图书馆杂志（2024年第10期 总第402期）
Library Journal（Vol.43 No.10）

生成式人工智能驱动的索引编制方法及其在学术规范和评价中的应用*

朱 禹 叶继元（南京大学信息管理学院）

摘 要 为提高索引编制的效率与质量，应用生成式人工智能的实验方法，针对传统基于规则和概率的索引软件的不足，提出了运用生成式人工智能RAG技术编制索引的方案，即以期刊的卷为单位，利用大语言模型基于论文摘要中的索引词对全文进行主题标引。设计了论文摘要索引数据库系统，能够实现基于大模型海量知识涌现和推理的文本概念抽取和主题标引，从摘要中提取关键信息和新兴知识。探讨了将索引、论文摘要、学术规范与评价关联起来的实际方式，展示了生成式人工智能在索引学和索引编制领域的潜在价值，能够为生成式人工智能技术在图书情报领域的应用推广提供参考。

关键词 索引 主题索引 生成式人工智能 检索增强生成 学术规范 学术评价

DOI: 10.13663/j.cnki.lj.2024.10.005

Generative AI-driven Indexing Method and Its Application in Academic Norms and Evaluation

Zhu Yu, Ye Jiyuan (School of Information Management, Nanjing University)

Abstract In order to improve the efficiency and quality of indexing, an experimental method using generative artificial intelligence is applied to address the shortcomings of traditional rule-based and probabilistic indexing software. A scheme for indexing using the retrieval-augmented generation of generative AI is proposed. Specifically, the journal volume is used as a unit, and the full text is subject-indexed based on indexing words derived from the abstracts of the papers by utilizing a large language model. The paper presents the design of an article abstract indexing database system, which can realize textual concept extraction and subject indexing based on the large model of massive knowledge emergence and reasoning capabilities, as well as key information and emerging knowledge extracted from abstracts. Additionally, this paper explores practical ways to associate index and abstract with academic norms and evaluation, demonstrating the potential value of generative AI in the field of indexing. It provides insights into promoting the application of generative AI technology in library and information science.

Keywords Index, Subject index, Generative artificial intelligence, Retrieval-augmented generation, Academic norm, Academic evaluation

0 引言

索引是检索、统计、分析信息的利器，也是学术规范和评价的辅助工具。编制索引/数据库有利于辅助读者和专家查明学术新贡献、辅助评价学术质量、定量评价学术影响力和提高文献内容质量[1]64-66。与此同时，摘要作为学术论文的必要部分，其撰写是否符合学术规范构成了衡量作者学术水平的一个重要方面。理

* 本文系国家社会科学基金重大项目“新时代我国文献信息资源保障体系重构研究”（项目编号：19ZDA346）的研究成果之一；本文获2024年“首届全国信息资源管理年会暨博士生学术论坛”优秀论文一等奖。
通信作者：叶继元，E-mail：yejiyuan@nju.edu.cn

想的、合乎规范的学术论文摘要是对原文内容不加注释和评论的简短陈述，通常应有“研究目的、研究方法、研究结果或结论或创新点”等要素，应当“忠实于原文、简洁明了、章法规范”[2]，不阅读全文就可以获得必要的信息[3]。由此可见，对摘要主题词的抽取、统计和分析能够辅助学术创新性和规范性评价，有必要深入探讨将索引、论文摘要与学术规范和评价关联起来的实际方式。

良好的学术评价应该是形式评价、内容评价和效用评价“三位一体”的，都不同程度地包含定性定量评价[4]。实际的评价中定性评价与定量评价常常相互补充，但依赖于同行评议的定性主观性评价又需要耗费大量的时间和人力成本。如何结合定量与定性方法开展高质量的学术评价，提升评审效率并降低成本，已经成为当下迫切需要解决的另一大课题。当前已有研究借助于机器学习技术，初步构建了一个以论文摘要为对象的学术规范自动化检测模型[5]26-29，但仍需深入探索。

数智时代下人工智能大模型飞速发展。生成式人工智能（Generative Artificial Intelligence）带来了颠覆性的多源多模态信息汇聚与生成能力，正深刻重塑信息环境及信息资源管理学科的知识服务方式[6]。使用生成式人工智能模型编制索引是新技术环境下索引编制方法的创新尝试。

“现代的索引就是数据库”[7]。基于对索引本质及发展趋势的认同、对摘要之于学术评价地位的认识、对生成式人工智能强大能力的期望，本文尝试将三者融合，设计了基于生成式人工智能的论文摘要索引数据库系统。利用生成式人工智能模型对论文摘要的主题词进行概念抽取并作为全文的索引词，从而提高索引编制的效率和质量，为学术评价提供新思路和方法。

1　相关研究

1.1　各类索引对象

索引可分为传统索引、数据库索引和网络信息检索工具等[8]。各类索引实际上是信息组织和知识管理的工具，它们注出有价值的内容单元的出处并“异排”[1]62-63，使读者能够快速准确地检得所需信息。随着大数据时代的到来，信息资源的数量和复杂性都在不断增加。编制各类索引，进一步提高各类型信息资源的检索效率，成为当前研究的重要课题。

