人工智能与中医药前沿研究专栏

DOI: 10.16305/j.1007-1334.2026.z20251126007

人工智能赋能的中医药标准信息化存储与检索系统设计与实现

邢　可[1]，陈立铭[2,3*]，吴蒙捷[1]，何心如[2,3]，吴　阳[1]，高启化[4]，李　静[2,3]，桑　珍[5]

1. 上海超级计算中心（上海　201203）；2. 上海中医药大学（上海　201203）；3. 上海市中医药国际标准化研究院（上海　201203）；4. 上海复楚智能科技有限公司（上海　201203）；5. 上海中医药大学附属曙光医院（上海　200021）

【摘要】 **目的**　构建一个基于人工智能的中医药标准信息处理管线，即中医药标准信息化存储与检索系统，实现中医药领域内标准文件的结构化读取、入库、检索，推动中医药标准体系化建设的智能化发展。**方法**　从中医药标准文件的结构特征出发，采用MinerU等格式转换工具进行信息提取，创新性地建立包含印章去除、文件类型判别和并行解析的处理流程，采用基于有限状态机的标准主干提取方法，采用分层策略处理高度非结构化的领域专业内容。依托大语言模型（LLM），结合检索增强生成（RAG）技术及特定提示词实现标准数据的精准检索，并进行验证。**结果**　研究构建的中医药标准信息化存储与检索系统可支持主流LLM对其进行RAG语义检索，利用包含中医药领域内的国家标准、地方标准、行业标准等400余条标准数据对系统进行验证，提出的文件格式转换算法、标准主干提取算法准确率均在95%以上。**结论**　研究设计的中医药标准信息化存储与检索系统能够满足标准信息提取、入库、检索等全流程的信息化需求。

【关键词】 中医药标准；结构化存储；检索；人工智能；大语言模型

Design and implementation of an AI-empowered informationized storage and retrieval system for traditional Chinese medicine standards

Xing Ke[1], Chen Liming[2,3*], Wu Mengjie[1], He Xinru[2,3], Wu Yang[1], Gao Qihua[4], Li Jing[2,3], Sang Zhen[5]

1. Shanghai Supercomputer Center, Shanghai 201203, China; 2. Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; 3.Shanghai Academy of International Standardization for Traditional Chinese Medicine, Shanghai 201203, China; 4.Shanghai Future DATech Co., Ltd., Shanghai 201203, China; 5.Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 200021, China

Abstract: Objective　To develop an artificial intelligence-based processing pipeline for traditional Chinese medicine (TCM) standard information, specifically, an informationized storage and retrieval system for TCM standards. This system aims to achieve structured extraction, database integration, and retrieval of TCM standard documents, thereby advancing the intelligent development of the TCM standardization system. **Methods**　Based on the structural characteristics of TCM standard documents, information extraction was performed using format conversion tools such as MinerU. The research group innovatively established a processing flow that included seal removal, file type discrimination, and parallel parsing. It adopted a standard backbone extraction method based on finite state machines and a hierarchical strategy to handle highly unstructured domain-specific content. Leveraging large language model (LLM) combined with retrieval-augmented

[基金项目] 上海市卫健委进一步加快中医药传承创新发展三年行动计划项目（2025年-2027年）[ZY(2025-2027)4-2-1]；2021年上海市标准化推进专项资金项目（2021-A-228）；上海市中医药国际标准化研究院内涵建设经费（BZH-N26-0103）

[作者简介] 邢可，男，硕士研究生，主要从事人工智能和高性能计算研究工作。*陈立铭为共同第一作者

[通信作者] 李静，副研究员，硕士研究生导师；E-mail：ivylee0109@126.com。桑珍，主任医师，博士研究生导师；E-mail：sangzhen8507@hotmail.com

generation (RAG) and domain-specific prompt engineering, the system achieved precise retrieval of standard data, followed by rigorous validation. **Results** The developed system for TCM standards supports semantic retrieval via mainstream LLM using RAG. Validation was conducted using over 400 standards, including national, local, and industry standards within the TCM field. The proposed document format conversion and backbone extraction algorithms both achieved accuracy rates exceeding 95%. **Conclusion** The AI-driven storage and retrieval system designed in this study effectively meets full-process informatization requirements, ranging from information extraction and database entry to retrieval of TCM standards.

Keywords: traditional Chinese medicine standards; structured storage; retrieval; artificial intelligence; large language model (LLM)

中医药标准作为中医药领域的一项关键性基础建设，不仅为中药材、中医药疗法、生产技术、运营管理等领域制定了统一的规范，也为中医药数据的跨平台、跨行业交流及促进贸易、防范公共卫生事件等提供了可行性指导[1-3]。随着中医药领域相关标准数量的不断增加，如何在国家标准、地方标准、行业标准等复杂类别下对标准进行统一有效管理，从而为后续对标准体系挖掘分析提供基础性支持成为难题。

在早期中医药信息化建设时期，一些专家学者对中医药标准体系框架展开了初步研究，促进了中医药标准体系的顶层设计建设[4-6]。有研究[7-8]针对中医药信息标准体系建设的现状进行分析，认为中医药信息标准之间协调性较差，应用推广受限。在医疗大健康标准的信息化领域，张柯欣等[9]以 cjk-unicode16 汉字编码对照表、框架词典表和术语数据表为基础，设计了标准数据库，并实现了基于中医临床术语标准分析中医临床文献的需求。李楠[10]在中医儿科标准建设成果的基础上，进行中医儿科标准信息数据库建设与研制，并针对建设过程中标准采集、加工、录入等工作中的重点、难点问题进行分析。在文档的信息化存储方面，有学者[11]利用 MongoDB 文档数据库搭建了信息存储平台，并使用 Elasticsearch 作为搜索引擎，从而支持复杂场景的查询需求。但是 Elasticsearch 的使用增加了系统的复杂性，给系统的稳定性、后期运营维护增加了成本。在信息提取领域，一些学者[12-14]进行了系统性调研，并开始使用机器视觉赋能信息提取。

