关工作进行概述。

1.1 融合出版研究概述

国内对“融合出版”的针对性研究要追溯到 2014 年 [5]。曹继东 [5] 提出，“融合出版”是在“媒介融合”学术语境下，顺应中国出版融合发展趋势，基于数字化技术和互联网思维产生的新兴出版现象，是解决传统出版和数字出版融合发展问题的新兴出版范式。

此前，国内在数字技术融合与出版创新这一领域的研究对象主要是媒介融合与出版融合。媒介融合 (Media Convergence) 的概念源于美国，1983 年美国马萨诸塞州理工大学的浦尔教授 [6] 在其《自由的科技》一书首次正式提出媒介融合的概念。2005 年，媒介融合的概念经蔡雯教授引入到国内。在她发表的有关“媒介融合”与“融合媒介”的文章中，引入了美国新闻学会媒介研究中心主任 [7] 对“融合媒介”的定义——印刷的、音频的、视频的、互动性数字媒体组织之间的战略的、操作的、文化的联盟。然而，媒介融合与融合出版存在较大差异。媒介融合的研究主要侧重在新闻传媒业和电信业等的融合，较少涉及书刊出版业 [8]。在该思想启发下，从出版的视角出发，出版业也被指出有探索和实践融合发展的需要与必然 [9]。

2010 年起，国内逐渐兴起关于数字出版与传统出版融合的研究。2011 年，《新闻出版业“十二五”时期发展规划》等都将数字出版纳入重要扶持领域，加快传统出版与数字出版的融合发展成为“十二五”时期产业发展的新目标 [10]。2012 年，开始有“出版融合”这一概念。在新闻出版体制改革和媒介融合背景下，出版融合成为新闻出版业的发展方向。但是同样，其与“融合出版”仍是两个不同的概念，前者侧重于“融合”而后者则更侧重于“出版”[11]。

融合出版作为一种建构在数字化技术和互联网平台基础上的新兴出版范式 [5]，要求传统出版与新兴出版在内容、渠道、平台、经营、管理等方面进行深度融合 [12]。融合出版的目的在于实现出版内容、技术应用、平台终端、人才队伍的共享融通，从而构建组织结构、传播体系和管理体系一体化发展路径 [8]。但这个新名词提出之后的几年间相关研究寥寥无几。据知网文献库的中文文献统计结果显示，2019 年后，国内有关“融合出版”的研究数量才有了较大幅度的增长，2020 年发表的文献数量已超百篇，且较 2019 年几乎翻了一番。至今，融合出版已成为较为热门的研究话题。越来越多研究者关注融合出版背景下优秀人才的培养 [13-16]，图书编辑的素养提高 [17-20] 以及版权保护工作等 [21]。然而，融合出版领域发展速度还是相对较慢，传统出版业的数字化转型正面临比较大的技术困境 [22]，人才建设和机制创新上也有待突破。因此，我们需要更加精准定位该领域的未来研究方向，从而促进该领域更好更快发展。但目前国内几乎还没有研究者关注融合出版的未来研究，为弥补这一缺失，本文将从此角度出发，结合机器学习对融合出版领域进行深入探索。

doi:10.3772/j.issn.2095-915x.2023.05.010

1.2 未来工作研究概述

国内外现如今针对未来工作方面的研究数量相对较少。Hu 等 [23] 在 2015 年以信息检索、文本挖掘和数字图书馆领域为例，开展未来工作挖掘问题。他们通过一种基于正则表达式的方法抽取学术文本中的未来工作句，并将其定义为问题、方法、评估和其他四个类别，通过对比不同特征与机器学习模型的组合，实现不同领域的未来工作句分类。这是对论文中的未来工作句展开的首次探索，其创新性研究成果极大地推动了未来工作的开展。

随后，Li 等 [24] 利用人工设定规则来识别未来工作句，从中提炼出关键词并与标题和摘要中关键词进行匹配，从而得到不同领域文献与未来工作二者的概念上的联系。Zhu 等 [25] 使用深度学习模型 BERT 对 2006—2016 年间 JASIST 期刊论文上的 1579 篇论文进行未来工作句抽取，并用层次聚类方法确定了未来工作句的四种类别，即支持性的、方法性的、识别潜在影响因素的和提出未来目标的。之后，也有一些研究人员开始使用规则匹配和 BERT 相结合的方法来提取未来工作句 [26]。近几年 Zhang 等 [27] 使用机器学习模型对 NLP 领域学术论文的未来工作句进行研究，成功训练出具有较优性能的自动识别与分类模型 [25, 27-29]。但总体来看，针对未来工作句的研究数量较少，其大多是基于规则和统计的方法。基于规则的方法的优势在于分类精度高，操作也比较灵活方便，但规则必须具备足够的代表性。而且，随着类目的扩大，需要设置的规则数量也会增加，从而使得规则的维护变得更加困难 [30]。而基于机器学习的方法从一定程度上可以解决这些问题。本研究中也采用基于机器学习的方法进行未来工作句挖掘研究。机器学习相较于以往传统方法在文本分类任务上往往都能表现出较好的性能。但是，使用单一模型进行分类难以全面地对文本进行特征提取，而且易忽略上下文语义关系，从而导致模型的分类效果欠佳 [31]。近几年越来越多的研究者开始探索模型的改进与融合，以提高分类效果 [31]。

2 研究内容

本研究通过对融合出版领域的学术论文未来工作句进行挖掘研究来分析该领域的未来研究趋势，探测融合出版领域的前沿主题，发现该领域新兴与热点话题。研究以知网文献库中的融合出版领域中文论文全文本为数据来源，利用人工标注得到未来工作句识别与分类语料库；之后在识别语料库上使用支持向量机、朴素贝叶斯和随机森林三种机器学习模型与 SelectKBest 特征选择方法结合训练性能最优的未来工作句自动识别模型；最后在分类语料库基础上针对未来工作句类别进行进一步占比与统计分析。本文的研究框架如图 1 所示。