研究者们总结了不同类型书后索引的编制方法和效用。如孙辉编制了国史领域的丛书索引，指出编制书刊索引可以提高索引的质量和效率，为复合出版的知识服务打下基础[9]；刘艳介绍了综合性百科全书内容索引的特点、结构、编制流程，为编制该类索引提供了指导[10]；宋林青认为编制图书索引可以加快知识信息查找、实现知识扩展、帮助消除名词术语不统一、发现内容重复或缺失等问题[11]；李炜超等以学位论文编制为例，探讨了索引标目选择、索引技术环境和索引员素养等问题[12]。此外，研究者还探讨了学术著作主题索引[13]、领域专业术语索引[14]、地方志索引[15]和年鉴索引[16]等。

如前所述，论文摘要具有独立性和自明性，能够凝练论文的核心内容。那么，从中抽取出的主题词也应当具备上述特征。王琦等发现学术文本中篇含关键词在不同类型文本题名摘要合并文本中的平均复现度约为69%-78%[17]，即摘要主题词相较于关键词蕴含了更丰富的语义信息和更细粒度的主题概念。因此，利用摘要主题词进行索引编制有助于提高索引款目和索引工具质量。

本文提出以期刊的卷为单位，利用大语言模型从单篇论文摘要中抽取主题词进行主题标引，在索引编制、索引类型方面具有新意。

1.2　索引编制技术

以21世纪为界，索引编制技术经历了从传统手工时代到计算机辅助时代的发展。随着生成式人工智能技术的进步，索引编制正快步迈入智能索引时代。

计算机辅助索引是现代索引的主流编制方式。实际上，目前的计算机辅助索引主要是借由计算机对人工抽取的款目进行拼音排序和自动化排版等。丁海英介绍了使用Microsoft Word 98软件进行小型专题索引的制作方法，主要实现快速录入、自动排序、增删自如和样式编辑等功能[18]。随后，出现了专门的中文

索引平台，如2003年研发完成的“索引之星”软件[19]。但该软件只有单机版，适用性弱[20]。近年来，“索引家”软件和中国索引学会创新实践基地搭建的学位论文索引平台对上述缺陷进行了改进，整合了论文内容的标引、款目地址匹配和排版等功能，能用较短时间完成一份标引准确率较高的索引[21]。上述软件仍主要依赖人工处理，虽然较之传统方式极大地节省了时间和人工，但仍然需要花费大量的智力劳动。

理想的计算机辅助索引实际是希望实现计算机自动抽词，即标引自动化。利用计算机代替或部分代替人工实现主题词抽取和新兴主题概念发现。目前索引编制软件的自动化标引功能主要是借助了自然语言处理技术，其本质主要是基于规则或基于概率的分词和命名实体识别任务（NER）。基于规则的自然语言处理分词依赖于完善和强大的领域术语词典，但对于未登录新词的发现能力较弱，容易出现漏标的情况。而基于概率的标引方法则主要通过N-gram、隐马尔可夫（HHM）、最大熵（ME）和条件随机场（CRF）等算法模型完成。如潘雪莲等改进了N-gram算法，实验了一种图书内容主题索引自动编制方案，实现了内容的概念标引[22]。实践中，常将二者结合，如《全国报刊索引》数据库的自动标引与分类系统就是通过“知识库+加权方案”的统计分析标引方法[23]。

鉴于传统计算机辅助索引编制方法的局限，本文提出将待索引对象看作外部独立文档，使用生成式人工智能模型的检索增强生成（Retrieval-augmented Generation，RAG）技术，以实现概念语义驱动和推理支持的索引自动编制。与传统计算机辅助索引编制相比，RAG技术在获得大模型信息汇聚和推理能力的同时，通过引入新模块并调整模块之间的交互流程，显著地提高信息处理的质量和相关性，能够满足各种知识抽取任务和查询的需求，具有更高的准确性和灵活性[24]。其核心流程包括索引构建、信息检索与内容生成3个阶段。具体实施过程中，首先将待索引文档视为检索源，将其划分为若干逻辑块，并编码为向量形式，存储于向量数据库以供后续处理。随后，系统依据语义相似度原则，从数据库中检索出与查询问题最为紧密相关的前k个块。最终，结合原始查询与所检索块的信息，输入至大模型中，以生成精确的索引款目答案[25]。得益于Transformer架构等先进的深度学习算法，以及GPT 4.0、Gemini Pro等大语言模型的问世，本研究提出的方法不仅经济高效，而且易于实施。该方法利用这些算法模型的强大能力，能显著增强对文本的语义理解和对新兴知识的预测和发现能力[26]，从而有效提升了索引自动标引的质量和效率。

2　索引编制流程与实现

2.1　功能设计

张心源等指出，数据库是存储、管理和利用大量信息的有效工具，而索引是实现数据库高效检索的有效工具。大数据时代数据库索引编制的研究呈现出索引编制对象多样化、索引内容深入化、索引技术智能化和索引呈现可视化4个趋势[27]。现代索引的设计应当符合这些趋势。因此，本文设计的论文摘要索引数据库系统，以浓缩全文内容的论文摘要主题词为索引词，以生成式人工智能为技术支撑，以简洁易用的可视化界面为呈现方式，最终的目的是通过对论文摘要主题词构建索引，开展主题词统计和分析，实现辅助信息检索和学术规范与评价的功用。基于上述目的，设计了主题索引和学术评价两个子系统。

