

# 生成式人工智能与数字档案资源开发

何苗

四平市福利彩票发行服务中心，吉林省四平市，136000；

**摘要：**本文首先从“感知—理解—生成—应用”的基本逻辑出发，阐述了生成式人工智能在档案资源管理中的核心价值：通过多模态数字化获取原始档案信息，依托深度学习和知识图谱实现语义理解与结构化表达，再运用生成模型自动产出摘要、目录及交叉关联，并通过智能检索与个性化推荐将内容精准推送，实现档案价值的再创造。随后，文章围绕数据驱动的关键环节——数据收集与整合、清洗与预处理、挖掘与知识发现——剖析了高质量数据对生成式 AI 的支撑作用，强调了全流程的自动化与质量控制。接着，深入探讨了基于 Transformer、跨模态融合与自监督学习的算法核心，以及混合精度训练和推理优化等技术细节。最后，从需求分析与架构设计、模型微调、容器化部署，到运营监控与持续优化，提出了系统化的实践策略，为档案机构构建高效、可扩展的生成式 AI 平台提供了全面指导。

**关键词：**生成式人工智能；档案资源；数据驱动；系统部署

**DOI：**10.69979/3029-2700.26.01.055

## 引言

随着数字化浪潮的不断推进，传统档案管理面临着资源碎片化、检索效率低下和价值难以挖掘等挑战。生成式人工智能以其强大的文本与多模态生成能力，正在为档案资源开发利用注入新的活力。通过对档案进行高精度数字化、多维度语义理解和智能生成，生成式 AI 不仅能够提升档案的可访问性，还能在摘要提炼、知识发现和个性化服务等方面实现突破。然而，要将这一前沿技术真正落地，需要从理论逻辑、数据支撑、算法选型到系统架构和运维管理等多层次进行系统化规划与实施。本文旨在为档案机构提供一套完整的生成式 AI 应用框架，助力档案资源在新时代焕发更大价值。

## 1 生成式人工智能技术赋能档案资源开发利用的逻辑理路

### 1.1 基本逻辑分析

在档案资源管理中，生成式人工智能（Generative AI）的基本逻辑可概括为“感知—理解—生成—应用”四个环节。首先，通过多模态感知技术对原始档案进行高精度扫描和数字化，将纸质文字、图片、音视频等各类载体转为可计算的数据。接着，依托深度学习模型和自然语言处理算法，对海量档案文本进行语义理解与知识图谱构建，提炼出主题、人物、时间、地点等关键信息，并在此基础上形成档案主体结构的可视化表达。在生成阶段，生成式模型如 GPT 系列能够基于已有知识与结构，自动生成摘要、目录、标签，甚至是深度解读与

跨文献关联，实现对档案内部价值的再创造。最后，通过智能检索、自动问答和个性化推荐，将生成内容以多样化形式推送给研究人员、社会公众或决策者，极大提升档案利用效率与价值转化。此“感知—理解—生成—应用”的链条，既确保了档案资源的完整与可追溯性，又赋予了档案管理新的智能驱动能力，为传统档案工作注入创新活力。

### 1.2 数据驱动

生成式人工智能对档案资源的依赖，本质上是对数据能力的深度挖掘和运用。档案数据驱动的核心在于对原始信息的全面收集、严格清洗、深度预处理，以及高效的数据挖掘与知识发现。只有充分发挥档案数据的价值，才能为生成式模型提供可靠支撑，实现高质量的文本生成与多模态融合。数据驱动不仅是技术层面的需求，也是流程重塑的契机：在信息采集、存储、标注、管理等各环节融入数据质量控制与自动化管道，构建一条端到端的数据智能流，为后续的算法优化和应用创新提供坚实基础。

#### 1.2.1 数据收集与整合

档案资源分散于不同载体和部门，数据格式多样且质量参差不齐。生成式人工智能赋能档案的第一步，是建立统一、高效的数据采集与整合平台。一方面，利用光学字符识别（OCR）、智能摄像和语音转写等技术，将历史手稿、胶片、录音录像等传统载体数字化；另一方面，通过 API、爬虫与数据共享协议，将各部门、各系统的电子档案、业务数据和元数据集中汇聚。过程中，

针对语言种类、格式规范、字符编码等差异，需设计灵活的数据标准与映射规则，确保各类档案之间的互联互通。集成后的多源异构数据，不仅为生成式模型提供了丰富原料，也为后续的清洗与挖掘打下了坚实基础。此环节的目标是实现“全局可视、统一管理、按需调度”的档案资源生态。