我国早期“互联网+中医药”标准信息化体系建设并不完善，存在可移植文档格式(PDF)文件不可识别、格式不统一、文件质量参差不齐、海量标准文件存储与复杂检索查询需求之间难以平衡等问题[15]。自 2022 年 ChatGPT 出现后，人工智能和大语言模型(LLM)技术飞速发展，为解决上述问题提供了可行性[16-17]。本研究从中医药标准文件的结构特征出发，在 LLM 技术支持下设计并构建中医药标准信息化存储和检索系统，其创新性体现在以下 3 个方面。一是针对扫描件、文本型 PDF 和图层型 PDF 等不同来源文件，建立包含印章去除、文件类型判别及并行解析的处理流程；二是依据中医药标准中“范围”“规范性引用文件”“术语与定义”等常见章节结构，采用有限状态机提取标准主干信息；三是采用分层策略处理高度非结构化的领域专业内容。具体介绍如下。

1　设计思路

为了解决信息化体系建设中的上述结构性问题，完成从标准文件可阅读到可分析再到可检索的跨越，我们旨在搭建一条基于人工智能的中医药标准信息处理管线，即中医药标准信息化存储与检索系统。

首先，采用 MinerU 等基于人工智能技术的格式转换工具进行信息提取；多维度 PDF 分类算法和多文件并行处理优化算法，让信息提取的过程更加高效和准确。

其次，利用计算机视觉技术实现印章去除功能，提高光学字符识别(OCR)[18]的准确性；采用基于自动机的标准主干提取算法，最大化地减少因 PDF 文件质量问题对提取结果的影响；使用 MongoDB 文档数据库[19]存储海量标准数据。

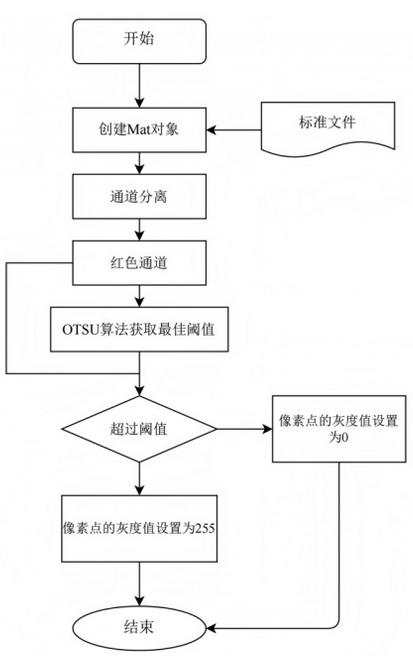
最后，依托最新的检索增强生成(RAG)技术，结合设置的特定提示词实现标准数据的精准检索[20]，并将检索得到的中医药标准关键信息提供给前端用户界面(UI)，实现标准文本的交互式操作。

2　系统构建

2.1　人工智能赋能信息提取　信息提取作为中医药标准信息化存储与检索系统信息处理的核心功能，承担着从原始 PDF 格式的标准文件提取关键信息并进行结构化表示的任务。传统的信息提取方式使用商业化接口，存在使用成本高、文件大小受限需要分割处理、难以针对中医药标准文件实现定

制化优化、数据安全性低、敏感标准文件不适合外部处理等问题。上海人工智能实验室的开源项目MinerU[21]是一款智能文档解析工具，能够将PDF文档转换为结构化的Markdown格式，进而弥补上述缺陷。该工具具有完全开源免费、支持大文件直接处理、可针对特定领域深度优化、数据处理全程本地化保障信息安全、持续技术更新等优点。本研究设计的中医药标准信息化存储和检索系统基于MinerU的2.1.7版本[22]完成定制化开发，在预处理阶段引入开源计算机视觉库OpenCV 4.10.0[23]的印章去除算法，并在后处理阶段采用有限状态机的标准主干结构提取技术。

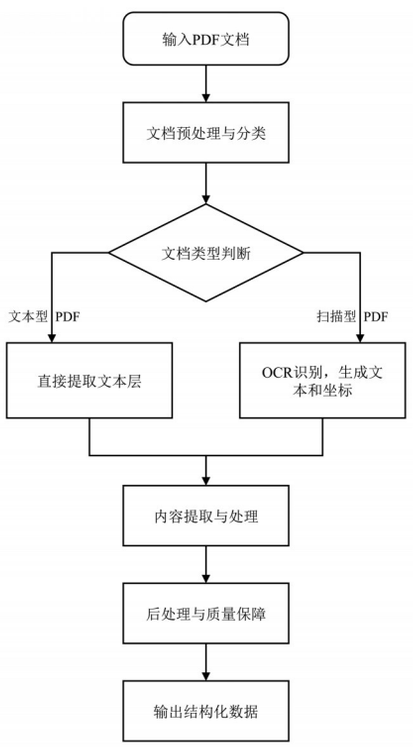
2.2 基于OpenCV的印章去除算法 由于防伪造、防篡改等要求，中医药领域的标准文件中往往含有印章，导致文字、表格等关键信息被遮挡，这给标准文件的信息提取工作带来了困难。基于OpenCV 4.10.0的印章去除算法，可以去除标准文件中的印章图样。该算法的核心流程如图1所示，首先对原始图片创建存储和处理图像的基本单元（Mat对象）[24]，然后使用通道分离技术[25]将原始图像分割为BGR三通道图形，获取红色通道，使用最大化类间方差（OTSU）算法获取最佳阈值，最后使用阈值分隔技术[26]获取到最终去除印章的标准文件并输入MinerU进行解析。



