

Research on the classification and retrieval technology of digital archives assisted by artificial intelligence

Yasen Lin

Suihua Open University (Suihua University for the aged, Suihua Community University), Suihua, Heilongjiang, 152000, China

Abstract

At the present stage of the rapid development of information technology, digital files are growing rapidly, like bamboo shoots after a spring rain, and how to efficiently classify and retrieve the massive digital files, has become a difficult problem to be solved. This paper mainly focuses on the artificial intelligence-assisted digital file classification and retrieval technology. In the paper, from optimizing the classification algorithm to make the file classification more accurate and efficient, to strengthening the application of natural language processing technology in retrieval, it puts forward related strategies that can improve the convenience and accuracy of retrieval. At the same time, it also puts forward the collaborative strategy of classification and retrieval technology, hoping to comprehensively improve the efficiency and quality of digital archives management, and then provide valuable reference for the in-depth research and practice in related fields.

Keywords

artificial intelligence; digital archives; classification technology; retrieval technology

人工智能辅助下的数字档案分类与检索技术研究

林亚森

绥化开放大学（绥化老年大学、绥化社区大学），中国·黑龙江 绥化 152000

摘要

现阶段信息技术在飞速地发展，数字档案则如雨后春笋一般迅速地增长，而如何高效地对海量数字档案进行分类与检索，已然成为亟待攻克的难题。本文主要聚焦于人工智能辅助的数字档案分类与检索技术，在文中从优化分类算法，让档案分类更精准高效，到强化自然语言处理技术在检索中的运用入手，提出能够提升检索便捷性与准确性的相关策略。同时还具备创新性地提出了分类与检索技术的协同策略，希望能够全方位地提升数字档案管理的效率与质量，进而为相关领域的深入研究与实践提供极具价值的参考。

关键词

人工智能；数字档案；分类技术；检索技术

1 引言

数字化时代当中，各行业产生的大量信息均以数字档案的形式进行保存，数字档案是重要的信息资源载体，要充分发挥其价值就要依赖有效的分类和检索。传统数字档案分类和检索方法多靠人工操作，效率低且准确性难保证。而人工智能技术，比如机器学习及自然语言处理等有很强的数据分析和处理能力，给数字档案分类与检索带来新机会。把人工智能技术用于数字档案管理领域能实现分类自动化和智能化，提高检索精准度和效率，更好满足用户对数字档案信息的需求，推动档案管理向现代化、智能化发展。

2 人工智能辅助数字档案分类技术策略

2.1 优化分类算法

在数字档案分类实践中，不同类型的档案数据有独特的特征分布，比如文本类档案有大量词汇组合和语义结构，图像类档案涉及色彩或形状等视觉元素。拿深度神经网络来说，对于文本档案分类可以用循环神经网络（RNN）及其变体长短时记忆网络（LSTM）。LSTM 能有效处理文本中的长距离依赖关系，通过门控机制选择性遗忘和记忆信息，准确捕捉文本语义特征。在图像档案分类中，卷积神经网络（CNN）效果很好，其卷积层和池化层能自动提取图像的局部和全局特征。迁移学习在实际应用中效果显著，例如某大型企业海量合同档案分类时，借助在通用文本分类数据集上预训练的 BERT 模型，只需少量企业合同数据微调就能快速建立高精度的合同分类模型，大幅缩短模型训练周期，

【作者简介】林亚森（1974-），男，中国黑龙江省绥化人，本科，档案馆员，从事档案管理研究。

进而提高分类效率^[1]。

2.2 提升数据质量

数据质量对分类模型性能起决定作用，在数据清洗阶段，对于数字档案中的数值型数据要检查有无异常值，比如财务档案中的金额字段，若出现负数或远超正常范围的值就需核实修正。对于文本数据，除了去除乱码和特殊符号还要进行词法和句法分析，纠正语法错误。在标注环节，以医疗档案分类标注为例，要组织专业医学人员依据国际疾病分类标准（ICD）等规范标注，为保证标注一致性，可以制定详细标注指南，明确不同类型医疗档案的标注规则。另外主动学习技术在实际操作中，模型会先从大量未标注的医疗档案中筛选出最具不确定性或代表性的数据，让标注人员标注，标注完后再用这些数据训练模型，使模型不断优化，减少对大规模标注数据的依赖，进一步降低标注成本。

2.3 增强分类模型的可解释性

在档案管理场景中，特别是涉及重要决策依据的档案分类可解释性很重要。以可视化技术在图像档案分类中的应用为例，通过热力图能直观看到卷积神经网络识别图像档案时，模型对图像不同区域的关注程度。比如在对历史文物图像档案分类时，热力图能清楚显示模型主要依据文物的关键部位（像独特的纹饰、造型特征等）进行分类决策。LIME 和 SHAP 等解释性模型在文本档案分类中作用很大，比如在对政府公文档案分类时，LIME 能针对某一公文被分类为特定类别的结果生成一段简短解释文本，指出文档中哪些关键词或短语对分类决策起关键作用，帮助档案管理人员理解模型为何这样分类，增强对模型分类结果的信任也便于在模型出现错误分类时排查改进。