### 1.2.2 数据清洗与预处理

在完成原始数据整合后，必须对数据进行清洗与预处理，以保障生成式人工智能模型输入的准确性和高效性。清洗阶段主要包括去除噪声与重复、纠正字符识别错误、填补缺失字段等操作；预处理阶段则涉及分词、词性标注、命名实体识别以及图像增强与分割等多项技术。针对档案中典型的手写体、老照片、录音噪声等复杂场景，还需定制化算法，如基于生成对抗网络（GAN）的图像修复或基于自监督学习的语音降噪。预处理输出不仅要保留原始信息的完整性，还要对文本、图像、音频等多模态数据进行特征提取与向量化，为后续的知识发现与模型训练提供优质特征。高效、准确的数据清洗与预处理，是生成式人工智能可靠运行的关键保障。

### 1.2.3 数据挖掘与知识发现

数据挖掘与知识发现环节，旨在从海量档案中揭示潜在规律与丰富关联。通过主题模型（如 LDA）、语义网络分析以及图数据库技术，能够识别档案文本中的主题分布、时间演变与人物关系；利用知识图谱，将分散在不同档案之间的事件、对象与概念形成结构化网络，支持跨文献的综合检索与语义推理。结合深度学习的表征学习能力，系统能够在无监督或弱监督环境下，自动发现档案资源间的潜在联系与知识空白，促进历史研究、决策参考甚至文化传播的新视角。生成式人工智能则基于这些挖掘成果，实现智能摘要、情境再现与虚拟对话，为不同用户群体提供个性化、可解释的档案知识服务。

## 1.3 算法核心

在生成式人工智能技术中，算法核心决定了系统的智能水平与应用边界。对于档案资源的处理，主流算法包括基于 Transformer 的预训练语言模型、多模态融合模型以及自监督学习框架。预训练语言模型通过海量文本预学习具备强大上下文理解能力，再通过在档案领域的微调，实现对专业术语、历史文本风格的精准生成；多模态融合模型则将文本、图像与音频特征在统一向量空间内对齐，支持跨模态检索与生成，如根据历史图片描述档案背景；自监督学习框架则借助自动构建标签的方式，降低人工标注成本，提高模型对档案中稀有材料和低频词的识别水平。除了模型结构，算法的优化策略

（如混合精度训练、分布式并行等）和推理加速（如量化、裁剪）同样关键，为档案馆等资源受限的环境提供了可行的技术方案。

## 1.4 应用导向

生成式人工智能技术在档案资源开发利用中的最终目的是服务多元化应用场景。对科研人员，可提供智能摘要、热点主题分析及虚拟助手问答；对广大公众，可打造沉浸式数字展览、互动式历史教育平台；对决策者，可生成基于档案数据的政策白皮书和风险评估报告；对档案管理者，则可实现自动目录生成、元数据完善及智能审查辅助。每一种应用，都基于前述逻辑与技术，实现从“数据—模型—服务”的闭环。通过持续迭代和场景验证，生成式人工智能与档案资源的深度融合将不断拓展，实现历史价值的数字再现与社会效益的最大化。

## 2 生成式人工智能技术的实践策略

### 2.1 部署前的需求分析与架构设计

任何技术落地均始于充分的需求调研与系统架构规划。首先，应与档案馆管理者、专业研究人员和终端用户开展多轮访谈与问卷调研，明确档案资源的核心价值、使用场景及业务痛点。例如，是侧重于大规模全文检索、还是更注重智能摘要与知识发现？用户群体是学术研究者、普通公众，还是内部管理人员？根据需求差异，才能确定系统需要支持的输入格式（文本、图像、音频）、响应时延、并发访问量、安全与隐私保护等级等关键指标。

在架构设计层面，应优先采用“分层解耦、模块化开发”的原则：数据层负责异构档案的采集、整合与预处理；算法层承载各类生成式模型的训练与推理；服务层提供检索、问答、可视化等前端接口；平台支撑层则涵盖容器化部署、弹性伸缩、日志与监控等基础设施。各层之间通过标准化 API 对接，既保证松耦合，也方便后续模块的更新与替换。此外，还需充分考虑灾备方案和数据安全：对敏感档案采用访问控制、加密存储与审计日志；对系统关键节点配置高可用集群和自动故障切换，确保档案业务连续性与系统稳定性。通过如此完整的需求与架构设计，才能为后续模型训练、系统集成和运营维护打下坚实基础。

### 2.2 模型选择与培训微调流程

在明确业务需求和系统架构后，接下来的核心是模型选型与训练流程设计。根据前期需求调研，如果侧重文本生成与智能摘要，可选用基于 Transformer 的预

训练语言模型（如 GPT、T5 或国内等效模型）；若需要同时处理图像、音频等多模态档案资源，则需引入 CLIP、BEiT 等视觉文本联合预训练模型，或专门的多模态融合框架。在模型训练前，需准备高质量的领域数据集：筛选具代表性的档案文本、历史图片、影音片段，并通过人工打标构建微调语料与任务样本。