注：Mat对象指OpenCV存储和处理图像的基本单元，OTSU算法指最大化类间方差算法。

图1 基于OpenCV的印章去除算法流程

2.3 多维度PDF文件分类处理逻辑规则 为了使信息抽取的过程更加高效和准确，同时避免额外的模型开销，本研究在MinerU中增加了多种逻辑规则，进而对PDF文件进行分类（图2）。首先通过检查PDF元数据判断文档类型，即图像型（扫描版）、文本型（可复制文本）或图层型（不可复制文本）。若为图像型PDF，则自动启用OCR模式；若为文本型PDF，则直接提取文本内容；若为图层型PDF，则采取智能分类与混合解析策略。然后分析检测图像占比、文本结构、排版方式等文档内容特征，根据检测情况选择不同的模型有利于提高解析质量。若文档存在大量的图像表格，则采用表格识别模型将表格图像识别为相应的源代码。此外，还可以通过配置参数指定解析方法，既考虑了自动化处理，又提供了手动控制的灵活性。



注：OCR指光学字符识别。

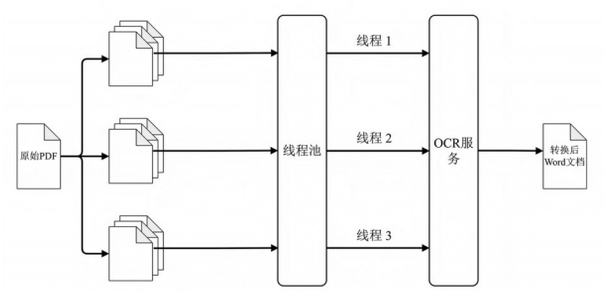
图2 可移植文档格式（PDF）文件分类处理算法流程

2.4 多文件并行处理优化 本研究采用分层批处理架构提高MinerU多文件处理的速度，从模型推理到资源管理都进行了深度优化。

首先是基础批处理配置优化。该过程包括4个优化参数：布局检测批处理，使用YOLO模型批量检测文档布局；公式识别批处理，并行处理数学公式检测与识别；OCR检测批处理，按语言和分辨率分组进行OCR；表格识别批处理，批量处理文档中的

表格结构。以上4个参数与批处理放大系数(batch_ratio)相乘,再结合实际硬件资源进行调参,即可找到最优参数。

其次是并行批处理。本研究通过增加线程池提高并行效率,如图3所示,给定包含n页的原始PDF文档,单次调用的PDF文档页数为m,则系统并发数为n/m,从系统线程池中使用n/m个线程发起调用,获得相应结果后,将n/m个转换结果按照原始顺序进行拼接,得到最终的转换后结果。



注:PDF为可移植文档格式。

图3　基于线程池的大文件光学字符识别(OCR)格式转换

此外,本研究还采用了内存管理优化和资源分组策略来最大化批处理效率。内存管理优化通过定期监控并周期性释放资源,从而有效降低内存占用;资源分组则采用"语言-分辨率"二阶段策略,确保图像尺度统一(对齐至32的倍数)且语言一致,从而最高效地发挥硬件计算能力。

2.5　基于有限状态机的标准主干结构提取　标准文件中目录主干结构包含范围、规范性引用文件、术语与定义等关键信息,能够反映中医药标准的作用对象、上下位文件、专有名词等关键内容,因此需要根据标准文件目录中存在的标号、描述的结构特性、排布规律识别目录结构以及关键内容,并进行结构化存储,以便进行后续分析工作。

本研究提出了一个基于有限状态机的标准主干提取算法,并由此来完成标准主干结构的提取工作。根据标准文件的格式特点,"范围""规范性引用文件"为通用固定格式,因此我们将某行匹配到"范围"并作为有限状态机的起始状态,当某行匹配到"规范性引用文件"时则跳转到下一状态。然后根据标号的变化特点确定状态转移逻辑:标号2的下一状态包含子标号"2.1"和同级标号"3";标号"2.1"的下一状态则包含"2.1.1"、"2.2"及"3"3种可能。据此构建状态转移图,最终完成标准目录结构的主干提取,如图4所示。

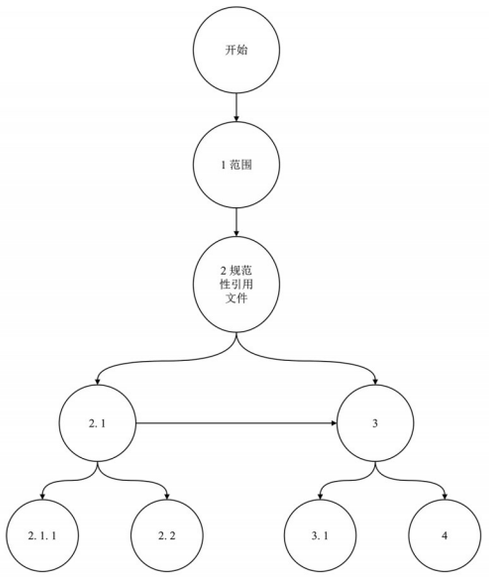


图4　基于有限状态机的标准主干结构提取状态转移图

2.6　领域专业内容处理策略　中医药标准文件中常包含复杂配方表、炮制工艺流程图及古籍引用等高度非结构化的领域专业内容,针对这些内容,本研究采用分层处理策略。

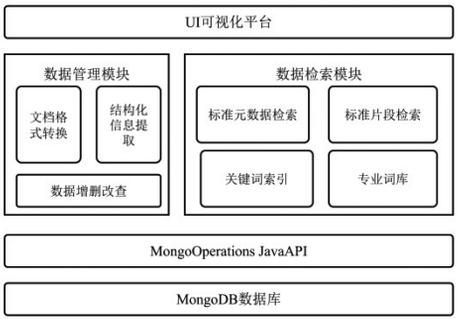
(1)复杂配方表处理。基于MinerU的表格识别模块,采用深度学习的表格结构识别算法,将配方表中的药材名称、剂量、配比等信息提取为结构化JSON格式。对于包含合并单元格、嵌套结构的复杂表格,保留其原始布局信息,支持后续人工校验与补充标注。

