

人工智能在数字档案系统智能分类与检索中的应用

王国亮

北京城建设计发展集团股份有限公司, 北京 100037

DOI:10.61369/SE.2025050029

摘 要： 在数字化时代，档案信息呈现出爆炸式增长。各类组织和机构积累了海量的数字档案，涵盖文本、图像、音频、视频等多种格式。如何对这些档案进行高效分类与快速检索，成为档案管理工作面临的关键问题。传统的基于人工或简单规则的分类检索方法，已难以满足日益增长的数据处理需求。人工智能技术，如机器学习、自然语言处理、图像识别等，具有强大的数据处理和模式识别能力，能够自动从海量数据中提取特征、发现规律，为数字档案系统的智能分类与检索提供了创新解决方案。

关 键 词： 人工智能；数字档案系统；智能分类；有效应用

The Application of Artificial Intelligence in Intelligent Classification and Retrieval of Digital Archives Systems

Wang Guoliang

Beijing Urban Construction Design & Development Group Co., LTD., Beijing 100037

Abstract： In the digital age, archival information has shown explosive growth. All kinds of organizations and institutions have accumulated a vast amount of digital archives, covering various formats such as text, images, audio and video. How to efficiently classify and quickly retrieve these archives has become a key issue faced by archive management work. The traditional classification and retrieval methods based on manual or simple rules have become difficult to meet the growing demands of data processing. Artificial intelligence technologies, such as machine learning, natural language processing, and image recognition, possess powerful data processing and pattern recognition capabilities. They can automatically extract features and discover patterns from massive amounts of data, providing innovative solutions for the intelligent classification and retrieval of digital archive systems.

Keywords： artificial intelligence; digital archive system; intelligent classification; effective application

一、人工智能技术在数字档案系统中的应用原理

（一）机器学习技术

1. 监督学习算法在档案分类中的应用

监督学习是机器学习中的一种重要方法，它通过已知的训练数据来构建模型，并使用该模型对未知数据进行预测。在数字档案分类中，监督学习算法可以利用已标注类别的档案数据作为训练集，学习档案特征与类别之间的映射关系。例如，支持向量机（SVM）算法能够在高维空间中找到一个最优分类超平面，将不同类别的档案数据分开。通过对大量档案文本进行特征提取（如关键词、词频等），将这些特征作为 SVM 模型的输入，经过训练后，模型可以对新的档案文本进行准确分类^[1]。

（二）无监督学习算法在档案聚类中的应用

无监督学习算法用于处理没有标注类别的数据，其目的是发现数据中的内在结构和规律。在数字档案系统中，无监督学习算法可用于档案聚类。例如，K-Means 聚类算法通过将档案数据划分为 K 个簇，使得同一簇内的数据相似度较高，不同簇之间的数据相似度较低。通过对档案的内容特征（如文本主题、图像颜色特征等）进行分析，K-Means 算法可以自动将相似的档案聚成一

类，有助于档案管理人员对档案进行组织和管理，也方便用户在检索时从聚类结果中快速找到相关档案^[2]。

二、自然语言处理技术

（一）文本分类与标注

自然语言处理（NLP）技术在数字档案文本处理中发挥着重要作用。对于档案文本，NLP 技术可以实现自动分类与标注。首先，通过词法分析、句法分析等预处理步骤，对档案文本进行结构化处理，提取文本中的关键词、短语等特征^[3]。然后，利用基于深度学习的文本分类模型（如卷积神经网络 CNN、循环神经网络 RNN 等）对预处理后的文本进行分类。这些模型能够学习文本的语义特征，根据文本内容判断其所属类别，并为其标注相应的类别标签，大大提高了档案文本分类的效率和准确性。

（二）语义理解与检索

在档案检索方面，NLP 技术的语义理解能力使系统能够更好地理解用户的查询意图。传统的基于关键词匹配的检索方式往往存在检索结果不准确、相关性低等问题^[4]。而利用 NLP 技术，系统可以对用户输入的查询语句进行语义分析，理解其含义，然

后在档案库中进行基于语义的检索。例如，通过词向量模型（如 Word2Vec、GloVe 等）将文本中的词语映射到低维向量空间，计算查询语句与档案文本之间的语义相似度，返回与查询语义最相关的档案结果，从而提高检索的精准度和用户满意度。

（三）图像识别技术

1. 档案图像内容识别

对于包含图像的档案，如图书扫描件、照片档案等，图像识别技术可以对图像内容进行识别。基于深度学习的卷积神经网络在图像识别领域取得了显著成果^[6]。通过构建大规模的图像数据集，对卷积神经网络进行训练，使其能够学习到不同图像的特征模式。例如，在识别档案中的人物图像时，模型可以学习到人物的面部特征、姿态等特征，从而准确识别出图像中的人物身份。这对于档案图像的分类和检索具有重要意义，用户可以通过人物名称等关键词检索到包含相应人物的档案图像。

