

# 档案数字化资源的智能检索与知识发现应用研究

刘德青

赤峰市融媒体中心 内蒙古赤峰 024000

**摘要:** 随着信息技术的飞速发展,档案数字化进程不断加速,海量的数字化档案资源为社会发展提供了丰富的数据支持。然而,如何高效地从这些资源中获取有价值的信息成为关键问题。本文深入探讨档案数字化资源的智能检索与知识发现应用,分析相关技术原理、应用现状、面临挑战及发展趋势,旨在推动档案领域智能化发展,提升档案资源利用效率,为各行业决策和研究提供有力支撑。

**关键词:** 档案数字化; 智能检索; 知识发现; 人工智能

## 引言

档案,作为历史长河中最为真实且珍贵的记录载体,不仅承载着海量的信息和知识,更是连接过去与未来的桥梁。在传统的档案管理模式下,面对着日益激增的档案数量,管理者往往面临着检索效率极为低下、知识挖掘过程异常艰难等诸多棘手问题,这些问题随着时间的推移愈发凸显,严重制约了档案资源的有效利用。

然而,随着数字化技术的迅猛发展和广泛应用,档案管理领域迎来了前所未有的变革,正式步入了全新的发展阶段。数字化技术的融入,使得档案资源的存储变得更为便捷高效,同时,档案的快速传播与共享也成为了现实。更为重要的是,在这一坚实基础之上,智能检索与知识发现技术的引入,犹如为档案管理插上了智慧的翅膀。这些先进技术能够帮助用户以更加精准、高效的方式获取到所需的档案信息,不仅极大地提升了检索效率,还能深入挖掘档案背后所隐藏的丰富知识关联,使得档案资源的内在价值得到充分挖掘和利用。

这种质的飞跃,对于推动社会各领域的持续发展、促进文化的有效传承以及助力学术研究的深入探索等方面,都具有不可估量的重要意义。可以说,智能技术的融入,不仅改变了档案管理的传统面貌,更为社会的全面进步注入了强劲动力。

## 一、档案数字化资源智能检索与知识发现的技术基础

### (一) 自然语言处理(NLP)技术

自然语言处理(NLP)技术在档案智能检索中扮演着至关重要的角色。它能够将用户以自然语言形式提出的查询请求,转化为计算机可以理解和执行的指令。这

一过程涉及到对查询语句进行深入的词法、句法和语义分析,从而实现了对查询意图的精准理解,并从中提取出关键信息,以便进行高效的检索操作。在知识发现领域,NLP技术同样发挥着重要作用。它能够对档案文本进行自动分类、聚类,以及生成摘要,从而帮助用户快速把握档案内容的核心要点,提高知识获取的效率。

### (二) 机器学习与深度学习算法

机器学习算法,如分类算法、聚类算法和关联规则挖掘算法等,在档案智能检索和知识发现中得到了广泛的应用。这些算法能够从大量的档案数据中学习潜在的模式和规律,从而提高检索的准确性和效率。而深度学习算法,作为机器学习的一种进阶形式,在处理复杂档案数据方面表现尤为出色。例如,卷积神经网络(CNN)可以用于图像档案的识别和检索,而循环神经网络(RNN)及其变体长短期记忆网络(LSTM)则能够有效地处理文本档案中的序列信息,从而进一步提高检索和知识发现的准确性。

### (三) 知识图谱技术

知识图谱技术通过图形化的方式展示实体之间的关系,为档案领域构建了一个全面的知识网络。当用户进行查询时,知识图谱能够提供与之相关的全面关联信息,帮助用户了解事件的全貌,发现知识之间的联系。此外,知识图谱还为智能推荐和知识推理提供了坚实的基础,使得档案检索和知识发现更加智能化。

### (四) OCR技术与多模态数据处理

OCR(光学字符识别)技术能够将纸质档案中的文字内容转换为电子文本,为智能检索和知识发现提供了宝贵的数据基础。随着OCR技术的不断进步,其识别的准确率也在不断提高,为档案的数字化和智能化处理提

供了有力支持。此外，多模态数据处理技术能够整合不同模态的数据信息，如文本、图像、音频和视频等，实现对这些不同类型档案资源的统一检索和知识发现，从而满足用户多样化的检索需求。

## 二、档案数字化资源智能检索与知识发现的应用现状

### （一）档案馆的智能检索服务升级

在当今时代，档案馆作为保存历史记忆的重要场所，正面临着数字化转型的挑战。为了提升服务质量，许多档案馆开始引入智能检索技术，以实现更高效、更便捷的档案查询服务。例如，昌平区档案馆通过接入DeepSeek大模型，实现了本地化部署，从而将传统的精准关键词检索升级为智慧问答，支持自然语言交互，大大提升了检索效率。此外，洛阳市档案馆也引进了AI大数据模型智能检索系统，该系统具有智能联想功能，能够提供便捷准确的查询服务，为用户节省了大量时间和精力。