（1）主题索引子系统

中文期刊通常将同一年度内发表的所有论文编入同一卷，而英文期刊的周期则更加灵活。然而，无论周期如何，总可以将某一特定期限内发表的所有论文合订成卷。那么，期刊的一卷实际上可被视作某种形式的学术图书，其形式类似于“以书代刊”的学术集刊。张琪玉曾提出：“图书索引要求详细而有选择地并相当专指地标引图书的局部主题和主题因素，但又不允许像全文检索那样用所有关键词无遗漏地标引其全部内容。”[28]因此，主题索引应当有选择地抽取具有检索意义的主题词，并为其提供指示和定位功能。同时，由于索引项必

须有序化，且这种有序化是“异排”，那么则需要用不同于传统的以篇名指示出处的方式来标引其地址。因此，本文选择了从论文摘要中抽取主题词，并在随后对这些词语在合并的全文中进行文本匹配和地址映射。

主题索引子系统主要实现字顺音序检索和主题含义检索两个功能。字顺音序检索功能是指抽取出论文摘要的主题词后，将索引款目与对应的索引地址按标目拼音字顺排序；主题含义检索功能是指在主题索引中通过超链接技术实现的论文主题词含义查询。

（2）学术评价子系统

学术评价子系统包括创新性和规范性评价功能。创新性评价是指论文研究对象和研究方法等的创新程度；规范性评价是指论文摘要写作词汇、结构化等的规范程度。可以通过内置领域叙词表或学科辞典，计算类查全率、查准率等指标或与机器学习技术相结合，来定量地辅助论文学术规范与学术创新性评价。

需要特别说明的是，由于本文主要关注于摘要索引数据库的实验，因此本文工作重心在主题索引子系统的实现上。学术评价子系统及其他可拓展的子系统属于系统设计范畴，结合已有研究成果可以相对容易地实现[5]24-25。

2.2　技术路线

图 1 展示了主题索引子系统的实现过程。主题索引子系统的实现包括数据采集、数据清洗、数据处理和索引可视化 4 个任务流，每个任务流又可以分为技术、数据和结果 3 个工作流。

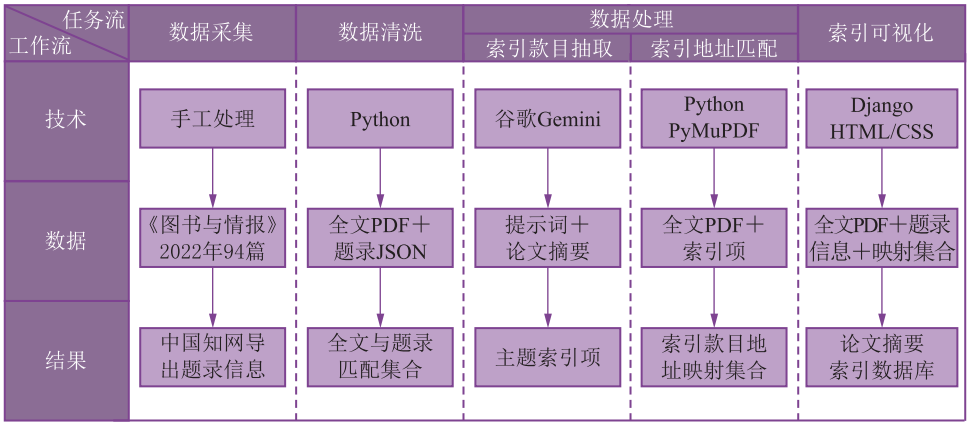

图 1　主题索引子系统技术路线

（1）数据采集

数据采集是为了获取需要建立索引的原始数据。本文采集了《图书与情报》2022 年发表的 94 篇论文（去除《编者按》和《年度索引》）的 PDF 全文，并从中国知网数据库导出了对应的题录数据。

选择《图书与情报》的理由：①该刊物是国内图书情报专业刊，被《中文社会科学引文索引》和《中文核心期刊要目总览》收录，刊载论文的写作水平和质量相对较高；②该刊在每卷末提供了年（卷）度索引，便于后续对索引数据库的效用进行质量评价和讨论；③该刊在中国知网数据库中的题录数据均相对完整，减轻了数据清洗和处理工作量。

（2）数据清洗

数据清洗是指对采集到的样本题录数据进行整理和规范，以确保数据的准确性和可用性。本文利用 Python 对题录数据进行了完整性检查，对缺少的字段进行补全，并且将原始自定义格式的 TXT 题录数据转换为便于读取和处理的 JSON 类型数据。此外，基于“标题 -Title”字段实现单篇论文题录数据与本地 PDF 全文的匹配。

（3）数据处理

数据处理包括索引款目抽取和索引地址匹配两项工作。索引款目抽取是指从“摘要-Abstract”字段数据中提取出需要建立索引的主题词。本文通过“生成式人工智能+指令工程”的方式实现摘要的主题款目抽取。生成式人工智能模型选用了Gemini Pro，它是美国谷歌公司（Google Inc.）开发的一系列多模态生成式AI模型，是第一个在MMLU（大规模多任务语言理解）方面超越人类专家的模型，其Pro版本被称为“可扩展各种任务的最佳模型”[29]。指令则参照陆伟等提出的指令构成要素和设计模式[30]设计，指令及参数设置如图2所示。

