

基于“智思体”的高校档案智能问答系统研发与应用研究

张 杰¹
ZHANG Jie

摘 要

针对高校档案服务智能化转型需求,文章构建了以“智思体”大模型为基础的智能问答系统,通过梳理其设计思路、关键技术与实现路径。与此同时,针对当前高校档案管理中检索效率低、反馈准确性不足等问题,结合实践场景引入 DeepSeek 等大语言模型,提出档案智能问答系统解决方案。依托山东建筑大学自主研发的“智思体”教育专用大模型,对校史、科研与教学等多类型档案数据进行整合,构建包含 2 万余实体与 30 余类关系的领域知识图谱,进一步结合 TextCNN+Transformer 融合架构的意图分类模型与 RAG 推理引擎,在提升检索效率的同时增强了结果可靠性。结果表明,该系统有助于推动档案服务由被动响应向主动供给转变,为高校档案管理的智能化实践提供了可复制的实现路径。

关键词

档案管理; 智能问答; 大语言模型; 知识图谱; 检索增强生成(RAG); 多模态服务

doi: 10.3969/j.issn.1672-9528.2026.07.046

0 引言

传统档案检索与利用模式存在效率低下、反馈准确性不足等问题,亟须引入新技术提升档案知识服务效能。“智思体”作为教育领域专业化的大模型平台,在数据处理、自然语言理解和知识生成方面表现突出,为档案管理的智能化转型提供了技术可能^[1]。山东建筑大学发布的“山大智思体”教育大模型平台,是山东省首个实现 DeepSeek 等最新主流大模型本地化部署的平台,且已全面应用于教学、科研、管理和服务等场景,为档案智能问答系统的开发奠定了坚实基础。

当前高校档案服务普遍存在重保管轻利用、智能化程度低、交互形式单一、知识关联薄弱等痛点。调研显示,高校师生单次档案查询平均耗时 8 min,“基建项目竣工图归档规范”等专业问题解答准确率不足 70%。国家档案局在《“十四五”全国档案事业发展规划》中明确提出,推动档案馆从静态存储中心向动态知识服务中心转型^[2]。但当前高校档案智能化建设仍面临数据治理能力不足、知识组织深度不够、交互服务单一三大瓶颈。

1 高校档案智能问答系统的现状分析

1.1 现有系统的特点与局限

现有档案智能系统存在不同特点,例如,西安交通大学档案馆与博物馆研发的“AI 智档”智能体具备三大核心功能:

1. 山东建筑大学 山东济南 250101

[基金项目] 2025 年山东省档案科技项目“融合知识图谱与智思体大模型的综合档案智能交互服务研究”支持(2025-18)

语境对话式检索、7×24 h 在线查档助手与人脸识别技术。进一步,西南大学“档案馆 AI 助手”可解答校史馆预约、新生档案归档、学生档案管理、档案利用预约和办展流程须知等常见问题。这些一站式服务平台大幅简化了档案查询和利用流程。

但现有系统仍存在明显局限。首先,系统功能较为单一,未针对高校特有的教师科研、教学档案实体进行深度关联,难以满足教师深度查询需求;其次,准确性有待提升,因缺乏高校档案领域专业语料训练,专业问题解答准确率普遍较低;最后,系统集成度低,难以实现档案信息的全面共享和联动,用户使用体验不友好。

1.2 技术架构与功能设计

从技术架构来看,现有档案智能问答系统多基于大语言模型通用框架,并结合档案领域特点进行定制化开发^[3]。需重点探索如何应用智能问答技术对用户的自然语言文本进行知识识别优化,同时结合信息融合技术对知识库、识别模型和语义环境等核心要素,提出具体建构策略。

在功能结构上,系统主要包含文本挖掘、知识表示、语义匹配和深度学习等模块。各模块围绕自然语言的处理流程展开意图解析、提取信息,并对信息进行组织输出,从而实现自动化问答支持。