(2)流程图处理。将标准文件中的流程图、示意图等图像内容作为附件资源进行关联存储,保留图像与上下文文本的位置关系,并提取图像周围的标题、图注等文本信息建立索引,支持基于图注的检索定位,以待未来采用。

(3)古籍引用处理。通过正则匹配识别书名号标记的古籍(如《黄帝内经》《伤寒论》)名称,初步建立古籍清单,并结合上下文提取引用片段,为后续构建古籍溯源知识库提供数据基础。

2.7　系统框架设计　如图5所示,本系统包含前端UI可视化平台、数据管理模块、数据检索模块等多个部分,底层基于MongoDB数据库实现标准数据的海量存储,同时使用GridFS分布式文件系统实现大文件存储的功能。

2.7.1　数据管理模块　数据管理模块主要负责对



注：UI指用户界面。

图5　系统框架设计

标准数据进行统一管理，包括标准数据的增加、修改及删除。该模块对读入的标准内容进行分析，采用基于有限状态机的标准模式识别算法读取标准主干信息，同时提供MinerU格式转换工具用于完成PDF文件到Markdown文件之间的转换。具体而言，数据以JSON形式存储在文档数据库MongoDB中，基于MongoDB动态存储的特性，可以对存储结构进行灵活调整，便于后续对存储结构进行更新、升级等操作。同时高性能、易伸缩、易扩展的特性可以满足未来平台数据不断膨胀、需要进行横向扩展的场景。为了实现大文件存储的功能，使用GridFS分布式文件系统作为大文件存储规范，将原始文件分割为256 KB大小的块进行存储，从而实现存储超过16 MB（BSON文件限制）的文件，见图6。为了实现标准数据的多元化信息存储能力，标准的存储采用如下格式。

元数据字段：存储编号、标准类型、标准号、标准名称、发布省市、发布单位、发布日期、标准内容、原始文件、关键词组、Word文件ID、PDF文件ID。

文件：Word原文件，PDF原文件。

其中Word原文件与PDF原文件通过ID索引的方式与标准元数据进行关联。

在对标准数据的结构化处理时，我们还提取了“规范性引用文件列表”“代替标准号”“被代替标准号”“废止日期”等字段，用于标识标准之间的引用关系以及代替关系。目前的系统以单文档的结构化入库和检索为主，待未来标准数据进一步扩充后，我们将进一步构建标准引用的网状结构，并在系统前端进行可视化。

2.7.2　*数据检索模块*　数据检索模块负责对标准数据进行检索，该模块在LLM结合RAG向量知识库的增强技术下，可实现检索引擎对多语种自然语言描述的可接受性，达到人工智能技术下语义增强检索的效果，从而满足多元化的查询需求。

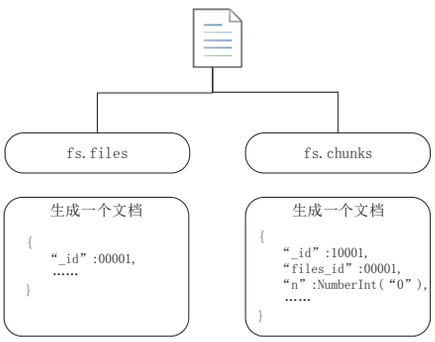
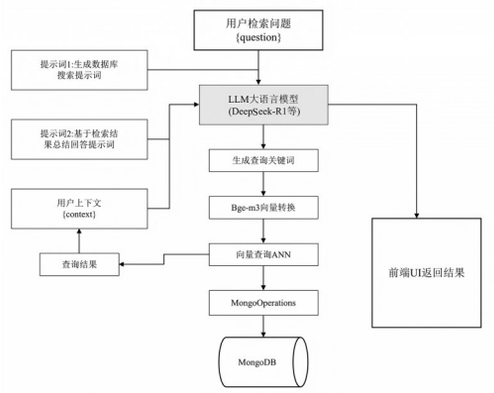


图6　GridFS文件存储示意

目前LLM基于表征学习和分布语义假设，将离散文本词（prompt提示词）映射到高维向量空间成为一个密集的embedding向量，由于在此空间语义相似的词或句子向量几何接近，模型基于Transfomer架构的自注意力机制来计算序列中每个位置（单词）和其他位置的“关联度”，即Query（Q）向量、Key（K）向量和Value（V）向量，通过将检索关键词的Q向量与其他所有词的K向量点积得到注意力分数，然后进行分数的Softmax归一化，得到关键词的注意力分布[27]。通过这种注意力机制，LLM可以实现上下文感知和标准文件中长文本情境下远距离的语义依赖建模，从而实现深度的语义检索。为了实现对MinerU已提取并结构化存储到MongoDB数据库的数据进行智能化检索，提升标准库的检索效率和准确度，本系统基于RAG技术结合最新的LLM来实现对MongoDB数据库的智能化语义检索。RAG技术利用基于Transformer架构的强大表示学习能力，通过稠密向量检索方法，将数据库中的内容编码成高维空间中的嵌入向量。在这种方法中，语义相近的文本在向量空间中的距离也更近，从而实现查询内容与相关内容的快速粗检索[28]。在粗检索结果的基础上，结合LLM的深度语义理解能力进行重排序与筛选，最终获得与用户问题最相关的精细检索结果[29]。RAG提供了检索的相关上下文数据库，这约束了LLM的生成空间，让LLM仅限于已存储的数据库内容进行检索，可降低“编造”答案的可能性[30-33]。检索技术架构如图7所示。