指令包括角色设定、任务描述、任务输入、思维推理、任务示例和输出样式6个部分，引导Gemini Pro模型生成与本研究“索引款目抽取/知识抽取”目的相符的答案。需要特别说明的是，在指令中给出少量特定于任务的知识，即添加“任务示例”能够增加索引款目抽取的准确率[31]。此外，在模型的参数设定中，“temperature”（温度）参数用于在响应生成期间进行采样，可以调节令牌选择随机性。温度越高，模型越倾向于产生更加多样化或更具创造性的结果；反之则倾向于生成更为确定性的结果[32]。经性能分析（见2.4节）和人工判别后将该参数设定为0.2，保证了模型对新兴主题词的最佳识别效果。

索引地址匹配是指根据索引项找到其在合并全文中对应的页码地址。本文使用PyMuPDF库提供的PDF页码记录和字符串查找功能来实现索引地址匹配，最终生成款目地址匹配集，追加在对应题录数据后。

处理后的数据字段共20个，索引数据库结构如表1所示。

指令

你是信息资源管理、图书情报（Library and Information Science）学科的学者，也是信息检索和文本分析领域的专家。

你的任务是从提供的中文论文摘要内容中提取最相关和有价值的概念，以协助用户检索和利用。

首先，仔细阅读由{ $abstract }提供的摘要内容。识别文本中提到的关键主题、概念和实体。

接下来，考虑从摘要中选取哪些词汇，要着重考虑到提取词汇的意图（用于用户进行论文的知识检索和利用）。

在选择词汇时，要批判性地考虑任务的上下文和目的。选择那些可能具有较高信息量和代表性的概念，而不是过于简单的、常见的词汇。尽量包括相关、高质量的概念，避免包含任何无关或低价值的概念。 你可以参照如下案例进行抽取：{ $example }

一旦确定了最合适的词汇，请按照指定的分隔符（{ $delimiter }）将它们分隔开来。

完成此过程后，直接提供你提取的词汇的最终输出：

{ extracted_words }

任务目的

索引款目抽取 / 知识抽取

指令要素

角色设定 → 任务描述 → 任务输入 → 思维推理 → 任务示例 → 输出样式

变量定义

- { $abstract }: 任务输入
- { $delimiter }: 输出样式
- { extracted_words }: 任务输出
- { $example }: 任务示例

参数设定

- temperature：0.2
- top_p：0.95
- top_k：0
- max_output_tokens：2048

图2　指令及参数设置

表1　索引数据库结构

字段名	含　义	数据类型	主键
id	记录号	bigint	✓
src_database	来源库	varchar	—
title	题名	varchar	—
authors	作者	varchar	—
organ	单位	varchar	—
source	文献来源	varchar	—
keywords	关键词	varchar	—
summary	摘要	longtext	—
pub_time	发表时间	varchar	—
first_duty	第一责任人	varchar	—
fund	基金	varchar	—
year	年	varchar	—
volume	卷	varchar	—
period	期	varchar	—
page_count	页码	varchar	—
clc	中图分类号	varchar	—
issn	国际标准刊号	varchar	—
url	网址	varchar	—
index_item	索引项	json	—
index_item_add	索引地址	json	—

（4）索引可视化

可视化是指将索引款目与索引地址的匹配集以 Web 可视化界面的形式呈现给用户。如图 3 所示，本文基于 Python Django 框架、SQL 数据库、HTML5 和 CSS3 技术构建了 Web 应用。最终能够实现在 Linux 操作系统部署与发布，用户可以通过任意浏览器访问该索引。

用户界面的设计遵循了直观性、一致性和反馈机制的原则，以确保用户友好性和提高互动性。根据《索引编制规则（总则）》（GB/T 22466-2023）对索引符号系统的要求[33]进行了统一处理和输出。按照设计心理学原理，优化了主题索引的呈现方式，按照主题词的音序分类排列，以助检符号确保信息清晰、有序并易于理解。此外，在信息架构的展示上，使用明确的菜单和描述语言，提供了直观的导航工具，帮助用户清晰地使用该索引数据库。