3 人工智能辅助数字档案检索技术策略

3.1 深化自然语言处理技术应用

自然语言处理技术在数字档案检索领域不断深入发展，以满足日益复杂的检索需求。传统关键词匹配方式只能简单比对字面信息，难触及语义核心，而现在词向量模型是理解文本语义的关键工具。例如 Word2Vec 通过学习大量文本数据，把每个词语映射成多维向量，向量间距离反映词语语义相似度，这意味着即使检索词和档案文本用词不同，只要语义相近，也能精准匹配。在此基础上 BERT 模型借助 Transformer 架构的强大能力，能深入理解查询语句和档案文本的上下文语义，当用户输入“查找有关 20 世纪经济危机的资料”，BERT 能综合考虑“20 世纪”“经济危机”及各词间语义关联，在海量数字档案中找到相关内容。同时针对自然语言多义性，语义消歧技术利用档案文本语境信息，比如在一篇金融档案中能准确判断“杠杆”是金融术语而非物理工具，大大提高检索结果准确性，让检索不受语言表达多样性限制^[2]。

3.2 优化检索模型

传统检索模型面对海量数字档案时局限性逐渐显现。

向量空间模型虽能把文档和查询转化为向量计算相似度，但不能充分考虑文档间复杂关系；布尔模型基于逻辑运算，检索结果太绝对，容易遗漏潜在相关文档。机器学习和深度学习技术的应用带来改变，学习排序模型借助用户点击数据，分析哪些检索结果更受关注来自动调整排序算法，使后续检索结果更符合用户期望。例如在学术数字档案检索中，若用户频繁点击高被引论文链接，模型会提高此类文档排序权重。而双塔模型，如 DSSM，把用户查询和文档分别编码到同一语义空间，以新闻档案检索为例，当用户搜索“近期体育赛事新闻”，DSSM 能快速计算不同新闻文档与查询语义向量的相似度，实现高效排序。除此之外知识图谱技术的融入能够把档案中的人物、事件、时间等实体及关系结构化，为检索模型提供丰富语义知识，进一步提高检索准确性和智能化，让检索从简单文本匹配进入智能语义检索时代。

3.3 提升检索结果的相关性排序

检索结果的相关性排序直接影响用户对检索系统的满意度。要实现精准排序需综合考虑多方面因素，文档与查询的相似度是基础，通过余弦相似度等算法计算文本匹配程度。但仅靠这个不够，文档权威性也很关键，在政府文件档案检索中，来自核心部门发布的文件权威性高应优先展示。同时用户历史检索行为也不能忽视，系统可分析用户过往检索词及点击内容来挖掘其兴趣偏好，若用户常检索科技类档案且对前沿科研成果页面点击频繁，当再次检索相关内容时，模型会把此类文档排序提前。除此之外强化学习技术为动态优化排序策略带来活力，系统把每次检索当作一次“行动”，用户反馈（点击、收藏、忽略等）作为“奖励”信号。系统不断尝试不同排序方案并根据累计奖励调整策略去逐渐找到最适合用户的排序方式，使检索结果排序持续优化，始终为用户提供最具相关性的数字档案内容。

4 人工智能辅助数字档案分类与检索技术的协同策略

4.1 建立统一的数据标准与接口

4.1.1 数据标准制定

数据格式标准化是基础，在数字档案领域，XML（可扩展标记语言）因为有良好的结构化和自描述性，是存储数字档案数据的理想格式之一。它能清楚定义档案内容的各个元素及其关系以方便不同系统间的数据交换，例如对于人事档案可用 XML 标签明确区分姓名、出生日期、工作经历等字段，在分类系统和检索系统中都能被准确识别和处理。元数据定义要全面且精确，以学术档案为例，除了基本的标题、作者、发表日期等元数据还应包含学科分类、关键词、引用文献等信息。另外还要有统一的学科分类编码，比如采用中图分类法，能保证不同来源的学术档案在分类和检索时一致，并且数据编码规则也很重要，像对档案密级进行标准化编码，“机密”“秘密”“公开”等分别对应特定编码，在整个数字档案管理流程中保持统一，从而为分类和检索提