本系统采用Bge-m3嵌入模型将标准文本分块（chunk大小为768）后生成向量嵌入，存储于MongoDB向量数据库并创建vectorSearch类型的向量索引[34]。检索时，系统首先调用LLM优化用户查询，然后通过近似最近邻搜索（ANN）算法[35]获取



注:LLM指大语言模型,RAG指检索增强生成,UI指用户界面。

图7　基于RAG技术结合LLM的智能检索架构

Top-10相关文档块作为上下文,最后结合提示词引导LLM生成检索结果。本系统测试了ChatGPT5、Qwen3-32B、Llama-4-Maverick-17B 和 DeepSeek-R1等模型,这些模型在文本信息提取和处理方面具有较高的准确度和性能[36-37]。提示词设计是高效利用LLM的关键,本系统设置了两类提示词,图8展示了用于数据库检索关键词生成的提示词,图9展示了用于基于检索结果进行总结回答的提示词。

```
你是一名中医药标准化技术文档检索专家。你的任务是将用户的日常提问，转换成与中医药标准文件内容高度匹配的、正式的专业术语和关键词，以便进行精确的向量数据库检索。

请遵循以下规则：
1. 术语标准化：使用最正式、最标准的技术名称。
2. 同义词扩展：涵盖核心技术的标准名称、别名、简称及其英文缩写。
3. 结构化解构：从标准文档的常见章节中提取关键词。
4. 输出格式：仅输出一串用逗号分隔的关键词，不要任何解释。

用户问题：{question}

请生成关键词：
```

图8　数据库搜索关键词生成的提示词

```
你是一名中医药标准规范的释义专家。请严格根据提供的<上下文>内容回答问题，该上下文来源于权威的中医药标准规范相关文件。

你的回答必须绝对遵循以下规则：
1.答案必须完全源自提供的<上下文>内容，不得编造、推测或添加任何外部知识
2.如果上下文中对某个部分（如“治疗器械”）没有明确描述，请在相应部分填写“未见明确标准”
3.语言需专业、简洁、客观。
4. 回答附上出处说明
*本回答内容均来源于提供的内部技术标准文档，仅供参考。具体操作请以最新版官方文件为准。*

<上下文>
{context}
</上下文>
```

图9　基于检索结果进行总结回答的提示词

2.8　图形界面设计　为了更加方便地使用中医药标准信息化存储检索系统,本系统设计了支持人机交互的前端可视化UI,用户能够在该界面中完成标准录入、格式转换、标准检索等操作。图10展示了标准列表页面。图11为标准数据导入页面,输入标准的元数据,上传标准PDF、Word文件后,系统自动完成信息提取,将其结构化存储至数据库中。图12为标准检索页面,可以输入标准内容中多个关键词组,系统按照关键词包含的数量对检索到的标准进行排序展示。图13展示了标准内容页面,检索的词组在标准文本中用黄色高亮表示。

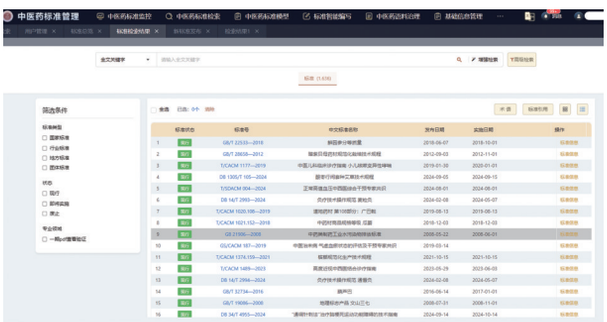


图10　标准列表页面

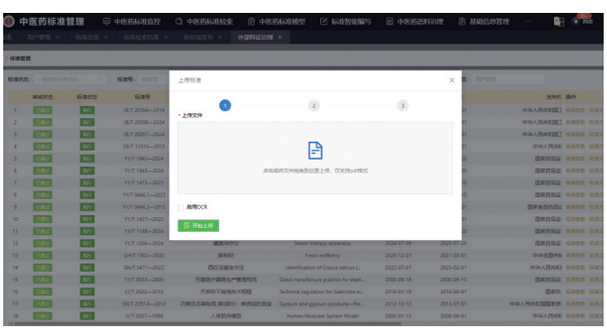


图11　标准录入页面

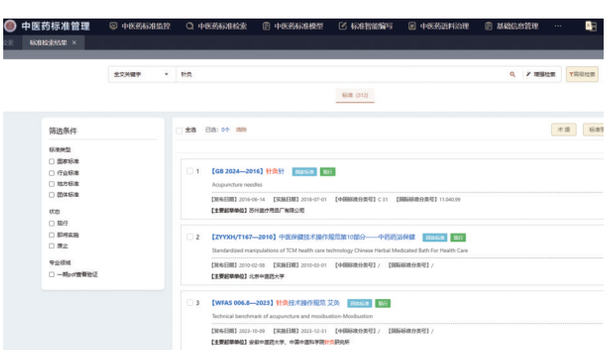


图12　标准检索页面

图13　标准内容展示页面

3　应用验证

3.1　数据集　验证数据主要包含中医药大健康领

域内的国家标准、地方标准、行业标准以及国家标准化指导性技术文件(GB/Z)4类,涵盖中药材、中医技法、生产技术规范、运营管理规范等,共计400余份,相关统计信息如图14所示。