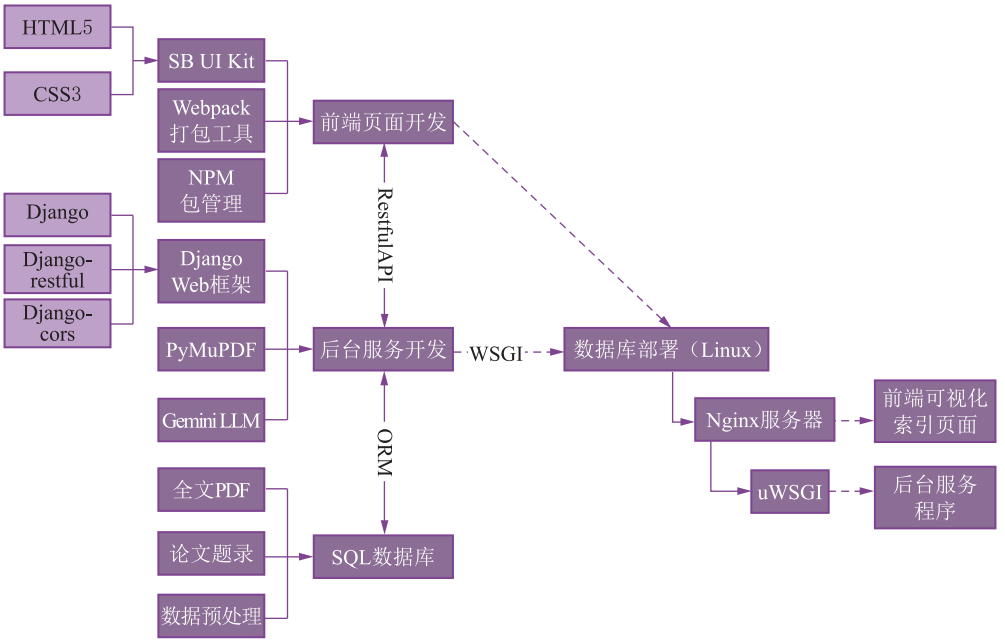

图 3　索引数据库开发技术路线图

2.3　索引数据库实现

论文摘要索引数据库旨在为用户提供结构化且友好的检索工具。为此，提供了索引编制说明、索引款目、主题索引总结和声明 3 个核心板块及索引款目中的主题词含义查询功能。分别简介如下：

（1）索引编制说明。索引编制说明是为了便于用户使用而前置的凡例，置于索引款目正文之前，说明了本主题索引采用的技术或方法。样例为：“本索引根据《图书与情报》2022 年全年刊出论文摘要编制，按标目拼音字顺排序。索引款目的结构为‘标目 + 页码’，页码以阿拉伯数字表示。相同标目不同页码只保留一个标目，页码按照从小到大的顺序依次连接，中间用半角逗号分隔。”

（2）索引款目。索引款目由标目和出处组成。此外，利用 HTML 标记语言的超链接技术，为各索引款目提供了实时在线的主题词含义查询功能。用户可以通过点击主题词快速访问其在《中国大百科全书》第三版网络版中的条目释义和相关词，从而增强了索引款目的交互性和解释性。助检符号“A”下的部分款目示例如下：

A

ABSA　73−76, 78, 80, 82

安全情报教育　32−38

奥地利图书馆协会　121−122, 118

（3）主题索引总结和声明。索引总结概述了索引对象的基本情况。更为重要的是，由于现阶段的生成式模型存在“幻觉”（Hallucination）问题，可能在主题款目抽取时引入部分非事实内容[34]。因此，在将生成式人工智能引入索引时，需要确保在诚信和透明原则下向用户充分、正确地披露索引编制者对生成式人工智能的使用情况。索引总结和声明样例如下：“《图书与情报》为双月刊，2022年共6期，刊载94篇论文。从中抽取出有检索意义的索引款目，按照音序排列，供检索使用。*声明：索引款目由Gemini Pro模型自动生成，属人工智能生成内容。作者在使用Gemini时仅考虑了提示词和参数设定，并未对生成内容进行编辑。”

2.4　款目抽取的性能分析

实验后，采用准确率（Precision，P）、召回率（Recall，R）和F1值（F-measure）[35]3个常用指标对款目抽取性能进行评价。图4展示了模型温度参数从0至1共11种情况下的性能情况。当温度为0.2时，出现极大值点，此时3项指标都接近0.9——P = 0.877、R = 0.900、F1 = 0.881。