2. 图像分类与索引

图像识别技术还可用于档案图像的分类与索引。根据图像的内容特征（如场景、物体等），将图像分为不同的类别，并建立相应的索引。例如，对于历史建筑档案图像，可以根据建筑风格、年代等特征进行分类，为每类图像建立索引^[6]。当用户检索相关建筑档案时，系统可以通过索引快速定位到符合条件的图像档案，提高检索效率。同时，图像分类与索引也有助于档案管理人员对图像档案进行整理和管理，使档案资源的组织更加有序。

三、人工智能在数字档案系统智能分类与检索中的应用优势

（一）提高分类准确性与效率

传统的档案分类主要依赖人工操作，不仅工作量大，而且容易受到人为因素的影响，导致分类准确性不高。人工智能技术通过机器学习算法能够自动从海量档案数据中学习分类规则，对档案进行快速、准确的分类。例如，基于深度学习的文本分类模型可以处理复杂的文本文义，准确判断档案文本的类别，大大提高了分类的准确性。同时，人工智能的自动化处理能力使得档案分类能够在短时间内完成大量任务，相比人工分类，效率得到了极大提升。

（二）增强检索的精准性与速度

在数字档案检索方面，人工智能的语义理解和智能匹配能力使检索结果更加精准。传统检索方式往往仅基于关键词匹配，无法理解用户查询的真正意图，容易返回大量不相关的结果。而利用自然语言处理和机器学习技术，系统能够深入理解用户的查询语句，在档案库中进行基于语义的精准检索，提高检索结果的相关性。此外，人工智能技术通过构建高效的索引结构和优化检索算法，能够快速定位到用户所需的档案信息，大大缩短了检索时间，提升了检索速度^[7]。

（三）实现个性化服务

人工智能可以根据用户的历史检索记录、浏览行为等数据，分析用户的兴趣偏好和需求模式，为用户提供个性化的档案分类与检索服务。例如，对于经常关注某一领域档案的用户，系统可以自动将该领域的最新档案信息推荐给用户；在检索时，根据用户的习惯和偏好，调整检索结果的排序，优先展示用户可能感兴趣的档案。

这种个性化服务能够更好地满足用户的特定需求，提高用户对数字档案系统的满意度和使用频率。

四、人工智能在数字档案系统智能分类与检索中的应用案例分析

（一）某大型企业数字档案管理系统

1. 系统架构与技术应用

某大型企业构建了基于人工智能的数字档案管理系统。该系统采用分层架构设计，包括数据采集层、数据存储层、人工智能处理层和用户应用层。在数据采集层，通过 OCR 技术将纸质档案转换为数字格式，并采集各类电子档案数据。数据存储层采用分布式数据库，确保数据的安全存储和高效访问。人工智能处理层集成了机器学习、自然语言处理和图像识别等技术。在档案分类方面，利用深度学习的文本分类模型对文档型档案进行分类，通过图像识别技术对图像档案进行分类；在检索方面，借助自然语言处理技术实现语义检索^[8]。

2. 应用效果评估

该系统应用后，取得了显著的效果。档案分类的准确性从原来的 70% 提高到了 90% 以上，大大减少了人工分类的工作量。在检索方面，检索效率提升了 5 倍以上，用户能够快速准确地找到所需档案。同时，通过个性化推荐功能，用户对系统的满意度提高了 30%，有效提升了企业内部档案管理的效率和员工对档案资源的利用效率，为企业的决策支持和业务发展提供了有力保障。

（二）某国家档案馆数字化项目

1. 项目实施过程

某国家档案馆实施了数字化项目，引入人工智能技术提升档案管理水平。项目实施过程中，首先对馆藏的大量历史档案进行数字化处理，建立数字档案资源库。然后，针对档案分类与检索需求，采用人工智能技术进行系统优化。在分类方面，利用监督学习算法对档案进行分类训练，结合专家知识对分类结果进行校验和优化；在检索方面，构建知识图谱，将档案中的人物、事件、时间等信息进行关联，通过自然语言处理技术实现基于知识图谱的智能检索^[9]。

2. 取得的成果与经验

通过该项目的实施，国家档案馆实现了档案管理的智能化升级。档案分类的效率提高了 80%，检索准确率达到了 85% 以上。同时，知识图谱的构建使得档案之间的关联关系更加清晰，用户可以通过更灵活的方式进行检索，深入挖掘档案中的历史信息。该项目的成功实施为其他档案馆的数字化建设提供了宝贵经验，证明了人工智能技术在大规模档案管理中的可行性和有效性。

五、人工智能在数字档案系统应用中面临的挑战及应对策略

（一）数据质量与安全问题

1. 数据质量对人工智能应用的影响

高质量的数据是人工智能模型训练的基础。在数字档案系统中，数据质量问题可能包括数据缺失、错误、重复等。如果训练数据存在质量问题，会导致人工智能模型学习到错误的模式，从

而影响档案分类和检索的准确性。例如，在文本分类中，如果训练数据中的类别标注错误，模型在预测新数据时也会出现错误分类。因此，确保数字档案数据的质量对于人工智能技术的有效应用至关重要。