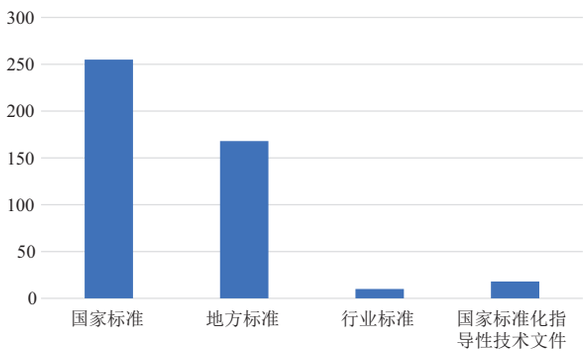


图14 验证数据集分类情况

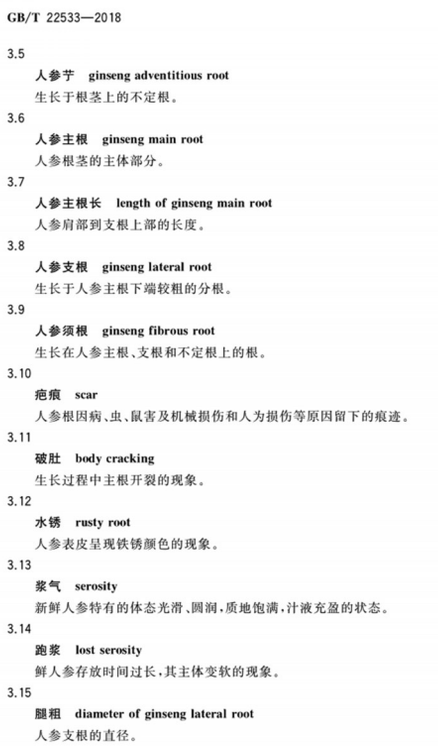
3.2 文件格式转换分析 图15、图16展示了使用印章去除前后的格式转换效果对比图。可以看到在开启印章去除功能之前,无法提取被印章覆盖部分的文字,并在word文档中只能以图片的形式展示,开启文档格式转换功能之后,能够有效提取全部关键信息。


GB/T 22533—2018

3.5
人参芋 ginseng adventitious root
生长于根茎上的不定根。
3.6
人参主根 ginseng main root
人参根茎的主体部分。
3.7
人参主根长 length of ginseng main root
人参肩部到支根上部的长度。
3.8
人参支根 ginseng lateral root
生长于人参主根下端较粗的分根。
3.9
人参须根 ginseng fibrous root
生长在人参主根、支根和不定根上的根。
3.10
疤痕 scar
人参根因病、虫、鼠害及机械损伤和人为损伤等原因留下的痕迹。
3.11
破肚 body cracking
生长过程中主根开裂的现象。
3.12
水锈 rusty root
人参表皮呈现铁锈颜色的现象。
3.13
浆气 serosity
新鲜人参特有的体态光滑、圆润,质地饱满,汁液充盈的状态。
3.14
跑浆 lost serosity
鲜人参存放时间过长,其主体变软的现象。
3.15
腿粗 diameter of ginseng lateral root
人参支根的直径。

图15 印章去除之前的光学字符识别(OCR)结果

3.3 多文件并行处理优化 我们基于MinerU 2.1.7版本进行性能优化,使用技术包括通过合理设置批处理放大系数、根据硬件资源调整batch_ratio参数、采用线程池并行执行多文件处理、定期清理显存、


GB/T 22533—2018

3.5
人参芋 ginseng adventitious root
生长于根茎上的不定根。
3.6
人参主根 ginseng main root
人参根茎的主体部分。
3.7
人参主根长 length of ginseng main root
人参肩部到支根上部的长度。
3.8
人参支根 ginseng lateral root
生长于人参主根下端较粗的分根。
3.9
人参须根 ginseng fibrous root
生长在人参主根、支根和不定根上的根。
3.10
疤痕 scar
人参根因病、虫、鼠害及机械损伤和人为损伤等原因留下的痕迹。
3.11
破肚 body cracking
生长过程中主根开裂的现象。
3.12
水锈 rusty root
人参表皮呈现铁锈颜色的现象。
3.13
浆气 serosity
新鲜人参特有的体态光滑、圆润,质地饱满,汁液充盈的状态。
3.14
跑浆 lost serosity
鲜人参存放时间过长,其主体变软的现象。
3.15
腿粗 diameter of ginseng lateral root
人参支根的直径。

图16 印章去除之后的光学字符识别(OCR)结果

针对不同类型资源进行分组等。经测试,与单文件串行处理模式比较,批量处理性能可提升1.76~3.76倍,见表1。

表1 多文件并行处理优化情况

处理模式	文件数量/个	总页数/页	处理时间	相对速度/倍
单文件串行	10	175	4 min 42 s	1
基础批处理	10	175	2 min 40 s	1.76
并行批处理	10	175	2 min 04 s	2.27
优化并行	10	175	1 min 15 s	3.76

3.4 标准主干提取分析 定义标准主干提取准确率为提取到的主干信息中真正位于标准主干结构上的数量所占的比例,标准主干覆盖率为标准主干结构中被准确提取的信息所占的比重。图17展示了不同种类的标准文件提取准确率和覆盖率情况,可以看到国家标准由于最为规范,准确率和覆盖率较高,行业标准、地方标准由于格式上与国家标准相比标准化程度较低,因此准确率和覆盖率较低。但是综合来看,所提出的基于有限状态机的标准主干提取算法能够较为准确、全面地提取出标准的关键信息。