2 基于智思体的档案智能问答系统设计

结合高校档案管理业务流程,本文以山东建筑大学“智思体”教育大模型为基础,设计面向档案场景的智能问答系统,推动档案信息由静态管理向动态管理转变,具体技术框

架如图 1 所示。

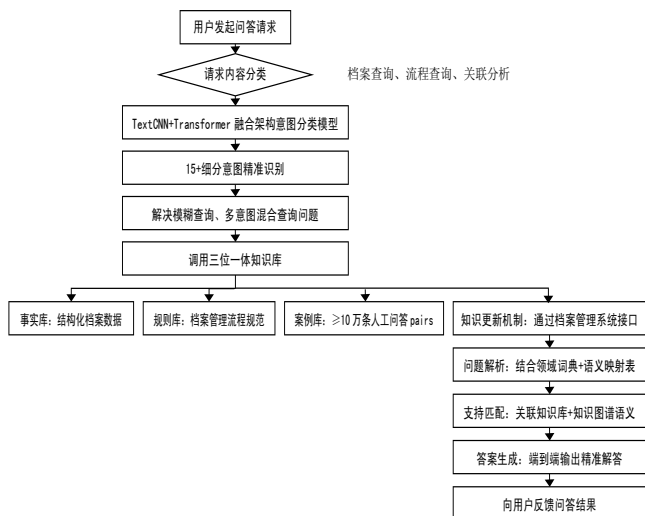


图 1 基于 TextCNN+Transformer 的档案智能问答技术架构

2.1 系统整体架构

在系统构架设计上，将功能划分为数据层、知识层、核心引擎层、交互层与应用层并采用分层架构设计，通过联动机制实现数据传递与功能调用，具体如图 2 所示。一是数据层，由 MySQL、Neo4j、Redis 及文件服务器组成的混合存储架构，以适配不同类型格式的档案数据存储与查询需求；二是知识层，主要包括知识图谱、知识库与专业术语词典库^[4]；三是核心引擎层，主要包括意图识别引擎、问答推理引擎、推荐引擎和数据治理引擎，其中意图识别引擎采用 TextCNN 与 Transformer 相融合的架构，问答推理引擎基于智思体与

RAG 技术进行搭建，以实现知识调用与结果生成^[5]；四是交互层，具有语音交互、文本交互、结果可视化展示等多模态交互能力，以适应不同使用场景下的交互需求；五是应用层，主要面向不同的使用场景，支持 PC 端及移动端（小程序 / APP）等访问，实现档案智能服务多终端接入。

2.2 技术基础与核心能力

“智思体”教育大模型为整个系统提供算力与模型支持，已完成对 DeepSeek、Qwen 等模型的本地化部署。该智能问答系统通过融合知识图谱与 RAG 技术，可对学籍、人事等六类档案数据关联查询。在检索效率方面，测试结果表明自然语言问答准确率不低于 88%，单次查询时间在 2 s 以内；在语音交互方面，可实现普通话、方言的语音识别，识别准确率分别不低于 85%，个性化推荐准确率达 80% 以上。上述结果表明，该系统能够满足高校档案信息查询的基本需求。

3 各模块功能详述

3.1 智能归档模块

针对高校档案中非结构化数据占比高的痛点^[6]，特别是基建图纸、合同文档等非结构化数据占比超过 90%。本模块针对多源异构档案资源的数据治理难题，制定包含 20 余项核心字段的《山东建筑大学档案元数据规范》，采用 Tesseract 开源引擎与百度 OCR 识别模型结合的识别方案，大幅提升图纸、扫描件、老照片等模糊或复杂版式资料的文字提取成功率。

