ICS
CCS

T/GXDSL

团　　体　　标　　准

T/GXDSL　— 2026

档案智能化分类与精准检索技术应用指南

Application Guide for Intelligent Classification and Precise Retrieval Technology of Archives

（工作组讨论稿）

（本草案完成时间：2026-01-22）

2026 - - 发布　　2026 - - 实施

广西电子商务企业联合会　发 布

目　　次

前　言

本文件依据GB/T 1.1-2020 《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西产学研科学研究院提出并宣贯。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

档案智能化分类与精准检索技术应用指南

1 引言

随着数字中国战略的深入推进和各行各业数字化转型的加速，档案资源作为国家与社会的核心信息资产，其形态正从传统实体向海量、多态、异构的数字化档案迅速演变。面对急剧增长的档案数据规模和日益复杂的利用需求，传统依赖人工和经验的管理与检索方式已难以满足高效、精准、智能的档案服务要求。人工智能、自然语言处理、机器学习等新一代信息技术的发展，为档案管理现代化提供了革命性工具。为推动智能技术在档案管理领域的规范化、深度化应用，提升档案分类的科学性、检索的精准性和服务的智能化水平，有效挖掘档案数据价值，特制定本指南。本指南聚焦档案智能化分类与精准检索的技术应用，对系统架构、关键算法、数据处理、应用实现及安全要求提出指导性规范，旨在为各级各类档案机构及相关技术服务机构开展智能化建设提供科学、可行的技术路径与实践依据。本指南由广西产学研科学研究院联合档案管理机构、高校及科技企业共同研制。

2 范围

本指南规定了档案智能化分类与精准检索技术应用的系统架构、数据处理要求、关键技术方法、应用实现模式、性能指标及安全管理要求。本指南适用于各级国家综合档案馆、专业档案馆、部门档案馆以及企业事业单位档案机构，在文书档案、科技档案、专业档案等各类档案数字化管理场景中，应用人工智能技术进行档案智能分类、著录、标引与精准检索的系统规划、设计、开发、部署与评估。其他信息管理机构对结构化与非结构化文档进行智能处理时可参照执行。

3 规范性引用文件

下列文件对于本指南的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本指南。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本规范。

GB/T 18894-2016 电子文件归档与电子档案管理规范

DA/T 18-2022 档案著录规则

DA/T 31-2017 纸质档案数字化规范

DA/T 46-2009 档案数字化工作规范

DA/T 58-2014 电子档案管理基本术语

GB/T 39784-2021 电子档案管理系统通用功能要求

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 35273-2020 信息安全技术 个人信息安全规范

《中华人民共和国档案法》（2020 年修订）

《国家档案馆档案开放利用办法》（国家档案局令第 19 号）

4 术语和定义

4.1 档案智能化分类：指利用自然语言处理、机器学习、深度学习等人工智能技术，自动或半自动地识别、分析和判断档案内容、形式及上下文特征，并依据预定义的分类体系（如档案分类方案、主题词表、职能分类等）将其归入相应类目的过程。

4.2 档案精准检索：指基于对用户检索意图的深度理解，运用语义分析、知识图谱、相关性排序等智能技术，从海量档案资源中快速、准确地查找并返回与用户需求高度相关档案信息的过程。

4.3 档案数据预处理：指在应用智能技术前，对数字化档案原文、图像、音视频等数据进行格式转换、文字识别、噪声去除、文本清洗、分词等操作，以形成适合机器处理的标准数据形式的过程。

4.4 训练数据集：指用于训练和优化智能分类或检索模型的、已由人工进行正确分类或标注的档案数据集合。

4.5 档案知识图谱：指以结构化形式描述档案实体（如人物、机构、事件、地点、时间等）及其之间语义关系的知识库，是实现语义关联和智能推理检索的核心组件。

4.6 置信度：指智能分类模型对档案归类结果的确信程度，通常以概率值表示。

5 总则

档案智能化分类与精准检索技术应用应遵循“辅助人工、提升效能、确保准确、保障安全”的基本原则。智能技术是辅助档案专业人员提升工作效率和质量的有力工具，而非完全替代人工判断，尤其在涉及复杂价值鉴定、敏感信息识别等领域需保留人工审核环节。系统设计与实施应以提升档案管理效能和利用服务水平为根本目标，确保技术方案与业务需求深度融合。必须高度重视智能处理结果的准确性，建立有效的质量控制和纠错机制。全过程应严格遵守国家档案管理、网络安全、数据安全及个人信息保护等相关法律法规，确保档案数据的完整性、安全性、可用性和保密性。系统建设应坚持开放、兼容、可扩展的技术路线，支持与现有档案管理系统平滑对接。