2.1 语料标注与预处理

2.1.1 用于未来工作句识别的分类语料标注

由于本研究的研究对象是融合出版领域的论文，为了保证数据的准确性与领域特性，本研究采用知网中通过查询词为“融合出版”的篇名搜索得到的融合出版论文为研究对象，从人工筛选 (筛选过程中过滤篇名中“融合出版”没有作为整体出现的论文) 后得到的 447 篇文章中抽取研究数据构建未来工作句语料库。未

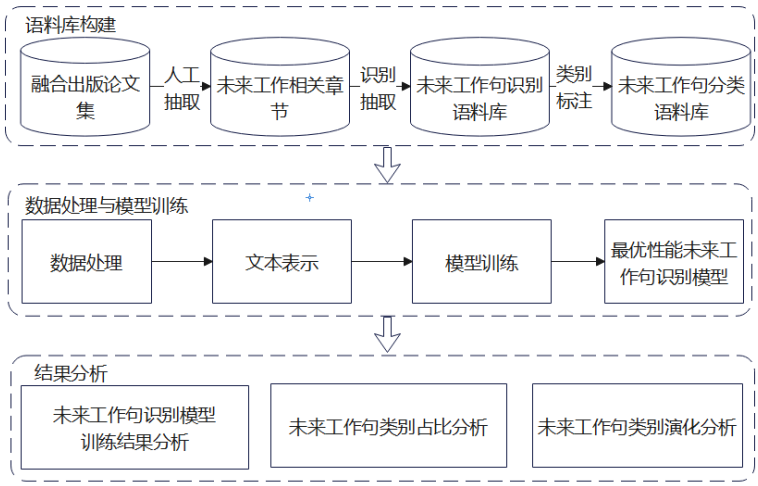


图 1 研究框架图

来工作句的标注主要分两个阶段，第一阶段抽取期刊论文的未来工作相关章节，第二阶段从相关章节中抽取未来工作句。

第一阶段是未来工作相关章节抽取。在抽取工作前，笔者观察过大量该领域相关论文，发现未来工作句出现的位置几乎都是文章末尾部分章节。这些章节主要分成两种情况，一种是作者将未来工作单独用一个章节来说明，章节名诸如“未来展望”“未来工作”等；另一类则是将未来工作放在文章总结性章节来论述，章节名诸如“小结”“总结”“结论”等。因此，本研究从论文的这些章节中抽取未来工作相关章节。之后笔者使用人工抽取的方式将每篇文章的篇名、发表年份以及未来工作相关章节抽出后用同一张工作表进行存储，方便后续的研究与分析。

第二阶段是未来工作句抽取，我们在之前构建的工作表的基础上进行后续的未来工作句抽取工作。通过对语料进行分析，总结出未来工作句总体特征，并且参考先前研究者在 NLP 领域的研究成果 [27]，总结出针对本研究语料库内未来工作句的判别标准。未来工作句的判别标准主要分成以下三种：

（1）对未来工作的直接提及，如“在未来工作中将进一步探寻…”“未来还需…”。例句：“目前科技期刊传统出版亟待数字化转型，而信息产业的数字出版还不成熟，足以看出拥有全面数字化特征的科技期刊融合出版模式将成为未来的发展方向 [32]。”

（2）包含引出未来工作的词语，如“后续我们将…”“下一步我们将…”。例句：“因此，如何培养兼具专业知识和新媒体操作技能的“两栖型”期刊编辑人才是后续研究的重要方向 [33]。”

（3）承上启下的连接句，如“本研究还需进一步完善和深入”“未来的研究工作将具体包含以下几个方面”。例句：“但是仍有不少可以开拓的空间 [34]。”

本研究中，未来工作句抽取与标注过程严格按照标注规范进行，此阶段主要是人工识别与标注，最终标注出未来工作句共 216 句。另外，我们也标注出非未来工作句共 388 句，最终形成标注语料库情况如表 1 所示。

doi:10.3772/j.issn.2095-915x.2023.05.010

表 1 标注语料库统计表

类型	未来工作句	非未来工作句	总计
数量	216	388	604

2.1.2 用于未来工作句类型分类的语料标注

为了更加深入了解融合出版领域的研究现状，从而更精准定位其未来研究方向，本研究对于未来工作句集依据事先构建好的分类体系人工判定类别，形成本研究的未来工作句类型分类语料库。本研究首先对语料进行阅读与分析，发现此语料与先前研究者在对 NLP 领域的未来工作句自动识别与分类研究 [27] 中采取的分类体系较为契合。该分类体系基于扎根理论构建，研究者将未来工作句分为方法、资源、评估、应用、问题和其他六大类别。本研究选取融合出版领域部分未来工作句进行预标注后发现，由于语料的领域特性，某些句子按照此分类体系归类不是很恰当。因此，我们在此分类基础上增设“管理手段”和“工具”两大类别，并在“管理手段”下设置了“人员”和“管理制度”两个子类别，具体分类标准如表 2 所示。

表 2 未来工作句分类表

类别	子类	含义	例句
方法	模型	对当前研究提出的模型、算法等进行扩展优化研究	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。
	体系	构建或完善相关机制或体系	另一方面要加强对人脸识别技术的监管与规制，构建起完备的行业规范体系。
	新方法	探索或引入更适宜的新方法	出版单位尤其大学出版社应秉承为教育服务的出版理念，在深耕内容的基础上不断探索更多的融合方式，为全民教育和文化传承传播作出更大的贡献。
	政策	国家或社会政策驱动与保障	建议新闻出版署在全国融合发展重点实验室工作的推进中，强化科技期刊融合发展相关工作。
	其他	不属于以上方法子类的其他方法类句子	
资源	优化资源	优化当前使用的数据、信息或语料	本文研究的数据是通过随机选出，如果能选择出更多的新闻评论类型，将有助于本方法的通用性的验证。
	扩展资源	扩大数据集或语料库	后续研究应扩大数据来源范围。
	更改资源类型	使用其他领域的语料资源	未来可以考虑利用其他领域数据进行研究。
	其他	不属于以上资源子类的其他方法类句子	
评估	评估当前工作	通过实验评估当前工作提出的模型或者算法的效果	本文研究还有不足，在后续的研究中，可以通过案例研究和实际调研，利用供应链绩效评价方法来进一步检验科技期刊融合出版模式的优越性。
	改善评估手段	优化或使用其他的评估指标或标准	因此，参与出版融合知识产权交易体系的各方仍然需要积极探索知识产权新标准，构建更加合理的知识产权评估、定价体系 …
	扩大评估范围	扩展到其他领域进行评估	未来，新媒体将助力期刊评价体系逐步从唯引文至上的学术评价体系逐步过渡到面向学术成果的全面影响力评价指标体系。
	其他	不属于以上评估子类的其他方法类句子	