3.5 智能信息检索结果分析 本系统基于“2.7.2”项下所述的RAG技术架构,采用图8、图9所示的提示词

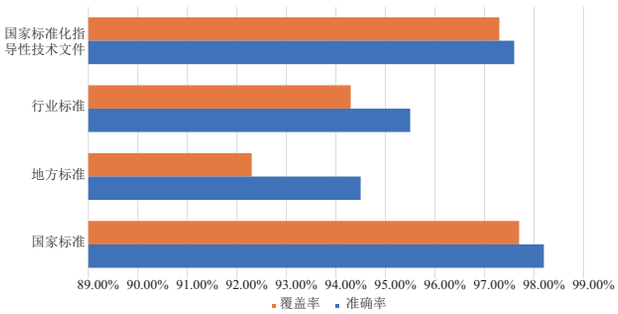


图 17　不同数据集下标准主干提取的准确率和覆盖率

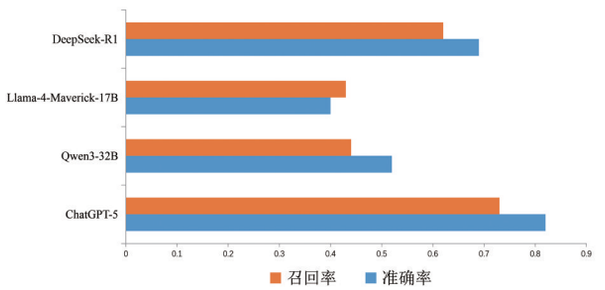


图 18　不同大语言模型(LLM)对标准文件检索的准确率和召回率

策略,从中医药标准文件中提取以下3类关键信息:①标准元数据信息,包括标准号、标准名称、发布单位、发布日期、适用范围等基础信息;②标准主干结构信息,包括规范性引用文件、术语与定义、技术要求、检验方法、标签标志等核心章节内容;③中医药专业领域信息,包括中药材名称、来源、性状鉴别、炮制方法、功效主治、用法用量、贮藏条件等专业内容。

为了验证系统的检索效果,本研究设计了4类共30个典型检索问题进行测试:①元数据查询类(8个),如“查询所有2020年后发布的艾灸相关国家标准”;②术语定义类(7个),如“根据标准,‘道地药材’的定义是什么?”;③技术要求类(8个),如“黄芪药材的性状鉴别要求有哪些?”;④跨标准检索类(7个),如“哪些标准涉及三七的质量控制指标?”。

测试过程中,检索准确率和召回率由系统基于文本匹配与语义相似度自动计算,并由2名中医药标准化领域专家进行抽样验证,以确保评估结果的可靠性。本研究基于不同的开源LLM对标准文件的信息检索结果进行了对比,定义检索准确率为LLM生成的结果真正符合提供的检索到的上下文(Top-10文档数据块)的数量所占的比例,检索召回率为LLM生成的结果占检索到的上下文中的所有相关信息的比例。图18展示了不同种类LLM对标准文件检索的准确率和召回率情况,综合来看,LLM结合RAG技术的方法能够较为准确、全面地检索出中医药标准的关键信息。

4　讨论

中医药标准领域长期缺乏高标准、高质量的数据库,其信息散在于出版物、全国标准数据库以及各地方、各团体标准数据库中,年代跨度长。随着中医药领域标准意识的不断深入,《中医药标准管理办法》《健康信息学—中医药数据集分类》等一系列法规和标准出台,近年来中医药领域的相关标准规范化程度大大提高[38]。然而,有部分年份较长的标准目前仍在中医药产业链中应用,进行替代升级仍需要时间。因此中医药标准领域的信息化需要兼顾新旧标准,形成标准间的网络、层级以及迭代关系,这有利于发现中医药标准体系中的不足之处从而提高标准升级的效率[39]。

本研究从中医药标准文件的结构化特征出发,设计并构建了一个中医药标准信息化存储和检索系统,选取了包含国家标准、地方标准、行业标准等400余份标准文件进行建库测试,结果显示采用本系统构建的中医药标准数据库,可支持主流LLM具备基于该系统的RAG语义检索能力。本研究验证所用的数据集包含400余条标准,截至投稿时,本文所述系统已经处理2 000余份各类各时期中医药标准,表现良好。

本研究存在一定局限性。首先,对于中医药标准中复杂配方表的深度语义解析尚不完善,例如“君、臣、佐、使”配伍关系、剂量换算等领域知识的自动抽取还需要结合中医药专业知识库进一步优化。其次,相关流程图目前以图像形式存储,尚未实现流程节点的自动识别与结构化表示。除此以外,古籍引用的版本溯源、异体字识别等问题也有待更深入的研究。对于中医药领域海量的历史遗留标准、团体标准、政策以及散见于各类出版物中的非标准规范性文件,系统的鲁棒性和泛化能力仍有待进一步验证。

本文的核心创新点有三:一是建立了包含印章去除、文件类型判别和并行解析的处理流程;二是采用基于有限状态机的标准主干提取方法;三是采用分层策略处理高度非结构化的领域专业内容。本研究所提出的文件格式转换算法、标准主干提取算法准确率均在95%以上,实现了从标准信息提取、入库、检索等全流程的需求。在后续的研究中,本研究团队将继续探索开源国产LLM在中医标准化数据库领域的应用,结合中医药领域知识图谱与多模态LLM,进一步深入

研究其在中医药标准信息检索中的表现。

参考文献：

[1] 李振吉，贺兴东，姜再增，等. 对中医药国际标准化建设的战略思考[J]. 世界中医药，2009，4(3)：167-171.

[2] 张盼，邓文萍，沈绍武，等. 中医临床大数据知识工程标准体系构建研究[J]. 时珍国医国药，2022，33(8)：2048-2050.

[3] 许吉，施毅，袁敏，等. 中医药信息基础标准框架构建探索[J]. 中国中医药信息杂志，2016，23(2)：8-10.

[4] 常凯，邓文萍. 中医药信息化标准体系框架研究[J]. 医学信息学杂志，2011，32(1)：14-18.

[5] 常凯. 中医药信息化标准体系构建研究[D]. 武汉：湖北中医药大学，2012：2-8.

[6] 李海燕. 中医临床信息标准体系框架与体系表的构建研究[D]. 北京：中国中医科学院，2012：5-8.

[7] 肖勇，田双桂，常凯，等. 中医药信息标准体系建设的思考[C]//中国中医药信息学会. 第四届中国中医药信息大会论文集. 北京：中国中医药信息学会，2017：205-208.