针对不同任务（摘要生成、分类、关联检索等），规划分阶段训练策略：先在大规模通用语料上做基础预训练，再在档案领域语料上进行微调；必要时可采用少量监督学习和自监督学习混合的方式，以提升对低频、冷门历史素材的理解能力。训练过程中，应关注模型的收敛速度、生成质量与推理效率，并通过自动化评估指标（如 ROUGE、BLEU、Embedding 相似度）与人工评审相结合，确保输出符合档案文本的严谨性与学术规范。最后，将训练好的模型封装为微服务，支持 RESTful 或 gRPC 调用，为上层应用提供高效的生成与推理能力。

## 2.3 系统集成与可扩展部署

模型训练完成后，需将其与档案管理系统无缝集成，并在生产环境中进行可扩展部署。首先，在开发环境中，利用 Docker 或 Kubernetes 将各模型服务、数据管道和 API 网关进行容器化，确保不同模块的版本隔离与资源调度；同时，通过 CI/CD 流水线实现代码、配置和模型版本的一键发布与回滚。上线前，应在预生产环境进行压力测试和安全扫描，验证系统在高并发检索、生成任务下的响应时延、内存与 GPU 占用情况，并修复可能的性能瓶颈。

在生产部署时，根据访问量与使用模式，设置自动扩缩容规则：如当请求量超过阈值时，自动添加更多推理计算节点；当低峰期时，释放资源以节约成本。对于敏感档案，可在私有云或本地机房部署，同时对外服务则可利用公共云加速内容分发。技术团队还需制定完善的服务发现、配置管理和权限认证方案，确保各组件之间的通信安全可靠。通过上述集成与部署策略，既让生成式 AI 功能平滑嵌入现有档案系统，也保证系统具有良好的伸缩性与运维可控性。

## 2.4 运营监控与持续优化

系统上线后，持续的监控与优化是保障长期价值的重要环节。要构建全面的监控体系，包括模型推理延时、错误率、资源使用率等指标，以及业务指标如用户点击率、满意度、搜索命中率等。通过可视化监控看板，运营团队能够及时发现系统瓶颈、模型漂移或用户体验下降。同时，应定期采集用户反馈和使用日志，分析生成

内容的准确性与实用性，识别模型在特定领域或题材上的弱点。

在监控基础上，实行“迭代—评估—优化”闭环：当发现问题时，调整模型微调数据，或尝试新的架构与算法（如引入 Retrieval-Augmented Generation 以提升事实准确性）；对系统配置进行性能调优，如 GPU 资源分配、并发策略优化等；同时不断丰富档案领域语料库，引入更多新近数字化档案。为了应对技术与业务的双重演进，还需制定定期的更新计划，确保生成式人工智能系统始终保持行业领先水平。通过科学的运营监控与持续优化，才能最大化生成式 AI 在档案资源开发利用中的实际效益。

## 3 结语

生成式人工智能在档案资源管理中的应用，不仅是技术层面的革新，更推动了档案学与信息学的深度融合。通过构建“数据—模型—服务”的闭环，档案机构能够实现从原始信息采集到智能内容生成的全流程升级。展望未来，随着模型算法的不断演进和数据生态的持续完善，生成式 AI 将在档案主题挖掘、知识图谱拓展、虚拟展览及决策支持等领域发挥更大作用。持续的监控与优化、跨学科协同与标准化建设，将成为驱动这一变革持续落地的关键因素。档案事业的数字化转型必将借助生成式人工智能，实现历史价值的创新再现与社会效益的最大化。

## 参考文献

- [1] 仇壮丽,石敏.生成式AI赋能档案知识服务创新路径——基于SECI模型视角[J/OL].浙江档案,2025,(05):1-10[2025-05-26].<https://doi.org/10.16033/j.cnki.33-1055/g2.20250430.001>.
- [2] 刘聪.生成式人工智能与数字档案资源开发[J/OL].图书馆,1-8[2025-05-26].<http://kns.cnki.net/kcm/detail/43.1031.G2.20250424.1940.002.html>.
- [3] 杨冬权.“深度求索”(DeepSeek)带给档案人的深度思索[J].四川档案,2025,(02):4-5.
- [4] 戴旸,唐亮亮,李佳轩.人工智能生成内容(AIGC)驱动下的智慧档案建设研究:应用场景、风险挑战与解决路径[J].档案学研究,2025,(02):111-120.DOI:10.16065/j.cnki.issn1002-1620.2025.02.013.
- [5] 王玉珏,樊静雅,温翰英.人工智能生成合成内容的可信存档策略研究——基于对电子档案“四性”的思考[J].北京档案,2025,(04):22-29.