2. 数据安全风险与防范措施

数字档案包含大量敏感信息，在应用人工智能技术过程中，面临数据泄露、篡改等安全风险。人工智能模型在训练和运行过程中需要访问和处理大量数据，如果数据安全防护措施不到位，可能导致数据被恶意获取或篡改。为防范数据安全风险，应采取一系列措施，如加强数据加密，对传输和存储的数据进行加密处理，确保数据的保密性；建立严格的访问控制机制，限制只有授权人员能够访问敏感数据；采用数据备份和恢复技术，防止数据丢失；定期进行安全审计，及时发现和处理潜在的安全漏洞^[10]。

（二）算法可解释性与模型优化

1. 算法可解释性的重要性

在数字档案系统中，人工智能算法的决策过程往往较为复杂，缺乏可解释性。例如，深度学习模型通过多层神经网络进行数据处理和预测，其内部的决策机制难以直观理解。对于档案管理人员和用户来说，算法的不可解释性可能导致对结果的不信任。在档案分类中，如果无法解释为什么某份档案被分到特定类别，可能会影响档案管理的准确性和规范性。因此，提高人工智能算法的可解释性，使人们能够理解算法的决策依据，对于增强对人工智能技术的信任和应用具有重要意义。

2. 模型优化与持续改进方法

随着数字档案数据的不断变化和更新，人工智能模型需要持续优化以保持良好的性能。模型优化包括对模型结构的调整、参数的优化以及训练数据的更新等方面。例如，可以采用迁移学习技术，利用已有的相关领域模型进行初始化，然后在数字档案数据上进行微调，提高模型的训练效率和性能。同时，定期收集新的档案数据，对模型进行重新训练，以适应数据的动态变化。此外，通过性能评估指标（如准确率、召回率等）对模型进行监

测，及时发现模型存在的问题并进行改进。

（三）人才短缺与技术培训

1. 档案管理领域人工智能专业人才现状

目前，档案管理领域既懂档案业务又熟悉人工智能技术的专业人才短缺。传统的档案管理人员大多缺乏人工智能技术方面的知识和技能，难以有效推动人工智能在数字档案系统中的应用和维护。而人工智能专业人员对档案管理业务的了解相对较少，在开发和应用过程中可能无法充分满足档案管理的实际需求。这种人才结构的不平衡制约了人工智能技术在档案管理领域的深入应用。

2. 加强技术培训与人才培养的措施

为解决人才短缺问题，应加强对档案管理人员的人工智能技术培训。通过组织专业培训课程、开展学术讲座和研讨会等方式，提升档案管理人员的技术素养，使其掌握人工智能的基本原理、常用技术和应用方法。同时，高校和职业院校应加强相关专业的建设，培养既懂档案管理又具备人工智能技术能力的复合型人才。此外，档案管理部门还可以与人工智能企业合作，引进专业技术人才，共同推动数字档案系统的智能化建设。

六、结束语

本文通过对人工智能在数字档案系统智能分类与检索中的应用研究，得出以下结论：人工智能技术，如机器学习、自然语言处理和图像识别等，为数字档案系统的分类与检索提供了创新解决方案。通过自动学习档案数据的特征和规律，人工智能能够提高分类的准确性和效率，增强检索的精准性和速度，实现个性化服务，并降低人力成本。实际应用案例表明，人工智能技术在数字档案管理中具有显著的优势和应用价值。然而，在应用过程中也面临数据质量与安全、算法可解释性与模型优化、人才短缺等挑战，需要采取相应的应对策略加以解决。

参考文献

- [1] 陈红. 智能化视域下数字档案管理的挑战与应对策略 [J]. 办公室业务, 2024, (15): 163-165.
- [2] 黄鹏. 基于人工智能技术的数字档案系统建设 [J]. 信息记录材料, 2024, 25 (04): 180-182.
- [3] 王鸿翔. 人工智能时代下数字档案资源存储与长期保存的新模式与新路径探索 [J]. 山西档案, 2024, (03): 175-178.
- [4] 刘宁宁. 基于区块链和人工智能技术的数字档案系统建设 [J]. 数字技术与应用, 2022, 40 (12): 214-217.
- [5] 基于数字档案的人工智能辅助档案开放审核系统实现研究 [J]. 浙江档案, 2022, (10): 40-43.
- [6] 付文娟. 大数据背景下的高校学生档案大数据管理创新 [J]. 山西档案, 2025, (01): 116-118.
- [7] 鲍玲莉. 人工智能在专题档案收集工作中的应用 [J]. 档案天地, 2024, (10): 58-60.
- [8] 刘洪, 王文雯. 人工智能在档案管理中的应用研究 [J]. 兰台世界, 2024, (12): 35-38. DOI: 10.16565/j.cnki.1006-7744.2024.12.10.
- [9] 负疆鹏, 加小双, 王妍. 人工智能在我国档案管理中的应用现状与对策分析 [J]. 档案与建设, 2023, (02): 62-65.
- [10] 李龙凤. 智能化环境中档案业务的转型与创新 [J]. 兰台世界, 2019, (08): 96-98. DOI: 10.16565/j.cnki.1006-7744.2019.08.24.