类别	子类	含义	例句
应用	应用到其他任务	将当前研究提出的方法应用到其他任务	未来在应用服务、数据服务、知识服务的基础上，也可以对云化进行探索，提供更加高效的服务。
	增强研究内容普适性	增强当前研究内容在不同领域内的普适性	随着 AR 技术的发展和读者需求的多元化，在未来，AR 技术还将更深刻地介入出版的各个领域，为出版行业注入新的活力。
	其他		
问题	待解决的问题	目前研究者仍在努力解决的一些棘手的问题	但在树立本土文化自信的征程中，本土文化资源要完成融合出版的变革，实现融合出版的可持续发展，仍亟需解决意识欠缺、机制不完善、人才储备不足等问题。
	有意义的新问题	提出一个新的有意义的研究问题	此外，随着 AI 技术的逐步成熟，AI 创作将成为一种日益重要的内容生产渠道，如何对机器自动生产的内容进行管理和安全保护，也是很有价值的研究方向。
	其他		
管理手段	人员	人员管理与人才培养	在未来，编辑尤其是年轻编辑，应该积极探索融合出版形势下的编辑业务提升渠道，采用线上学习 + 线下实操、理论 + 实践、传统 + 创新的模式，不断突破自我，寻找编辑的真正生存之道。
	管理制度	对管理手段的改进或提出新的管理手段	一是融合出版的管理制度亟须完善。
工具	-	采用工具、软件或平台	当前，我国大多数科技期刊未实现全流程数字化改造，运作模式仍与纸刊时期大同小异，未来应充分利用已有数字平台和新媒体等手段真正实现出版全流程数字化。
其他	-	不属于上述七个类别的其他未来工作句包含在该类别中。	“融合出版”范式下出版规制、产业和组织将迎来一次革命性的变革，传统出版单位将在经历转企改制的阵痛后，通过“融合出版”这一新兴出版范式实现中国出版业的整体涅槃和浴火重生。

2.1.3 数据预处理

经过预处理的数据，可以获得更加准确的语言特征，从而更好地支持机器学习模型的训练，并且能更加快速地获得更为准确的训练结果。为此，我们必须先完成数据的清洗、分词、去停用词等工作，从而为机器学习的准确性奠定坚实基础。

（1）数据清洗

本文首先需要对文本进行一些去空去重操作，即处理一些无意义的空格、空行，其次，去除一些无用的标点符号，以方便后续进一步数据处理。

（2）分词

由于中文没有明确的句子分割标准，因此需要使用更复杂的分词模型进行分析。本研究使用 jieba[①]进行中文文本的分词。为了提高分词的准确性，本文通过去停用词来进行分词的优化。

2.2 模型建立与训练

2.2.1 文本表示

通过使用文本表示，我们可以把数据转换成计算机能处理的形态，比如向量或矩阵。本文采用的离散式中的 TF-IDF[35]，它是最为常用

① https://pypi.org/project/jieba/

doi:10.3772/j.issn.2095-915x.2023.05.010

的文本特征权重计算方法。TF-IDF 可以用以评估文档集合中的某个字词或是语料库中的某份文档中的某个字词的重要程度。当某个词在一篇文章中出现的频率 TF 高，并且极少出现在其他文档中，则我们可以判断此词具有出色的类别区分能力 [36]，其计算公式如下：

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

其中，$W_{i,j}$ 表示文档 j 中第 i 项词项的权重，N 则表示集合中的文档数量，此外，$tf_{i,j}$ 表示文档 j 中第 i 个词的词频，df_i 表示集合中出现当前词项 w_i 的文档数 [37]。

2.2.2 文本特征选择

特征选择可以减少冗余特征，保留具有较强区分能力的特征从而提高分类器的表现并且防止分类器过拟合 [38]。其方法可以归纳为三类：过滤式、包裹式和嵌入式 [39]。

特征选择方法种类很多，本研究中主要采用 SelectKBest。原因在于其他常用算法如特征递归消除算法（Recursive Feature Elimination，RFE）和随机森林（Random Forest, RF）等，分别存在稳定性和选择偏向的问题 [40]。SelectKBest 是一种基于统计学原理的过滤式特征选择方法 [41]，用于从 n 堆数据中寻求价值最优的 k 类数据 [42]。它可以根据给定的评价函数和得分，来选择和排名特征。在使用 SelectKBest 时，如果数据集中含有不止一个特征，可以采取评分函数进行特征筛选 [43]。在本研究中，采用卡方检验（Chi-Squared Test) 作为评分函数。此种特征选择方法在分类任务上已有成功应用 [44-45]。

2.2.3 文本分类模型训练

未来工作句自动识别任务本质上属于二分类问题。在本实验中，我们选择采用支持向量机（Support Vector Machine，SVM）、朴素贝叶斯（Naive Bayesian，NB）和随机森林训练未来工作句自动识别模型，通过对比模型性能选择最优模型。

支持向量机是一类广义线性分类器，它采用监督学习的方式对数据进行二元分类。SVM 首先利用以内积函数进行定义的非线性变换将输入空间变换到一个高维空间，之后在这个空间中来求解（广义）最优分类面 [46]。在本实验中，采用核函数为 linear 的 LinearSVC 模型，最大迭代次数 maxiter 设为 5000，惩罚参数 C 设为 1.0。

朴素贝叶斯是一种基于概率统计的机器学习算法，其原理在于通过类别的先验概率以及特征分布相对于类别的条件概率来计算未知文档属于某一类别的概率 [47]。本实验中选择的是伯努利朴素贝叶斯 (BernoulliNB)，拉普拉斯平滑系数 alpha 设为 0.0001。