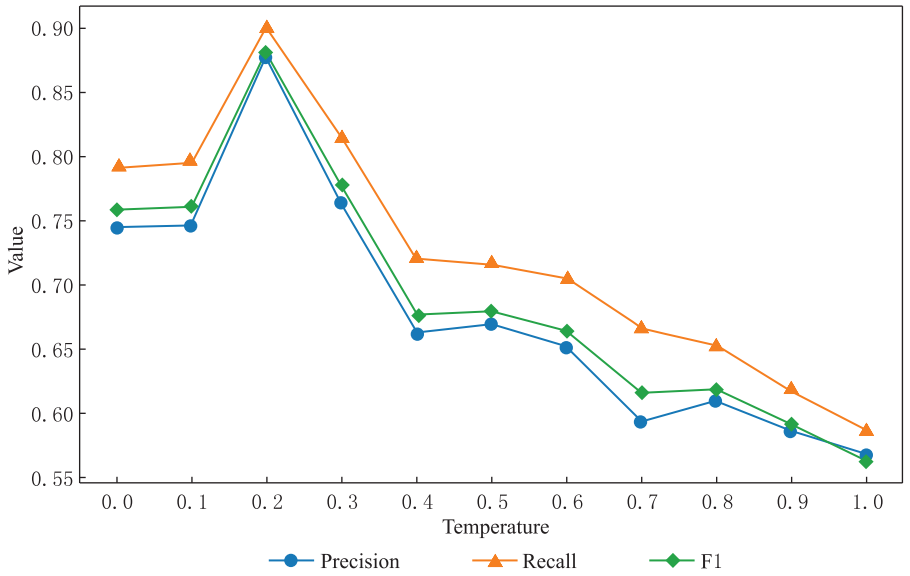

图4　各温度参数下的得分计算结果

与同类学术文献实体抽取任务相比，基于大模型的款目抽取性能表现优于条件随机场和SciBERT+CRF等模型效果[36]。尽管本方法性能略逊于基于领域知识库构建和训练的RoBERTa-BiLSTM-CRF模型[37]，但本方法在其不依赖于复杂模型训练的使用简便性以及向不同学科领域扩展的适用性上显示出一定应用优势。在深入内容维度评价款目抽取性能时，发现该方法能够有效识别新兴知识实体，如成功识别了“乡村飞阅计划”等具有特色的名词。此外，本方法可以通过引入如GPT 4.0等更强大的或经过微调的预训练模型来进一步提升款目抽取性能。

综上所述，可以认为使用“生成式人工智能＋指令工程”方式完成的论文摘要索引款目抽取任务，具有较高的精确度、覆盖度，能够较好地满足自动索引编制中的质量要求。

3　效用讨论

3.1　创新性

（1）索引方法创新。该数据库利用生成式人工智能RAG技术，深度理解和分析文献摘要所蕴含的知识概念，从而供给更为深入和全面的信息，能为用户提供更为准确的文献检索服务产品。这一索引方法创新不仅提高了大规模索引编制效率，还有助于索引自动编制工具更好地适应多样的学科领域。

此外，该数据库系统还提高了索引编制的经济性，主要成本是调用生成式人工智能服务企业提供的 API 的开销。随着技术和硬件的进步，这一成本会进一步降低，克服了传统手动编制的耗时和成本高的问题。

（2）索引对象创新体现在对摘要进行深度挖掘和应用。传统的索引方法通常以整篇文档文本作为主题词抽取对象，受限于抽词技术，可能导致识别不全、识别错误、过拟合的问题 [36]958–959。此外，全文内容的语料规模（中文论文通常为 10 000 字左右）又可能导致错误抽取、不宜抽取或过度抽取的主题词充斥整个词表，会降低抽词质量，造成大量的信息冗余 [38]，给人工校对、质量评价工作和用户检索带来较多不便。

由于检索关键词与文摘相较于检索海量全文具有更准确的结果和更高的效率，能更大程度地节省用户获取有用信息时间 [39]。因此，本文不使用全文索引范式，提出了以信息量更大、更准的摘要主题词为索引项。其关键在于将索引主题词标引的重点放在由论文作者书写的关键信息和新兴知识中，以确保索引的内容更贴近论文的核心思想。选择的索引款目更具有高度的检索价值，从而增强索引的检索质量。

（3）索引应用创新体现在与学术规范和评价的关联上。叶继元教授率先将索引、数据库与学术规范与评价的关系明确地关联起来，并指出索引 / 数据库在学术规范和学术评价中各有 3 个有利于：在学术规范中有利于提高文献内容质量、有利于“辨章学术、考镜源流”、有利于科研诚信建设；在学术评价中有利于辅助查明学术新贡献、有利于定量评价学术影响力、有利于辅助评价学术质量 [1]64–66。本文可被视为对叶继元教授这一理念的实践尝试和深入检验。通过将生成式人工智能技术应用于论文摘要索引数据库的设计与建设，不仅实现了检索工具的快速构建。更重要的是，能以“索引 / 数据库 + 人工智能”的组合尝试解决期刊论文形式规范和内容评价等问题。

此外，在面对大规模的摘要撰写规范检测和期刊评价需求时，论文摘要索引数据库能够引入机器学习算法自动检测论文创新性和规范性，减轻了专家阅读和判断的负担。

3.2　适用性

使用生成式人工智能进行主题标引实际上采用了基于理解的自然语言处理算法，不依赖人工编制的领域术语词典和数据字典。因此，该方法不受限于某一特定领域，而具有向不同学科扩展的适用性，突破了传统索引编制方法对领域知识的依赖。这种适用性的提升使得该数据库的索引方法更具通用性，为广泛且全面的索引编制提供了更灵活和高效的支持。

更为直观的是，通过与《图书与情报》编辑部编制的年度索引 [40] 对比（见图 5），年度索引的“专题：大情报观视野下的国家安全”以论文标题为索引款目揭示了该论文所在