6 系统总体架构与技术要求

档案智能化应用系统宜采用分层解耦的微服务架构，主要包括数据资源层、智能引擎层、应用服务层和用户交互层。数据资源层负责存储和管理结构化目录数据、非结构化全文数据、多媒体档案数据以及知识图谱等，应支持海量数据的高效存储与访问，推荐采用分布式文件系统和关系型与非关系型数据库混合架构。智能引擎层是系统的核心，封装了分类模型、检索算法、自然语言处理工具、OCR 识别服务等各类智能算法模块，应以标准化 API 接口方式提供服务。应用服务层封装具体的档案业务逻辑，如智能分类任务调度、检索请求处理、用户权限管理、日志审计等。用户交互层为档案管理员和利用者提供 Web 端、移动端等多渠道交互界面。系统应具备高可用性和可扩展性，关键服务集群化部署，单点故障不应导致核心服务中断，系统整体可用性不低于 99.5%。平均无故障时间（MTBF）应大于 10000 小时。系统响应性能需满足：简单检索请求平均响应时间不超过 2 秒，复杂语义检索或跨库联合检索平均响应时间不超过 5 秒。系统应支持至少 100 个并发用户的在线智能检索请求。

7 档案数据预处理与特征提取要求

高质量的数据预处理是智能应用成功的基础。对于图像类档案，应首先采用光学字符识别技术（OCR）将其转换为文本。中文 OCR 的字符识别准确率对于印刷体应不低于 99.5%，对于清晰度较高的手写体应不低于 85%。OCR 后需进行文本清洗，包括纠正识别错误字符、去除无关符号、分段分句等。对于已数

字化的文本档案，需进行格式标准化处理，统一编码为UTF-8。预处理后的文本需进行分词和词性标注，推荐使用专业领域词典以提高分词准确性。在此基础上，需进行深入的特征提取，为智能模型提供输入。特征应包括但不限于：文本内容特征，通过词袋模型、TF-IDF、词向量（如Word2Vec、BERT等预训练模型生成的向量）表示；元数据特征，如文件标题、责任者、形成日期、文种等结构化字段；版面与格式特征，如公文版头、发文字号位置、印章区域等视觉信息；上下文特征，如该档案在案卷或全宗中的位置、前后档案的关联信息等。多媒体档案（如照片、录音、录像）应提取其元数据、文字解说信息，并可利用图像识别、语音识别技术提取关键视觉或听觉特征标签。

8 智能化分类技术应用指南

智能分类主要包括自动归类、智能标引和主题提取。自动归类是指根据档案内容及特征，自动将其归入预先设定的分类体系（如《中国档案分类法》或机构自定义分类方案）的相应类目。推荐采用监督学习方法，如支持向量机、深度学习文本分类模型（如TextCNN、BERT等）。首先需构建高质量的标注训练集，训练集应覆盖所有目标类目，每个类目的样本量原则上不少于500份，且样本分布应尽可能均衡。模型训练完成后，应在独立的测试集上进行评估，宏观平均准确率（Macro-F1）应不低于0.90，对于核心或高频类目，准确率应力争达到0.95以上。系统应输出分类结果及置信度，对于置信度低于设定阈值（如0.75）的档案，应自动标记为“存疑”，交由档案人员审核确认。智能标引是指自动从档案内容中提取关键主题词或关键词，并映射到规范的档案主题词表（如《中国档案主题词表》）。可结合基于规则的方法（如词频统计、位置权重）和基于神经网络序列标注的方法（如BiLSTM-CRF）进行实体识别和关键词抽取。主题提取旨在自动概括档案的核心内容，生成简明的摘要。系统应支持多级分类和复合分类，并能记录分类的依据（如触发分类的关键特征），确保过程可追溯、可解释。