随机森林是一种具有较高预测准确率的抽样方法，利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。本实验中基评估器数量 n_estimators 设为 200。

3 结果分析

本节中我们结合未来工作句自动识别模型的训练结果和未来工作句类别进行进一步分析。

3.1 未来工作句自动识别模型训练结果评估

在研究过程中，需要对模型进行评估。通过采用 K 折交叉验证（K-Floder Cross Validation）[49]，我们可以将大量的数据加入模型的训

练和预测，同时避免划分训练集和测试集时的随机性，从而大大减少模型的不准确性，并且更好地体现出交叉验证的概念。

本研究将数据集按 9：1 划分为训练集和测试集，进行十折交叉验证[50]，并将结果进行平均，来比较判别分类模型的优劣。

在本研究中，我们将正确率（Accuracy）、精确度 (Precision)、召回率 (Recall) 以及 F_1 值作为评估指标[51]，以期获得更准确的结果。

以上指标的详细定义如下：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

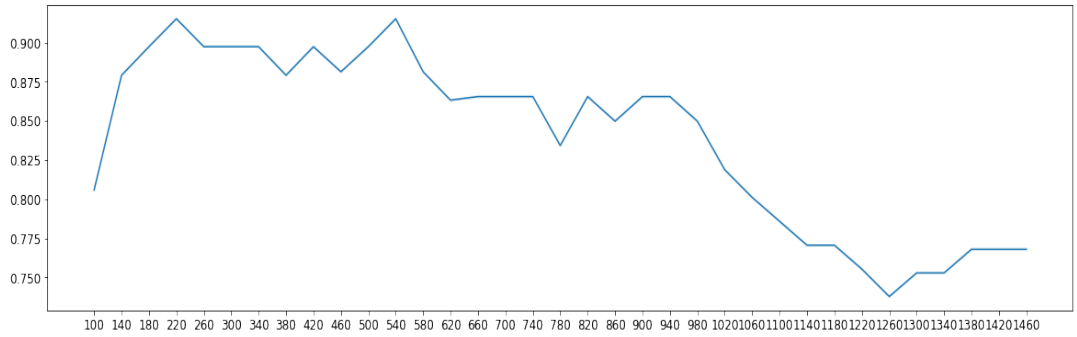
$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

由于精确率和召回率是一对矛盾的度量，模型评估中又往往需要同时考虑这两项指标，F_1 值即为人们设计的满足这一需要的性能度量指标[52]。故本实验中我们最终以 F_1 值来选定最优模型。

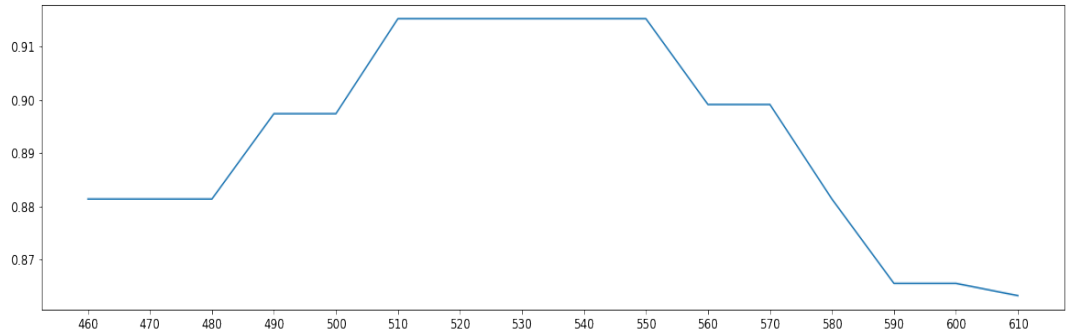
3.2 未来工作句自动识别实验结果分析

我们将 LinearSVC、BernoulliNB 和 RF 三种模型分别与 SelectKBest 特征选择方法进行组合，使用网格搜索的方法调整超参数 k。

以 LinearSVC 模型的超参数 k 的调整为例，我们首先将起点与终点分别设为 100 和 1500，步长设为 40，得到 k 在（460，620）内取值模型性能可能最优，如图 2（a）所示，之后我们在（460，620）区间上将步长设为 10 继续调参，得到 k 的最优取值区间为（510，550），如图 2（b）所示，本实验中我们取 530 作为 k 值。



（a）第一次 k 值调整验证曲线



（b）第二次 k 值调整验证曲线

图 2　LinearSVC 模型 k 值调整验证曲线

doi:10.3772/j.issn.2095-915x.2023.05.010

根据调整结果，最终得到 k 值为 530 时，LinearSVC 性能最优，加权平均 F_1 达到了 92.08%；k 值为 48 时，BeroulliNB 性能最优，加权平均 F_1 达到了 91.77%；k 值为 192 时，RF 性能最优，加权平均 F_1 达到了 88.91%。具体结果如表 3 所示。

对比支持向量机、朴素贝叶斯和随机森林三种模型的训练结果，我们得到 LinearSVC 模型在未来工作句自动识别任务中性能最佳，加权平均 F_1 值达到 92.08%。这表明此模型可以很有效地区分未来工作句和非未来工作句。今后我们可以使用此模型在更大规模语料库中进行未来工作句的自动识别，比人工抽取能节省更多时间与精力。

表 3　未来工作句自动识别模型训练结果评估

模型	类别	Precision	Recall	F_1
LinearSVC	非未来工作句	1.0000	0.8864	0.9398
	未来工作句	0.7727	1.0000	0.8718
	宏平均	0.8864	0.9432	0.9058
	加权平均	0.9367	0.9180	0.9208
BernoulliNB	非未来工作句	0.9231	0.9474	0.9351
	未来工作句	0.9091	0.8696	0.8889
	宏平均	0.9161	0.9085	0.9120
	加权平均	0.9178	0.9180	0.9177
RF	非未来工作句	0.9744	0.8636	0.9157
	未来工作句	0.7273	0.9412	0.8205
	宏平均	0.8508	0.9024	0.8681
	加权平均	0.9055	0.8852	0.8891