### （二）基于知识发现的档案资源深度开发

为了更好地挖掘档案资源的价值，一些档案馆开始尝试基于知识发现的档案资源深度开发。例如，浙江省金华市档案馆打造了“红色档案知识库”，通过对红色档案数据化成果进行语义分析，关联标注信息，实现了“知识检索—档案溯源”的闭环服务。此外，还有一些档案馆利用数据挖掘和人工智能训练挖掘分析馆藏档案数据，建立专题数据库，为用户提供精准推送服务，从而更好地满足用户的需求。

### （三）企业档案管理中的智能应用

在企业档案管理领域，传统的管理模式也面临着数字化转型的挑战。为了应对这一挑战，一些企业开始尝试引入智能应用。例如，重庆移动与江苏移动项目聚焦AIGC技术应用，构建档案知识引擎，通过智能检索系统快速查找所需信息，利用知识发现技术挖掘知识，为企业发展提供战略指导。这些智能应用不仅提高了企业档案管理的效率，也为企业的发展提供了有力支持。

### （四）数字人文领域的档案应用拓展

数字人文是信息技术和人文学科深度融合的新兴领域，为档案数据资源提供重构策略。关联数据、数据挖掘、数据可视化等数字人文技术可为挖掘档案数据的知识价值提供有效工具。例如，针对档案文本内容，运用词频分析、共现分析、聚类分析等文本分析方法，绘制档案信息的词云图、关键词共现矩阵、聚类谱系图等，帮助厘清特定档案中的关键信息、内容要义和语词共现关系。聚焦档案中的“人、地、时、事、物、情感”等

记忆要素，运用社会网络分析、地理位置分析、时序分析、情感分析等方法，梳理人物网络关系、时空演化、情感倾向、事件脉络等，通过对记忆要素的组织与关联，再现记忆场景，阐释记忆内涵和情感取向。在生成式人工智能的引领下，语音识别、图像识别、机器学习等技术支持文本类、声像类多模态档案的内容提取和挖掘日益深入，实现多媒体档案数据的检索与利用。生成式人工智能的自然语言处理技术还可提高文本类档案数据实体识别和语义关系揭示的效率，推动档案数据挖掘的效率提升、深度凸显，为数字人文研究提供丰富的资料和研究视角。

## 三、档案数字化资源智能检索与知识发现面临的挑战

### （一）数据质量问题

档案数字化过程中，数据质量参差不齐。部分档案由于年代久远、保存不善等原因，OCR识别准确率低，导致文本错误、信息缺失。不同来源的档案数据格式不统一，字段定义不一致，给数据整合和分析带来困难。数据录入过程中的人为错误，如错别字、信息遗漏等，也影响了数据的准确性和完整性。这些数据质量问题严重制约了智能检索与知识发现的效果，可能导致检索结果不准确、知识发现不全面。例如，在进行知识图谱构建时，如果数据质量存在问题，可能会导致实体关系错误，影响知识图谱的可靠性和实用性。

### （二）技术瓶颈

虽然当前人工智能等技术发展迅速，但在档案领域的应用仍存在一些技术瓶颈。自然语言处理技术在理解档案领域的专业术语和复杂语义时，还存在一定误差。机器学习和深度学习算法需要大量的高质量数据进行训练，而档案数据的标注和预处理工作繁重，且标注标准难以统一。知识图谱构建过程中，实体抽取和关系识别的准确性有待提高，特别是对于一些历史档案中的模糊信息和隐含关系，识别难度较大。多模态数据处理技术在融合不同模态数据时，还面临数据对齐、特征融合等技术难题，影响了多模态检索和知识发现的效果。

### （三）安全与隐私保护

档案数字化资源包含大量敏感信息，如个人隐私、商业机密、政府机密等。在智能检索与知识发现过程中，数据的传输、存储和使用面临安全风险。网络攻击、数据泄露等问题可能导致档案信息被非法获取和利用。同时，如何在保护用户隐私的前提下，实现档案资源的合理利用，也是一个亟待解决的问题。例如，在进行档案数据分析和知识发现时，可能会涉及到个人信息的关联