9 精准检索技术应用指南

精准检索旨在超越传统的基于关键词的字面匹配，实现基于语义的深度检索。系统应支持多种检索模式：关键词检索，作为基础功能，应支持布尔逻辑、短语检索、模糊匹配等；语义检索，核心是理解查询语句的真实意图，通过查询扩展、语义向量相似度计算（如通过Sentence-BERT计算查询与档案的语义相似度）返回相关结果；关联检索，基于构建的档案知识图谱，发现并推荐与检索目标相关联的人物、事件、地点等其他档案实体；跨媒体检索，支持“以图查档”、“以音查档”等。检索系统应构建

高效的索引机制，对文本内容、元数据、特征向量、知识图谱关系等分别建立倒排索引或向量索引，以实现毫秒级响应。相关性排序算法至关重要，应采用融合多种特征的排序学习模型，综合考虑文本相关性、语义相似度、档案价值权重、利用热度、时间新鲜度等因素进行综合打分与排序。检索结果应提供清晰的排序列表，并可按照相关度、时间、分类等多种方式灵活筛选和排序。系统应提供检索词建议、相关搜索推荐、检索结果聚类分析等辅助功能，提升用户体验。检索命中结果的查准率（Precision@10）在标准测试集上应不低于 0.85，查全率（Recall）在可控范围内应持续优化。

10 应用实现与部署要求

智能分类功能可应用于档案接收环节的自动预归类、数字化加工后的批量自动著录标引、存量档案数据的智能整理与深度编目等场景。精准检索功能应无缝集成到档案利用服务平台，面向内部管理人员和社会公众提供高效服务。系统部署可采用本地化部署、私有云部署或与可信公有云服务结合的混合部署模式。涉及国家秘密、工作秘密和个人敏感信息的档案，其智能处理与检索系统必须实行完全的物理隔离或逻辑强隔离的本地化部署，并符合分级保护或等级保护相关要求。系统应提供完善的管理后台，允许档案管理员对分类体系、词表、检索模型参数、权限规则进行配置和管理。系统需具备模型更新和迭代能力，能够定期利用新的标注数据对模型进行增量训练或重新训练，以保持和提升其性能。应建立人机协同机制，设置便捷的人工干预和反馈入口，将人工对智能处理结果的纠正信息作为新的训练数据，持续优化模型。

11 安全、隐私与伦理要求

安全是智能技术应用的生命线。系统必须满足 GB/T 22239-2019 中相应安全等级的要求。在数据安全方面，训练数据的采集、存储和使用需获得合法授权，严禁使用未授权的档案数据进行模型训练。处理包含个人信息、商业秘密的档案时，应采取数据脱敏、去标识化等技术措施，符合 GB/T 35273-2020 的要求。在算法安全方面，应关注算法的可解释性与公平性，避免因训练数据偏差导致对特定群体、特定主题档案的分类歧视或检索偏见。应建立算法审计机制，定期评估算法决策的合理性与公平性。系统所有操作均应记录详尽的日志，包括模型调用、分类检索行为、数据访问记录等，日志保存时间不少于 6 个月，以满足审计和追溯要求。应制定应急预案，应对模型失效、检索结果异常、系统被攻击等安全事件。

12 实施、评估与运维

实施前应进行详细的业务需求分析、数据现状评估和技术可行性论证。制定分阶段实施方案，可先选取部分类别或部分全宗的档案开展试点，验证效果后再逐步推广。系统正式上线前，必须进行严格的第三方测试与评估，评估指标至少包括：分类准确率、检索查准率与查全率、系统响应时间、并发处理能力、资源占用率等。应建立持续的运维保障体系，包括日常监控、性能调优、模型维护、数据备份与恢复等。定期（如每年一次）对系统应用效果进行业务评估，评估其对档案管理效率、查档利用满意度提升的实际贡献度。加强对档案业务人员和技术人员的培训，使其能够理解、管理和有效利用智能系统。

13 附则

13.1 本指南自发布之日起实施。

13.2 各相关单位在开展档案智能化分类与精准检索系统建设时，可参照本指南执行。

13.3 本指南所引用的国家标准和行业标准，其最新版本（包括所有的修改单）适用于本指南。

13.4 随着人工智能技术与档案学理论的发展，本指南将适时进行修订和完善。