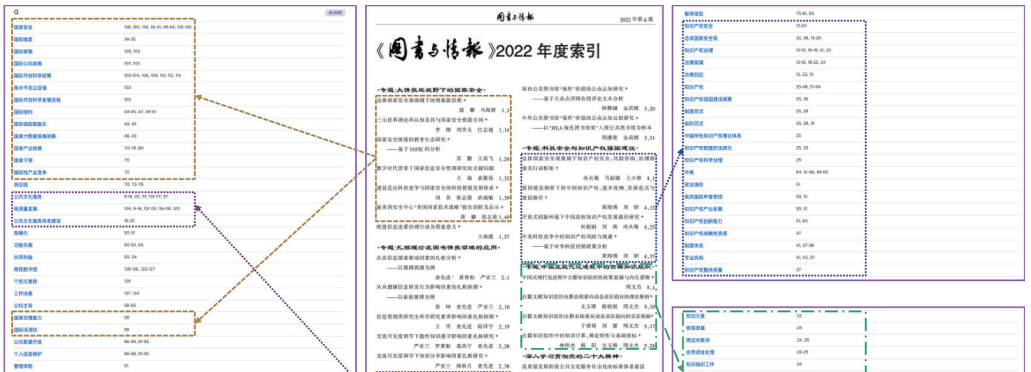

图 5 《图书与情报》年度索引与论文摘要索引对比对照

期次和当期起始页码，属于对论文外部形式特征的标引。而本文提出的索引编制方法则抽取到“国家安全”“国别维度”“国际政策”“国际规约”“国家干预”等主题词作为索引款目并将它们归入助检符号“G”下，从内容纬度实现了更细粒度的信息提取和整理能力。可以说，本文提出的索引编制方法为用户提供了基于论文主题内容的信息检索途径。若能将《年度索引》和论文摘要索引结合起来提供给用户，将有助于提高期刊的整体质量和吸引力。

3.3 扩展性

本文提出的索引编制方法的扩展性体现在索引目的、索引类型和索引功能 3 方面。

索引目的方面，本文以期刊的整卷检索为切入点，旨在帮助读者迅速检索该刊物整卷刊载的全部文献。此外，若以基金项目成果检索或评审为目的，则可将一个或一类基金项目所发表的全部论文集结并编制索引，为相关项目的研究提供检索和评价工具，从而满足多样化的信息检索需要。

索引类型方面，除主题索引外，还可利用大模型抽取人物、地点、机构和事件等实体，进而编制人名索引、地点索引和机构索引等作为主题索引的补充。多样化的索引类型可以丰富用户的检索纬度，满足用户多元化的检索需求，丰富了数据库的信息服务呈现形式。

索引功能方面，本文在功能设计阶段预留了学术评价子系统的扩展空间。这意味着未来可以通过添加评价指标和算法，帮助同行学者评价论文，乃至期刊的学术贡献、影响力。这种功能的扩展使得论文摘要索引数据库不仅是一个检索工具，更是学术评价和规范有力工具。同时，在收录数据量足够大的情况下，又或许可以适当将评价范围从论文评价拓展至横向的期刊评价。

4 总结与展望

由于论文，特别是核心、权威期刊刊载的论文通常代表了该领域的最新创新性成果，因此可能存在大量过去未知的新名词和新思想。传统的自然语言处理技术受限于内置词典或规则，可能无法完全、准确地识别出新兴主题词；而大语言模型则是基于深度学习算法，通过海量数据集训练、创建的，涌现出一定的对未知信息的推理能力。基于上述认识，本文采用生成式大语言模型的 RAG 技术，更全面、更深度地理解和分析论文摘要，提取其中的关键信息和新兴知识作为索引词，进而对全文进行主题标引。

论文摘要索引数据库不仅弥补了过去基于规则和概率的索引编制方法中的不足，还特别适用于大规模、大批量的索引编制，更展示了生成式人工智能在索引学和索引领域的潜在价值，为未来生成式人工智能技术在图书情报领域的应用提供了参考。

值得指出，由于该索引是人工智能生成内容（AIGC），或更准确地称其为：人工智能生成索引（AIGI，AI Generated-Index），存在受人工智能算法影响而造成的一定的“幻觉”问题。然而，随着算法的优化和精准算据的补充，相信由生成式人工智能驱动所编制的索引将更精当和全面。

（本文数据链接地址：http://hdl.handle.net/20.500.12304/11356）

参考文献

[1] 叶继元 . 索引的本质属性及其在学术规范与评价中的作用 [J]. 图书情报知识，2023，40（6）：61-67.

[2] 高建群，吴玲，施业 . 学术论文摘要的规范表达 [J]. 东南大学学报（哲学社会科学版），2003（2）：114-117.

[3] 全国信息与文献标准化技术委员会（SAC/TC 4）. 学术论文编写规则：GB/T 7713.2-2022[S]. 北京：中国标准出版社，2022.

[4] 叶继元 . 人文社会科学评价体系探讨 [J]. 南京大学学报（哲学・人文科学・社会科学版），2010，47（1）：97-110；160.

[5] 吕远，叶继元 . 论文摘要学术规范自动检测模型研究初探 [J]. 图书馆，2019（4）：24-29.

[6] 陆伟，刘家伟，马永强，等 . ChatGPT 为代表的大模型对信息资源管理的影响 [J]. 图书情报知识，2023，40（2）：6-9；70.

[7] 张琪玉 . 现代的索引就是数据库 [J]. 图书馆杂志，2001（12）：6-7.