供清晰判断依据^[3]。

4.1.2 接口规范设计

接口协议很关键, RESTful API 因为简洁易理解及可扩展性强是数字档案系统间通信常用的接口协议。以 HTTP 为基础, 它通过定义 GET、POST、PUT、DELETE 等标准方法, 实现对数字档案资源的获取、创建、更新和删除操作。例如检索系统向分类系统请求特定分类下的档案列表时可用 GET 方法, 在 URL 中指定分类标识, 分类系统按协议返回符合要求的档案数据。同时数据传输格式也需规范, JSON (JavaScript Object Notation) 因为轻量级、容易解析常被用于数据传输, 它能把复杂的数字档案数据结构以键值对形式组织, 方便在不同系统间快速传输和处理。在传输过程中还要对数据进行必要校验, 比如用哈希算法生成数据摘要, 保证数据完整性, 防止传输中数据丢失或篡改, 保障分类与检索系统间数据交互准确可靠。

4.2 构建一体化的智能平台

4.2.1 系统架构设计

分层架构给智能平台提供清晰逻辑结构, 数据层负责安全存储海量数字档案数据及其元数据。采用分布式文件系统, 如 Ceph 就可实现数据高可靠存储, 把档案数据分散存储在多个节点上以避免单点故障。另外业务逻辑层集成分类和检索的核心算法, 例如分类算法中的随机森林算法和检索算法中的基于语义的向量空间模型都在这一层运算。通过微服务架构去把分类和检索功能拆分为独立服务模块, 每个微服务可独立开发、部署和扩展。比如检索需求增加时可单独对检索微服务进行水平扩展, 增加服务器节点, 提高检索性能。各微服务之间通过轻量级通信机制, 如消息队列等, 进行数据交互以保证系统灵活性和可维护性。

4.2.2 功能整合与优化

分类结果与检索功能紧密结合能大大提升用户体验, 用户检索数字档案时可根据分类标签筛选。例如在企业档案管理平台中, 用户检索财务相关档案时可通过“财务报表”“税务档案”等分类标签缩小检索范围, 快速找到所需档案。同时智能推荐功能基于用户行为分析实现, 利用协同过滤算法来分析用户的历史检索和浏览记录, 若发现多个用户都频繁访问某类档案, 当新用户访问相关内容时系统会自动推荐该类档案。如在高校档案管理系统中, 当学生用户查看了某专业的教学大纲档案后系统根据其他学生的行为去推荐该专业的课程安排、考试试卷等相关档案, 在为用户提供个性化服务的同时提高数字档案利用效率。

4.3 强化安全与隐私保护

4.3.1 数据加密技术应用

对称加密算法在数据存储安全中作用很大, AES 算法因高效性和安全性被广泛应用。在数字档案数据库中, 对档案内容加密存储, 例如对于医疗档案中的患者病历信息, 用 AES 算法加密后存储在数据库表中, 只有有正确密钥的授权用户才能解密查看。同时非对称加密保障数据传输安全, 在数字档案传输过程中, 如从分类系统向检索系统传输数据时用 RSA 算法对 AES 加密密钥进行加密传输。发送方用接收方的公钥对密钥加密, 接收方用自己的私钥解密获取密钥再用该密钥解密档案数据, 确保密钥传输安全, 从而防止数据被窃取或篡改。

4.3.2 访问控制与权限管理

用户认证机制采用多因素认证提高安全性。在数字档案管理系统登录时用户不仅要输入密码, 还需通过指纹识别或短信验证码二次验证, 例如政府档案管理系统中, 工作人员登录时, 除密码外需通过指纹识别确认身份, 防止密码泄露导致非法访问。权限管理根据用户角色严格分配权限, 在企业中普通员工可能只有对自己相关工作档案的只读权限, 而档案管理员有对所有档案的读写和管理权限。通过权限管理系统能够精确设置每个用户对不同类型档案的操作权限, 定期审查和更新用户权限以确保档案数据访问安全, 进而防止敏感档案信息被不当访问或泄露。

5 结语

人工智能技术给数字档案分类与检索带来创新解决方案, 通过优化分类算法、加强自然语言处理技术应用、构建协同策略等能明显提高数字档案管理效率和质量。不过目前人工智能在数字档案领域应用还面临一些挑战, 比如模型的可解释性、数据隐私与安全等问题。未来还需要进一步加强相关技术研究与创新, 完善数字档案管理标准和规范, 推动人工智能与数字档案管理深度融合, 让数字档案更好地服务社会 and 用户, 进一步发挥更大价值。

参考文献

- [1] 朱云辉. 人工智能技术在数字档案信息资源分类与检索中的应用[J]. 资源信息与工程, 2018, 33(06): 200-201+204.
- [2] 黄丽清. 浅析人工智能技术在数字档案信息资源分类与检索中的应用和发展[J]. 中国管理信息化, 2019, 22(14): 164-165.
- [3] 兰洋, 张玥. 人工智能技术在档案管理信息化中的应用[J]. 科技视界, 2024, 14(31): 108-111.