在智能挖掘与知识沉淀层面，依托成熟的 BERT+CRF 实

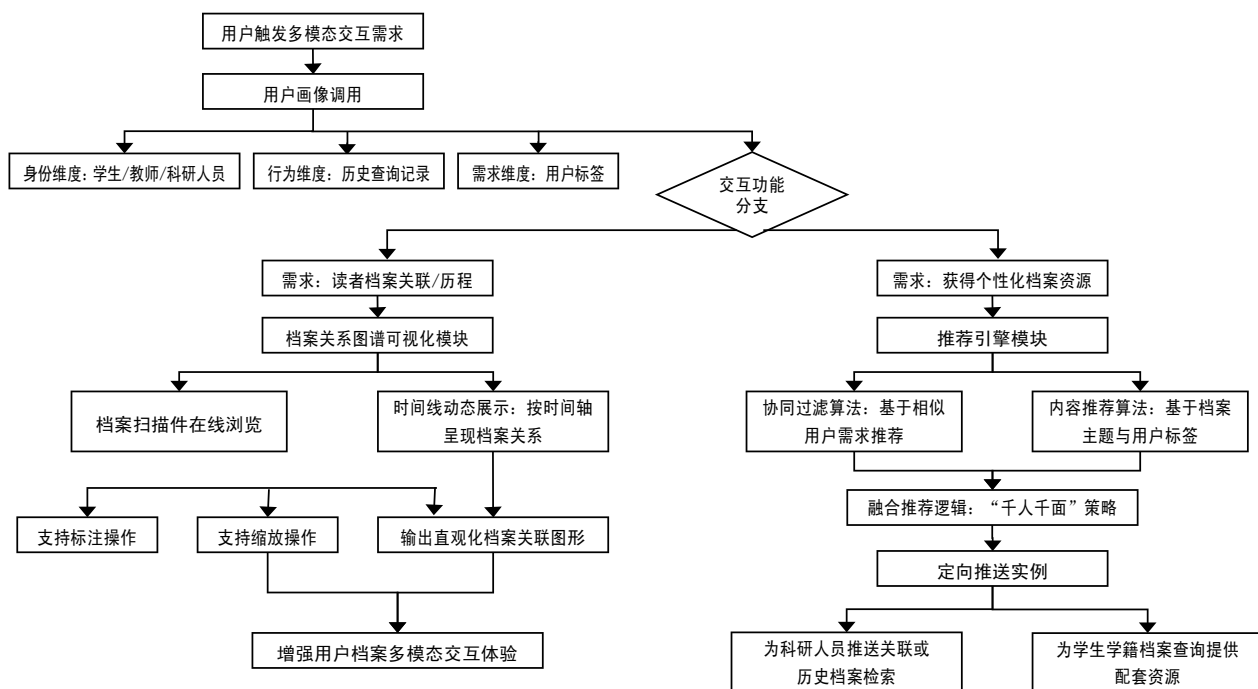


图 2 档案智能问答系统多模态交互架构

体识别模型，自动抓取名称、责任人等关键档案要素，综合识别准确率达到92%以上；再通过RE-BERT关系抽取算法，梳理并构建30余种典型业务关联链路，形成可动态更新的档案领域专属的知识图谱。同时，智能归档模块可以自动生成档案目录，并支持制度文件一键溯源并直达原文查看。

3.2 智能服务模块

构建“事实库+规则库+案例库”三位一体知识库。其中，事实库存储结构化元数据与知识图谱关系；规则库存储100余条档案管理流程规范；案例库共收集了10万余条人工问答样本用于模型微调。问答推理引擎基于智思体大模型与RAG技术，实现“问题解析-知识匹配-答案生成”的端到端处理。本系统提供7×24 h在线查档助手，支持名人档案深度解读，支持中英文双语精准问答，可链接相关音视频平台资源，为用户提供多模态档案体验。

3.3 智能编研模块

智能编研模块是系统的亮点之一，在西安交通大学“校缘拾忆”功能（利用人脸识别技术匹配校友历史影像）基础上，进一步开发基于ECharts的档案关系图谱可视化功能，支持实体拖拽、关系高亮与时间轴筛选（如“某教学楼1990年立项→1992年竣工→2020年改造”），图谱生成时间≤2 s，支持扫描件在线预览与标注，同时可以构建多维度知识网络，为校史研究和社会教育提供有力支持。具体如表1所示。