[8] 邱均平，马力，杨强 . 数字出版环境下书后索引发展研究 [J]. 图书馆杂志，2016，35(3)：68-73.
[9] 孙辉 . 利用信息组织技术编制书刊索引探析 [J]. 现代情报，2015，35(1)：96-99；103.
[10] 刘艳 . 综合性百科全书内容索引的编制 [J]. 出版发行研究，2018(5)：79-81.
[11] 宋林青 . 编制图书内容索引的价值研究 [J]. 出版与印刷，2019(1)：54-57.
[12] 李炜超，王雅戈 . 学位论文内容索引编制研究——以《萨都剌生平及著作实证研究》索引编制为例 [J]. 图书馆杂志，2018，37(6)：38-43.
[13] 张淑文 . 编辑视角的学术专著书后主题索引编制浅谈 [J]. 编辑之友，2018(8)：90-94.
[14] 贺玢，杨学红，王鲁燕 . 农机化专业术语索引编制 [J]. 农业工程，2021，11(10)：112-117.
[15] 杨斐，叶继元，王雅戈，等 . 地方志索引管窥 [J]. 图书馆论坛，2019，39(11)：47-50.
[16] 张卓杰 . 县级综合年鉴索引编制的设计——以绍兴地区年鉴为例 [J]. 中国年鉴研究，2022(1)：26-32；78.
[17] 王琦，靳嘉林，王曰芬，等 . 词序视角下学术文本关键词分布特征及其差异研究 [J]. 情报杂志，2022，41(8)：171-178.
[18] 丁海英 . 用 Word 软件编制小型专题索引 [J]. 图书馆理论与实践，2002(4)：86-87.
[19] 王彦祥 . 中国索引软件的开发与应用 [J]. 中国索引，2009(2)：53-57.
[20] 王雅戈，叶继元，黄建年，等 . 中文索引平台建设——以“索引家”开发为例 [J]. 图书馆论坛，2019，39(11)：37-40.
[21] 王雅戈，叶继元，黄建年，等 . 学位论文索引的演进 [J]. 图书馆论坛，2019，39(11)：44-46.
[22] 潘雪莲，侯汉清，许扬威 . 图书内容主题索引的自动编制实验 [J]. 大学图书馆学报，2008(3)：28-33.
[23] 侯汉清，薛鹏军 . 基于知识库的网页自动标引和自动分类系统的设计 [J]. 大学图书馆学报，2004(1)：50-55；64.
[24] Shao Z，Gong Y，Shen Y，et al. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy[C/OL]// Findings 2023. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.620/.
[25] Gao Y，Xiong Y，Gao X，et al. Retrieval-Augmented Generation for Large Language Models：A Survey：arXiv：2312.10997[EB/OL]. arXiv，2024. http://arxiv.org/abs/2312.10997.
[26] Khattab O，Santhanam K，Li X L，et al. Demonstrate-Search-Predict：Composing retrieval and language models for knowledge-intensive NLP：arXiv：2212.14024[EB/OL]. arXiv，2023. http://arxiv.org/abs/2212.14024.
[27] 张心源，邱均平 . 国内外数据库索引编制研究的进展与趋势 [J]. 图书馆杂志，2016，35(3)：60-67.
[28] 张琪玉 . 图书索引软件的功能要求与编制难题 [J]. 中国索引，2004，2(3)：41.
[29] Gemini-Google DeepMind[EB/OL]. [2024-01-05]. https://deepmind.google/technologies/gemini/#introduction.
[30] 陆伟，汪磊，程齐凯，等 . 数智赋能信息资源管理新路径：指令工程的概念、内涵和发展 [J]. 图书情报知识，2024，41(1)：6-11.
[31] Dai Z，Zhao V Y，Ma J，et al. Promptagator：Few-shot Dense Retrieval From 8 Examples：arXiv：2209.11755[EB/OL]. arXiv，2022. http://arxiv.org/abs/2209.11755.
[32] Google AI. About generative models[EB/OL]. [2024-05-13]. https://ai.google.dev/gemini-api/docs/models/generative-models#model-parameters.
[33] 全国信息与文献标准化技术委员会(SAC/TC 4). 索引编制规则(总则)：GB/T 22466-2023[S]. 2023.
[34] Filippova K. Controlled Hallucinations：Learning to Generate Faithfully from Noisy Data[C/OL]// Findings 2020. Association for Computational Linguistics. November 2020. Online. https://aclanthology.org/2020.findings-emnlp.76.
[35] 刘仕阳，王威威，化柏林 . 多源数据环境下公共文化服务机构年报的数据抽取研究 [J]. 图书馆杂志，2020，39(12)：52-60.
[36] 章成志，谢雨欣，张恒 . 学术文献全文内容中的方法实体细粒度抽取及演化分析研究 [J]. 情报学报，2023，42(8)：952-966.
[37] 刘懋霖，赵萌，王昊 . 面向古诗词的物象库构建方法及其分布规律研究 [J]. 图书馆杂志，2024，43(1)：96-108.
[38] 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.
[39] 蔡迎春，赵心如，朱玉梅，等 . 我国文献标引技术的回顾与展望 [J]. 图书馆杂志，2022，41(3)：18-31.
[40] 图书与情报编辑部 .《图书与情报》2022 年度索引 [J]. 图书与情报，2022(6)：142-144.

朱　禹　南京大学信息管理学院，硕士研究生。研究方向：信息资源建设、人工智能生成内容。作者贡献：论文撰写与修改。E-mail：zhu.yu@smail.nju.edu.cn　江苏南京　210023
叶继元　南京大学信息管理学院，教授，博士生导师。研究方向：信息资源建设、图书情报学理论与方法、学术规范与评价。作者贡献：论文指导与提出修改意见。江苏南京　210023

（收稿日期：2024-04-15　修回日期：2024-06-04）