和挖掘，如何确保这些操作符合隐私保护法规，不侵犯用户权益，是档案管理部门面临的重要挑战。

#### （四）人才短缺

档案数字化资源的智能检索与知识发现需要既懂档案业务知识，又掌握信息技术的复合型人才。目前，档案行业中这类复合型人才短缺。传统档案工作人员对新技术的掌握程度不足，难以有效运用智能检索和知识发现技术开展工作。而信息技术人员对档案业务流程和知识体系缺乏深入了解，在开发相关技术应用时，可能无法充分满足档案管理的实际需求。人才短缺限制了档案数字化资源智能应用的推广和深化发展。

### 四、档案数字化资源智能检索与知识发展的发展趋势

#### （一）多技术融合创新

未来，自然语言处理、机器学习、知识图谱、区块链等技术将在档案领域实现更深度的融合创新。例如，利用区块链技术保障档案数据的安全和可信，确保智能检索和知识发现过程中数据的真实性和完整性。将知识图谱与自然语言处理技术深度结合，进一步提高对用户查询意图的理解和知识关联的准确性，为用户提供更智能、全面的检索和知识服务。多模态数据处理技术将不断完善，实现文本、图像、音频、视频等多种模态数据的无缝融合，为用户提供更丰富、便捷的检索体验。通过技术融合创新，不断提升档案数字化资源智能检索与知识发现的效率和质量。

#### （二）个性化与精准化服务

随着人工智能技术的发展，档案智能检索与知识发现将更加注重视觉化和精准化服务。通过对用户行为数据和查询历史的分析，深入了解用户需求和兴趣偏好，为用户提供个性化的档案检索结果和知识推荐。例如，为科研人员提供与其研究领域相关的档案资料 and 知识发现成果；为企业用户提供与企业业务发展相关的档案信息和决策支持。利用智能算法对档案数据进行深度挖掘，精准定位用户所需信息，提高档案资源的利用效率，满足不同用户群体的多样化需求。

#### （三）跨领域应用拓展

档案数字化资源的智能检索与知识发现将不断拓展跨领域应用。在文化旅游领域，通过对历史档案的知识发现，开发具有文化内涵的旅游产品和线路，为游客提供更丰富的文化体验。在教育领域，利用档案资源开展历史、文化等课程教学，通过智能检索为教师和学生提供丰富的教学素材和研究资料。在医疗领域，通过对医

疗档案的智能分析和知识发现，为疾病研究、医疗决策等提供支持。跨领域应用拓展将进一步发挥档案数字化资源的价值，推动各行各业的发展。

#### （四）智能化与自动化发展

未来，档案数字化资源的管理将朝着智能化和自动化方向发展。智能检索系统将具备更强大的自主学习能力，能够根据用户反馈和数据变化不断优化检索算法和知识发现模型。档案的分类、编目、著录等工作将实现自动化，通过人工智能技术自动提取档案关键信息，完成相关工作流程，大大提高档案管理效率。在档案保护方面，利用智能技术实现对档案存储环境的实时监测和自动调控，保障档案实体和信息安全。智能化和自动化发展将使档案工作更加高效、便捷，为档案资源的开发利用提供更好的支持。

### 结束语

档案数字化资源的智能检索与知识发现应用是档案领域顺应时代发展的必然趋势。通过自然语言处理、机器学习、知识图谱等先进技术的应用，在档案馆服务升级、档案资源深度开发、企业档案管理以及数字人文等领域取得了显著成效。然而，在发展过程中也面临数据质量、技术瓶颈、安全隐私保护和人才短缺等挑战。未来，随着多技术融合创新、个性化精准化服务的推进、跨领域应用拓展以及智能化自动化发展，档案数字化资源的智能检索与知识发现将不断完善和深化，为社会发展、文化传承和各行各业决策研究提供更强大的支持，进一步发挥档案资源的重要价值。档案管理部门和相关技术人员应积极应对挑战，抓住发展机遇，推动档案数字化资源智能应用的持续发展。

### 参考文献

- [1] 郭乐. 新质生产力视域下档案资源的价值重构与开发利用[J]. 山西档案, 2025, (07): 59-61.
- [2] 柏松. 档案文化资源数字化保护、创造性转化与创新性发展路径[J]. 山西档案, 2025, (07): 87-90.
- [3] 沙莎. 数字环境下档案馆特色档案资源的整合路径[J]. 山西档案, 2025, (07): 162-164+172.
- [4] 李京举. 档案文化资源数字化开发与传播研究[J]. 山西档案, 2025, (07): 65-67.
- [5] 徐望. 基于文旅资源档案数字化的智慧文旅发展路径探索[J]. 江苏商论, 2025, (07): 53-57. DOI: 10.13395/j.cnki.issn.1009-0061.2025.07.030.