4 关键技术与实现挑战

“智思体”大模型的档案智能问答系统将涉及多项技术难题和现实挑战，具体表现在以下几方面：

（1）知识图谱构建与检索优化

围绕高校档案业务，搭建融合自然语言理解与知识图谱的档案知识库。通过数据清洗、实体识别与关系抽取，将其转化为结构化数据，并构建反映业务逻辑与语义关系网络以便不同场景调取分析。同时采用“检索优先、生成辅助”的实用问答策略^[7]，将档案领域知识图谱嵌入检索过程。同时能处理跨卷宗、多节点的复杂关联问题，在实际测试中，复杂关联问答准确率较传统检索提升11.8%。此外系统采用多文档联合关联解析技术，能自动对识别拆片段做统筹比对

与串联分析，重建档案信息的上下文业务逻辑。

（2）领域模型适配与文档整合能力提升

依托“智思体”教育专用大模型，我们开展深度定制优化：一是利用5万多条高校档案业务语料对模型进行微调；二是搭建了包含500多条专业术语的高校档案词典，强化了模型对归档范围、规则等专业内容的识别与理解能力。同时，在答案生成环节，引入模板约束机制，通过溯源原始档案为检索结果提供支撑，降低生成偏差，使输出内容与原始档案记录保持一致，提升结果的可解释性与规范性。

（3）安全管控与精准保障

数据安全是档案智能化建设的底线，系统采用本地化部署，将数据存储与模型推理限定在自有服务器集群中，搭配国产化适配模型，保证数据的安全性。在此基础上，构建包含身份认证、权限控制、数据脱敏与传输加密的安全机制，以保障数据使用过程的可控性。进一步建立自动同步、人工审核、更新发布的知识更新流程，使日常内容审核通过率达到95%；在问答输出中设置置信度约束，对低置信结果进行过滤与提示，提升系统输出的可靠性。

5 应用场景

5.1 多维应用场景

5.1.1 全天候智能查档服务

系统支持7×24 h在线智能查档服务，大幅简化传统线下签字、人工翻阅与排队等待的繁琐流程。在山东建筑大学3个试点学院中，用户单次查询操作一般不超过3步，零基础新手5 min就能熟练使用；系统全年稳定可用率达99.5%以上，师生使用满意度超过90%。参考天津商业大学“天商百事通”平台，进一步融合成绩单打印、教职工请假审批与财务报销指引等多项高频行政服务查询，有助于提升校园服务效率与资源利用水平。

5.1.2 校史文化挖掘与编研应用

档案系统升级能有助于校史挖掘、史料考证与专题编研。西安交通大学依托“校缘拾忆”模块，实现了校友历史学籍、毕业合影和荣誉档案等资料串联分析，从而提升档案资源的利用深度。

表 1 基于智思体的档案智能问答系统各模块功能对比

模块名称	核心功能	技术支持	应用场景
智能归档模块	语境对话式检索、一键溯源	多模型融合 OCR、BERT+CRF 实体识别、RE-BERT 关系抽取、知识图谱构建	档案目录生成、文件溯源
智能服务模块	24 h 查档助手、双语问答	三位一体知识库、TextCNN+Transformer 意图识别、智思体 +RAG 推理引擎、语音交互（普通话 / 方言）	查档服务、名人档案解读
智能编研模块	人脸识别、关联记忆	知识图谱、ECharts 可视化、PDF.js 预览	校史编研、文化展览

5.1.3 档案陈展宣教与服务延伸

四川省档案馆在档案展示与公共服务场景中,利用“档小河”AI智能体实现了语音实时交互、场馆自主导览和展品智能讲解等功能,拓展了档案利用渠道,提升档案资源在公共教育与文化传播中的应用价值。