3.3　未来工作句类别分析

在未来工作句类别标注过程中，为了控制标注质量，我们采用双人隔离标注的方式，之后将结果进行对比，出现分歧时，通过小组讨论或者专家评议确定最终结果，确保标注的一致性；并且，标注结果再经由专家审核，从而保证标注质量。由于每个未来工作句都只能标注唯一的类别标签，若一个句子中含有多于一种类别的未来工作句，则需要拆分后再进行类别标注。我们在标注后得到的分类语料库基础上对未来工作句类别进行进一步探究，主要包括未来工作句类别占比分析和分布分析。

3.3.1　未来工作句类别占比分析

（1）一级类目类别占比分析

笔者针对未来工作句识别语料库中的未来工作句类别占比进行统计，结果如图 3 所示。

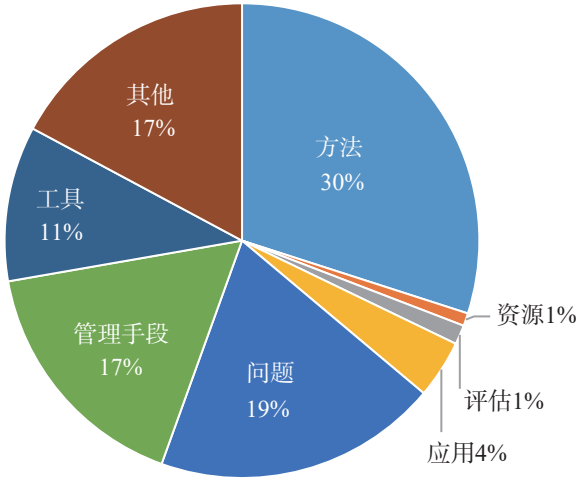


图 3　未来工作句类别分布图

从图 3 中可以看出，方法类未来工作句占比最大，这表明，融合出版领域的研究者对于方法的改进与创新较为重视，关注新方法的探索，体系的建立与维护等。其次，问题类、管理手段和工具类未来工作句也占有一定比例，说明研究者也较关注该领域尚未解决的难题，并关心人员管理以及管理手段上的革新。同时，希望利用软件、平台和工具来加速融合出版领域的发展。但是，我们不难发现，评估类和资源类未来工作句占很小，仅有 1%，这一方面表

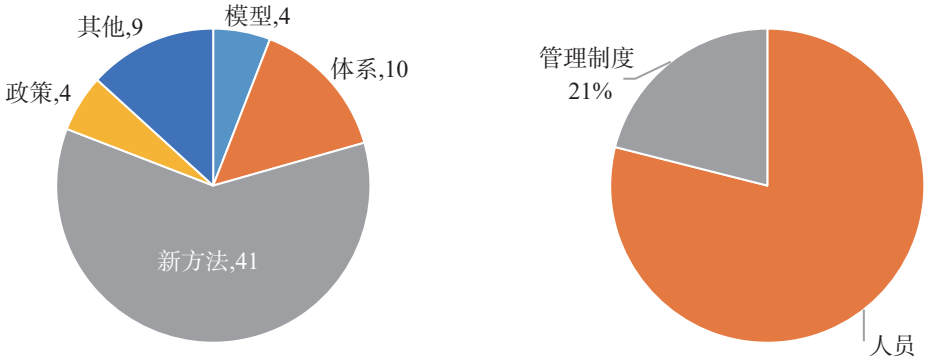
明研究者可能认为这两个方面研究意义不大，即便改进和创新对该领域的发展也无法起到较大的推动作用。但另一方面，正是融合出版领域发展至今这两个方向的未来研究有所欠缺，导致可能有些好的想法被忽略，而这往往可能正是突破点所在。这也为研究者的未来研究提供了一个很好的思路。

通过分析结果，我们可以预测未来融合出版领域的研究方向应该还是以方法、问题、管理手段和工具为主；而对于资源和评估类，未来可能需要在评估研究价值性后考虑是否要在这些方面投入更多精力，寻求突破。

（2）二级类目占比分析

为了更精准定位融合出版领域研究者的未来研究倾向，我们选取了占比最大的“方法”类别和我们比较感兴趣的“管理手段”类别进行子类别占比分析。由图 4（a）中可知，在方法类别的研究中，研究者多倾向于探索或引入更适宜的新方法，因为融合出版本身就要求出版业在传统方法上结合数字技术进行革新。此外，对相关体系建设与维护也是该领域一个较为重要的发展方向，研究者提议结合时代环境构建一个良好的体系环境，从而保障融合出版更好更快发展。这一方面需要出版业做出努力，另一方面也需要政府和社会提供政策支持与保障。

我们在管理手段下又分为人员和管理制度两个子类，从图 4（b）中我们可以看到，人员管理与人才培养非常有必要。不仅需要编辑人员提高素质，不断学习来自我提升，管理决策者也需要把握好行业形势，重视优秀人才的培养，更好助力该领域发展。另外，在管理制度上也不能固守成规，需要适时创新。



（a）方法类目分布图　　（b）管理手段类目分布图

图 4　未来工作句子类分布图

3.3.2　未来工作句类别分布分析

此外，我们分别统计了 2014—2023 年各年份的不同类别的未来工作句数量，进行分布分析，结果如图 5 所示。

从总体来看，方法类未来工作句的占比在近五年表现出明显优势，自 2019 年后，该类别未来工作句数量大幅增长。由于近几年各种新技术飞速发展，互联网、云计算与人工智能逐渐渗透到各个领域，传统出版业试图在传统方式上利用数字技术寻求新的突破。我们也可以预测到融合出版领域的未来研究应该还是以方法为主，在传统方法基础上加以改进与创新，相关体系建设与政策支持也是未来研究的重点。