[8] 王松，朱佳卿. 中医药信息标准体系建设现状、问题与对策[C]//中国中医药信息学会. 第五届中国中医药信息大会——大数据标准化与智慧中医药论文集. 成都：中国中医药信息学会，2018：201-208.

[9] 张柯欣，石岩，曲超. 中医临床术语标准数据库的设计研究[J]. 大众标准化，2019(8)：27-29.

[10] 李楠. 中医儿科标准数据库建设研究[D]. 南京：南京中医药大学，2011：21-23.

[11] 崔健骅，黎耕，赵永恒. 中国古代天象记录数据库的设计与实现[J]. 天文研究与技术，2023，20(2)：179-188.

[12] 田翠华，张一平，胡志钢，等. PDF文档表格信息的识别与提取[J]. 厦门理工学院学报，2020，28(3)：70-76.

[13] 唐锐，邓建新，叶志兴，等. PDF文件的表格抽取研究综述[J]. 计算机应用与软件，2021，38(7)：1-7.

[14] 于丰畅，陆伟. 基于机器视觉的PDF学术文献结构识别[J]. 情报学报，2019，38(4)：384-390.

[15] 常凯，肖勇，邓文萍，等. “互联网+中医药”标准体系研究与思考[C]//中国中医药信息学会. 第五届中国中医药信息大会——大数据标准化与智慧中医药论文集. 成都：中国中医药信息学会，2018：193-198.

[16] 崔骥，许家佗. 人工智能信息技术在中医四诊现代化研究中的应用现状与展望[J]. 上海中医药杂志，2025，59(1)：7-12.

[17] 龚凡，刘波，沙航宇，等. 基于大语言模型检索增强的中医“药食同源”饮食建议生成方法[J]. 上海中医药杂志，2025，59(6)：1-8.

[18] Nguyen T T H，Jatowt A，Coustaty M，et al. Survey of post-OCR processing approaches[J]. ACM Comput Surv，2021，54(6)：124.

[19] Chauhan A. A review on various aspects of MongoDB databases[J]. Int J Eng Res，2019，8(5)：580-583.

[20] Shi W，Tan H，Kuang C，et al. Pangu DeepDiver：adaptive search intensity scaling via Open-Web reinforcement learning[PP/OL]. arXiv (2025-05-30) [2025-11-01]. https://arxiv. org/abs/2505.24332.

[21] Wang B，Xu C，Zhao X，et al. MinerU：An Open-Source Solution for Precise Document Content Extraction[PP/OL]. arXiv (2024-09-27)[2025-11-01].https://arxiv.org/abs/2409.18839.

[22] Opendatalab. MinerU at release-2.1.7 [EB/OL]. [2025-10-16]. https://github.com/opendatalab/MinerU/tree/release-2.1.7.

[23] OpenCV. Release OpenCV 4.10.0·opencv/opencv[EB/OL]. [2025-10-16].https://github.com/opencv/opencv/releases/tag/4.10.0.

[24] OpenCV. cv：：Mat Class Reference [EB/OL]. [2025-10-16]. https://docs.opencv.org/4.x/d3/d63/classcv_1_1Mat.html.

[25] OpenCV. Operations on arrays [EB/OL]. [2025-10-16]. https://docs.opencv.org/4.x/d2/de8/group__core__array.html.

[26] OpenCV. Image Thresholding [EB/OL]. [2025-10-16]. https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html.

[27] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need [C]//Curran Associates. Advances in Neural Information Processing Systems. NY：Curran Associates，2017：1-11.

[28] 曹启航，郑欣. 面向RAG的文本分块语义连贯性检测方法研究[J]. 自动化应用，2025，66(8)：124-128.

[29] Salemi A，Zamani H. Evaluating retrieval quality in retrieval-augmented generation[C]//Association for Computing Machinery. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. NY：Association for Computing Machinery，2024：2395-2400.

[30] 王葩，郑凯，王烽杰. RAG技术在大语言模型时代的应用——以RAGFlow与水利行业应用为例[J]. 科技视界，2025，15(23)：72-75.

[31] Klesel M，Wittmann H F. Retrieval-Augmented Generation (RAG)[J]. Bus Inf Syst Eng，2025，67(4)：551-561.

[32] 高雅奇. 基于大语言模型和RAG技术的高校知识库智能问答系统构建与评价[J]. 电脑知识与技术，2024，20(29)：18-20.

[33] Lewis P，Perez E，Piktus A，et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C]//Neural Information Processing Systems Foundation. Advances in Neural Information Processing Systems. NY：Curran Associates，2020：9459-9474.

[34] MongoDB. 如何为向量搜索创建索引字段[EB/OL]. [2025-10-17]. https://www. mongodb. com/zh-cn/docs/atlas/atlas-vector-search/vector-search-type/#std-label-avs-types-vector-search.

[35] Aumüller M，Bernhardsson E，Faithfull A. ANN-Benchmarks：A benchmarking tool for approximate nearest neighbor algorithms [J]. Inf Syst，2020，87：101374.

[36] DeepSeek-AI，Liu A，Feng B，et al. DeepSeek-V3 technical report [PP/OL]. arXiv (2024-12-27) [2025-11-01]. https://arxiv.org/abs/2412. 19437.

[37] Yang A，Li A，Yang B，et al.Qwen3 technical report[PP/OL].arXiv (2025-05-14)[2025-11-01]. https://arxiv.org/abs/2505.09388.

[38] 肖勇，李智一，沈绍武，等. 中医药信息标准数字化发展路径与策略思考[J]. 医学信息学杂志，2025，46(4)：40-44.

[39] 刘嘉艺，赖鸿皓，潘蓓，等. 生成式人工智能在中医药标准化研究中的应用前景与挑战[J]. 上海中医药杂志，2025，59(4)：1-7.

编辑：严　林

收稿日期：2025-11-26