5.2 应用场景实现

5.2.1 多模态交互能力拓展

为持续丰富系统的交互能力,可以扩展包括图像识别、语音问答及视频内容解析等多模态处理功能^[8],扩充更多方言识别技术,同时优化多轮对话机制,助力实现复杂问题分步拆解、解答等功能,以增强交互能力的连续性与适应性。

5.2.2 动态学习与模型优化机制

为提升系统动态学习能力,可以搭建长效动态学习机制,对查询日志、反馈信息及纠错记录进行周期性分析,利用分析数据调整优化模型参数,同时结合不同用户群体的特征提供差异化服务,提高系统响应的针对性。

5.2.3 搭建跨校跨馆协同服务网络

借鉴四川省档案馆红色档案资源知识库的推广经验,搭建高校档案馆联合智能服务协作网络,打通校际档案互查通道,针对特色资源进行互享并开通优质知识库,实现跨馆档案资源互补^[9]。

5.2.4 深化档案数据挖掘辅助管理决策

在基础查档功能之上,进一步开展存量档案数据挖掘工作,对校史政策文件、基建变迁档案、科研项目数据进行系统梳理与结构化分析,为学校制度的修订、优化资源配置、长效发展规划等提供数据支持与经验参考。

6 结论与实施建议

“山建大智思体”的档案智能问答系统是推动高校档案工作从传统应用向智能数字化运用的有效探索。该项目融合档案管理、计算机与人工智能安全等多领域技术,通过TextCNN与Transformer融合的意图分类模型、RAG智能推理引擎与多模态人性化交互模块等实现多项技术创新,破解了档案工作查档难和效率低等痛点。该系统已在山东建筑大学完成试点应用,减轻了档案管理人员日常检索、咨询答疑与史料整理压力,同时提升了师生查档效率,为高校开展档案智能化管理提供了一套可复制的技术路径与实践范本。

高校档案智能化建设应立足学校实际情况,通过科学规划分步有序推进。优先上线成熟落地与刚需高频的智能查档基础服务,解决师生查询需求,待基础架构与知识库搭建后,再逐步拓展校史编研、智能宣教等复杂高阶应用。同时,持续完善系统底层技术架构、扩充知识库存量、严控内容安

全精准度,让系统功能始终贴合高校动态发展需求。后续将坚持长期迭代、稳步升级,不断优化功能体验与拓展应用边界,真正让高校档案活用起来,切实服务师生、赋能管理与传承校史,促进学校高质量建设发展。

参考文献:

- [1] 中华人民共和国教育部.中国智慧教育白皮书[R].北京:教育部,2025.
- [2] 中共中央办公厅、国务院办公厅.“十四五”全国档案事业发展规划[Z/OL].(2021-06-08)[2026-06-16].<https://www.saac.gov.cn/daj/yaow/202106/899650c1b1ec4c0e9ad3c-2ca7310eca4.shtml>.
- [3] 教育部.教育数字化战略行动实施方案[Z].北京:教育部,2022.
- [4] 郑慧,刘思含.人工智能与档案开发利用:应用、愿景与进路[J].山西档案,2022(5):5-10.
- [5] 赵莹莹.融合知识库与语境信息的档案文本生成式问答模型构建研究[J].山西档案,2024(1):132-134.
- [6] 周林兴,林凯.大数据时代档案数据质量治理:因素、框架和路径[J].档案学研究,2023(2):111-119.
- [7] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[R]. San Francisco: OpenAI, 2018.
- [8] 丁咚,杨文昌.知识库问答场景中大语言模型私有化研究与应用实践[J].计算机应用与软件,2025,42(10):133-138.
- [9] 杨建梁,王一多,黄美雯,等.基于大语言模型的红色档案资源交互式知识发现研究:以《南方局党史资料大事记》为例[J].图书情报工作,2025,69(15):112-123.

【作者简介】

张杰(1970—),女,山东青岛人,本科,副研究馆员,研究方向:档案管理。

(收稿日期:2026-05-14 修回日期:2026-07-06)