doi:10.3772/j.issn.2095-915x.2023.05.010

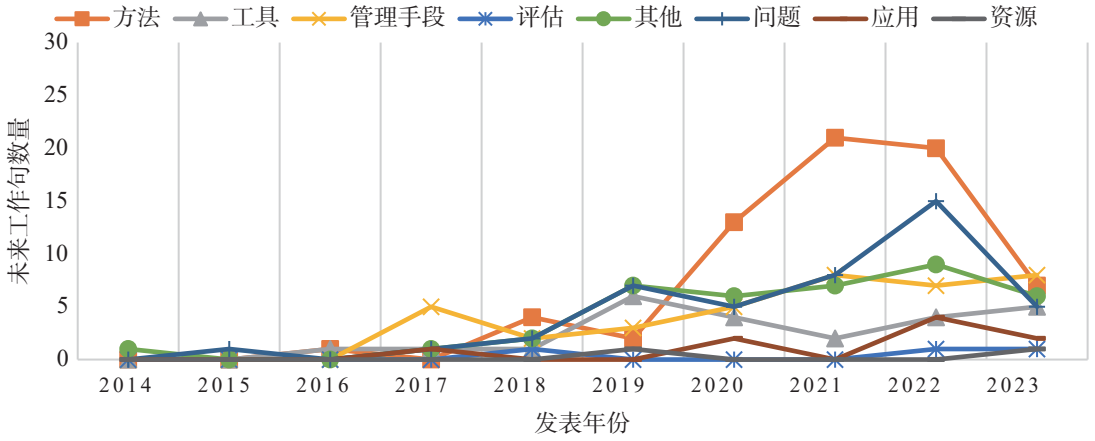


图 5　未来工作句类别分布图

另外，从图中我们也可以看到，近几年问题类的未来工具句数量也在不断增长。随着技术的飞快革新，研究者的创新想法愈来愈多。但数字技术的融入过程也会带来很多问题。例如平台建设方面资金投入不及时，版权保护与个人信息保护仍存在不少漏洞等。这些问题都有待未来研究者深入探索，寻求最佳解决方案。

而对于某些类别，譬如资源类和评估类，从 2014 年至今未来研究方面都有所欠缺，说明此领域对数据和评估的依赖可能不如其他领域显著；还有一种可能是对该领域的研究开始得较晚，这两个方面还未有大规模研究者对其进行深入挖掘。未来可能需要该领域研究者评估价值性后考虑是否要加大投入。

3.4　未来工作句内容演化分析

为了更加深入了解融合出版领域未来工作的具体内容，探析该领域的前沿主题及其变化规律，我们对 2019 至 2023 年的未来工作句具体内容进行深入分析。首先我们提取各年份未来工作句数据集计算逆文档频率并筛选后的 Top20 的词项进行分析，总结近五年融合出版前沿主题以及演变趋势，为该领域后续发展提供借鉴。

在基于逆文档频率的未来工作句内容分析时，我们首先抽取了 Top50 的词项，经过观察发现，其中有些词是“融合出版”和“未来工作”相关词，以及一些无实际意义动名词和程度副词。为了更好地分析前沿主题本身，我们在所有词项提取结果中均删除了这些词，并在筛选后的结果中取 Top20 的词项进行分析，词项内容如表 4 所示。

从各年份的 Top20 词项内容中，我们可以很清晰探察到各年份融合出版领域的前沿主题的差异及变化趋势。

2019 年，融合出版领域重点关注的前三位分别是“媒体”“科技期刊”与“传播”。结合之前的背景研究，2019 年融合出版领域研究论文数量大幅增长，这一年发表的论文中，研究者大多还是从“媒介融合”视域出发去研究融合出版领域问题，并且重视领域相关概念、方法等的传播。其次是对期刊和编辑等的要求，鼓励期刊突破传统出版方式，发挥“领头羊”作用，积极利用数字出版优势；鼓励编辑努力提升素养，转变观念，致力于融合出版领域的发展。另外，这一时期研究者也开始重视技术

表 4　未来工作句分类表

年份	Top20 词项
2019	“媒体” “科技期刊” “传播” “行业” “模式” “期刊” “时代” “编辑” “利用” “技术” “互联网” “信息” “创新” “图书” “如何” “更新” “形式” “文化” “数字化” “阅读”
2020	“技术” “内容” “创新” “媒体” “出版社” “转型” “渠道” “管理” “营销” “问题” “读者” “模式” “平台” “学术期刊” “保护” “体系” “传播” “优质” “企业” “人才培养”
2021	“编辑” “技术” “问题” “媒体” “内容” “图书” “需求” “平台” “期刊” “创新” “出版社” “传播” “渠道” “机遇” “用户” “能力” “方式” “出版业” “产品” “服务”
2022	“技术” “媒体” “内容” “图书” “编辑” “模式” “需求” “纸质” “转型” “服务” “实践” “创新” “文化” “科技期刊” “科普” “形式” “少儿” “少儿 科普” “多元” “知识 服务”
2023	“技术” “内容” “创新” “教育” “行业” “传播” “媒体” “问题” “优质” “方法” “高质量” “平台” “实践” “编辑” “体系” “图书” “需求” “优质 内容” “高质量 发展” “素养”

的利用，主要是基于互联网信息和技术与传统出版的融合来实现创新与突破。

2020 年，融合出版领域重点关注的前三位分别是“技术”“内容”与“创新”。这一时期该领域在建立好领域根基后，开始寻求技术层面的突破。另外，我们看到“管理”“问题”与“平台”的排名也较靠前，研究者也开始意识到管理与工具的重要性，融合出版不仅仅依靠出版社与编辑提升与转型，相关企业也需要在营销管理、体系构建与人才培养上倾注更多精力，另外，融合出版发展遇到的一些问题也不容忽视，需要多方共同深入探究并寻求解决方案。

2021 年，融合出版领域重点关注的前三位分别是“编辑”“技术”与“问题”。这一时期，融合出版领域对“人”提出了更高的要求，更加注重人才培养与管理，“技术”与“问题”仍旧是该领域关注的焦点。另外，我们观察到，该时期的Top20词项中出现了“需求”“用户”“产品”“服务”这类词，该领域开始考虑到用户需求与服务，这说明融合出版领域开始重视应用层面的问题。

2022 年，融合出版领域重点关注的前三位分别是“技术”“媒体”与“内容”。在考虑到应用层面后，研究者发现要着重解决的还是技术上的问题，并且主要是出版内容的融合，这才是提升应用性能的最佳途径。这一时期融合出版的前沿问题主要是如何改进与扩展技术来解决应用层面的问题。另外，我们发现“少儿”与“少儿科普”排名也较高，这也是融合出版领域服务视野拓展的表征。由于少儿是图书阅读的一类较大数量群体，所以也是传统出版与数字出版融合浪潮中受影响较大的一类群体，要更加重视这类群体的服务层面的问题。

到了 2023 年，融合出版领域重点关注的前三位与前一年相差无几，分别是“技术”“内容”与“创新”。说明融合出版领域聚焦点没有发生太大变化。值得注意的是，这一时期，“教育”一经出现便排到了第四的位置，说明这一时期融合出版领域从开始的“少儿科普”考虑到更为全面深入的问题，例如融合出版以何种方式融入教育中，如何在青少年培养中发挥最大作用等。另外，“高质量”“高质量内容”和“高质量发展”的出现让我们看到该领域发展到这一时期，基础层面的构建工作已基本完善，开始追求高质量高水平发展，这也是一个领域发

doi:10.3772/j.issn.2095-915x.2023.05.010

展趋于成熟的象征。

结合融合出版领域近五年的未来工作句中Top20 词项，我们可以看到该领域从起步走向成熟的演化过程，这五年中该领域发展侧重点也经历了一个较为鲜明的变化。另外，每年都有新的前沿主题的出现，这些前沿主题从开始受到关注到问题的发现与方法的探究再到应用层面的完善恰恰也是该领域一步步发展与壮大的具象表示。

4 结语

本文为探测融合出版领域前沿主题与未来有意义研究方向，通过采集知网文献库中的中文文献数据，结合机器学习进行未来工作句挖掘研究。研究主要利用融合出版领域论文为数据构建了未来工作句识别与分类语料库，在此识别语料库基础上使用支持向量机、朴素贝叶斯和随机森林三种机器学习模型与 SelectKBest 特征选择方法进行未来工作句自动识别模型训练，对比模型训练结果选择出性能最优的 LinearSVC 模型。该模型的 F_1 值达到了 92.08%，表明该模型可以很好地从文本中识别出未来工作句。另外，我们还基于未来工作句分类语料库对未来工作句的八大类别进行更进一步的研究，主要包括占比分析与分布分析。结果表明，融合出版领域的研究者更倾向对方法和问题类未来工作的研究，他们更关注该领域方法的改进与创新，并关心尚未解决的难题及有意义的新问题，对于资源和评估方面则关注较少。由此我们预测该领域未来的研究方向也是偏向方法与问题方面，而资源和评估方面可能需要研究者衡量研究价值后决定是否需要对其展开更加深入的研究。最后，本研究还通过对融合出版领域未来工作句本身内容研究来分析并总结该领域近五年的前沿主题的演化趋势，从而更好探析该领域过往前沿问题研究重点及其变化，并为未来研究提供借鉴。

本文也存在一定的不足，例如语料库规模较小且未采集英文论文数据，训练的模型还有待在更大规模语料库上测试抽取效果，另外，还未使用深度学习模型对比模型效果。后续我们将进一步扩大语料库规模，拟加入英文论文数据，且拟采集知网文献库之外的数据，例如微信公众号文章和新闻报道等；之后也将在更大规模语料库上测试模型的自动识别效果，也可进一步测试模型在其他领域语料上的适应性；此外，未来将使用深度学习模型与传统机器学习模型作对比，从而训练性能更优的未来工作句自动识别模型，提高识别准确率。

参 考 文 献

[1] 陈美华 . 全媒体视域下的区域出版产业竞争力评价与提升研究 [D]. 南昌 : 南昌大学 , 2018.
[2] 陈伟军 . 媒介融合视野中的新闻出版强国建设 [J]. 中国出版，2010, 33(21): 39-42.
[3] 刘荣 . 媒体融合背景下学术图书出版的困境与出路 [J]. 传播与版权 , 2023, 11(17): 8-11.
[4] 范亚芳 , 渠芳 . 对我国高校机构知识库建设的几点思考 [J]. 情报杂志 , 2007, 26(9): 132-134.
[5] 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, 33(9): 15-18.
[6] 刘颖悟 , 汪丽 . 媒介融合的概念界定与内涵解析 [J]. 传媒 , 2012, 14(1): 73-75.
[7] 李弘 . 基于矛盾论视角的出版融合发展要义辨析 [J]. 出版与印刷 , 2023, 34(3): 1-10.
[8] 蔡雯 , 王学文 . 角度·视野·轨迹——试析有关“媒介融合”的研究 [J]. 国际新闻界 , 2009, 49(11): 87-91.

[9] 王军 . 论融合出版的发生及其内涵 [J]. 新媒体研究 , 2019, 5(23): 36-38.
[10] 柳斌杰 . 深入学习贯彻党的十七届五中全会精神 , 为实现新闻出版业“十二五”时期发展目标努力奋斗 [J]. 中国编辑 , 2011, 9(2): 4-8.
[11] 于殿利 . 从融合出版到出版融合——数字传媒时代的出版新边界探析 [J]. 出版发行研究 , 2022, 38(4): 5-15.
[12] 李弘 . 面向知识服务的出版融合发展浅析 [J]. 科技与出版 , 2016, 35(12): 12-16.
[13] 杨石华 . 马克思主义出版观视域下融合出版人才培养的理念革新与路径优化 [J]. 出版与印刷 , 2023, 34(2): 84-90.
[14] 李亚欢 , 李宁 . 融合出版背景下青年医学编辑人才培养与个人成长探析 [J]. 科技传播 , 2023, 15(7): 34-36, 40.
[15] 张米 . 加强数字出版编辑人才建设推动地方出版产业融合发展 [J]. 采写编 , 2023, 33(3): 151-153.
[16] 方卿 . 守正创新：学科交叉融合背景下的出版人才培养 [J]. 科技与出版 , 2023, 42(1): 6-11.
[17] 丁子涵 . 融合发展环境下传统出版社编辑素养提升策略 [J]. 中国编辑 , 2021, 19(8): 93-96.
[18] 闫翔 . 融合出版时代编辑应该具备的素养分析 [J]. 中国编辑 , 2018, 16(7): 48-50.
[19] 侯培东 . 融合出版时代编辑的坚守与转变 [J]. 编辑学刊 , 2017, 34(4): 16-20.
[20] 冯颖 . 科技图书编辑在融合出版时代的历史使命与担当 [J]. 中国传媒科技 , 2023, 31(2): 108-112.
[21] 李士振 , 周昇亮 . 融合出版背景下加强版权保护的对策与措施 [J]. 出版参考 , 2023, 26(1): 60-63.
[22] 周海晨 , 郑德俊 , 郦天宇 . 学术全文本的学术创新贡献识别探索 [J]. 情报学报 , 2020, 39(8): 845-851.
[23] HU Y, WAN X. Mining and analyzing the future works in scientific articles[J]. arXiv preprint arXiv: 1507. 02140, 2015.
[24] LI K, YAN E. Using a keyword extraction pipeline to understand concepts in future work sections of research papers[C]//Proceedings of 17th International Conference on Scientometrics &Informetrics. Kisbon, Portugal, 2019: 1-11.
[25] ZHU Z, WANG D, SHEN S. Recognizing sentences concerning future research from the full text of JASIST[C]//Proceedings of the Association for Information Science and Technology, 2019, 858-859.
[26] 徐彤阳 , 尹凯 . 基于深度学习的数字图书馆文本分类研究 [J]. 情报科学 , 2019, 37(10): 13-19.
[27] ZHANG C, XIANG Y, HAO W, et al. Automatic recognition and classification of future work sentences from academic articles in a specific domain[J]. Journal of Informetrics, 2023, 17(1): 101373.
[28] HAO W, LI Z, QIAN Y, et al. The acl fws-rc: A dataset for recognition and classification of sentence about future works[C]//Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. 2020: 261-269.
[29] QIAN Y, LI Z, HAO W, et al. Using future work sentences to explore research trends of different tasks in a special domain[C]//Proceedings of the Association for Information Science and Technology, 2021: 532-536.
[30] 李渝勤 , 孙丽华 . 基于规则的自动分类在文本分类中的应用 [J]. 中文信息学报 , 2004, 19(4): 9-14.
[31] 代雨聪 . 基于情感分析的个性化推荐研究 [D]. 上海 : 东华大学 , 2022
[32] 刘爽 , 吕柔 , 武爱 . 基于供应链管理的科技期刊融合出版模式研究 [J]. 数字图书馆论坛 , 2018, 14(8): 45-52.
[33] 高存玲 , 赵星耀 . 海洋科学类期刊融合出版现状、问题与对策研究 [J]. 中国科技期刊研究 , 2019, 30(12): 1316-1323.
[34] 习妍 , 孔丽华 , 姜璐璐 . 科技期刊融合出版中网络平台效能的发挥——以《中国科学数据 (中英文网络版)》为例 [J]. 编辑学报 , 2019, 31(S2): 169-173.
[35] SALTON G, BUCKLEY C. Term weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513–523.
[36] 施聪莺 , 徐朝军 , 杨晓江 . TFIDF 算法研究综述 [J]. 计算机应用 , 2009, 29(S1): 167-170, 180.
[37] 王明文 , 付翠琴 , 徐凡 , 等 . 基于词项共现关系图模型的中文观点句识别研究 [J]. 中文信息学报 , 2015, 29(6): 185-192.
[38] 向鸿鑫 , 杨云 . 不平衡数据挖掘方法综述 [J]. 计算

doi:10.3772/j.issn.2095-915x.2023.05.010

机工程与应用 , 2019, 55(4): 1-16.
[39] 周成 , 魏红芹 . 专利价值评估与分类研究——基于自组织映射支持向量机 [J]. 数据分析与知识发现 , 2019, 3(5): 117-124.
[40] 刘逸竹 , 李晴 , 吴文斌 . 遥感提取灌溉耕地的特征优选——以中国北方为例 [J]. 中国农业资源与区划 , 2021, 42(9): 27-35.
[41] 文大鹏 . 基于机器学习与 LIBS 技术的中药材定性与定量分析研究 [D]. 兰州 : 西北师范大学 , 2021.
[42] 朱城 , 苏前敏 , 郭晶磊 , 等 . 基于改进随机森林优化算法在医疗数据中的应用研究 [J]. 智能计算机与应用 , 2021, 11(8): 130-134.
[43] 宋建 , 王文龙 , 李东 , 等 . 基于 Stacking 集成学习的注塑件尺寸预测方法 [J]. 华南理工大学学报 (自然科学版), 2022, 50(6): 19-26.
[44] 马栋林 , 张澍寰 , 赵宏 . 改进 Relief-C5. 0 的恶意域名检测算法 [J]. 计算机工程与应用 , 2022, 58(11): 100-106.
[45] 赵潇雅 , 郜志英 , 周晓敏 , 等 . 基于 FDA-LSTM 的冷轧过程多源异构时序数据处理及颤振预测 [J]. 振动与冲击 , 2022, 41(22): 202-210.
[46] 张学工 . 关于统计学习理论与支持向量机 [J]. 自动化学报 , 2000, 38(1): 36-46.
[47] 徐军 , 丁宇新 , 王晓龙 . 使用机器学习方法进行新闻的情感自动分类 [J]. 中文信息学报 , 2007, 22(6): 95-100.
[48] 方匡南 , 吴见彬 , 朱建平 , 等 . 随机森林方法研究综述 [J]. 统计与信息论坛 , 2011, 26(3): 32-38.
[49] 井晓丹 , 朱永忠 , 井睿 . 一类改进的贝叶斯模型选择方法 [J]. 统计与决策 , 2019, 35(1): 29-33.
[50] 袁睿豪 , 廖玮杰 , 唐斌 , 等 . 数据驱动的航空发动机材料设计研究进展 [J]. 航空制造技术 , 2021, 64(18): 22-30.
[51] 柳靓 . 社交网络中信息传播预测研究 [D]. 重庆 : 重庆邮电大学 , 2018.
[52] 刘怀亮 , 张治国 , 马志辉 , 等 . 基于 SVM 与 KNN 的中文文本分类比较实证研究 [J]. 情报理论与实践 , 2008, 31(6): 941-944.