基于深度学习算法的电网档案智能检索

尹丽鹃[1] 宋 超[2] 刘剑磊[2]
（1. 国网甘肃省电力公司综合服务中心 2. 甘肃同兴智能科技发展有限责任公司）

摘要：电网档案检索依赖于用户输入的关键词与档案中关键词的精确匹配，忽略了档案内容的深层次联系和语义关系，导致相关档案无法被检索出来，降低检索效率，因此研究基于深度学习算法的电网档案智能检索方法。建立电网档案索引数据库，为后续检索提供数据基础，利用深度学习算法高效提取档案中的关键特征，以实现深度内容理解与识别，最后，通过匹配用户查询与档案特征，系统自动输出精确的检索结果。实验结果表明，该方法在所有测试档案类型上均显著提升检索速度，平均减少检索时间约 50%~70%，尤其在大规模数据处理上表现更为突出，证明了其在优化检索算法和提升系统性能方面的卓越成效。

关键词：深度学习算法；电网档案；电网档案检索；档案智能检索；深度学习

0 引言

随着电网规模的不断扩大和复杂度的日益提升，传统的电网档案管理方式面临着检索效率低、信息整合难度大、智能化水平不足等挑战。因此提出多种现代化手段进行档案检索。其中，网络档案信息检索可视化内容研究提出了网络档案信息检索可视化的多个维度，其中，档案资源分布可视化细分为馆藏档案实体分布、档案信息资源分布和档案记录对象分布；检索结果可视化则包括著录信息可视化、内容信息可视化和知识可视化[1]。但过多的可视化维度可能增加用户的认知负担，用户可能需要较长时间来理解和利用这些可视化信息，从而影响检索效率。面向数字记忆开发利用的档案检索模型构建研究则从数字记忆开发利用的角度出发，构建面向该目标的档案检索模型。该模型主要包括数据存储元素的提取，元素的语义关系的抽取，以及索引的构建及匹配三个核心部分，实现记忆的完整再现[2]。但面对海量、异构的档案数据，对现有语义理解的不足可能导致检索结果与用户实际需求之间存在偏差，进而影响检索效率。为此，基于深度学习算法的电网档案智能检索系统应运而生，为电网管理带来了革命性的变革。深度学习，作为人工智能领域的核心技术之一，以其强大的特征提取与模式识别能力，在图像识别、自然语言处理等方面具有很好的应用前景。将这一技术引入电网档案管理领域，可实现对海量电网档案数据的高效整合与精准检索，为电网规划、运维决策提供有力支持。

1 建立电网档案索引数据库

建立电网档案索引数据库的重要性在于，它确保所有电网档案能够以结构化的方式存储，并为后续的

快速检索提供可能。通过构建索引数据库，不仅提高数据的可访问性，还降低直接处理原始数据所带来的复杂性和成本。

数据采集通过先进的传感器网络、自动化仪表以及远程通讯技术，实时捕捉电网运行中的各项关键参数和状态信息。数据采集架构如图 1 所示。

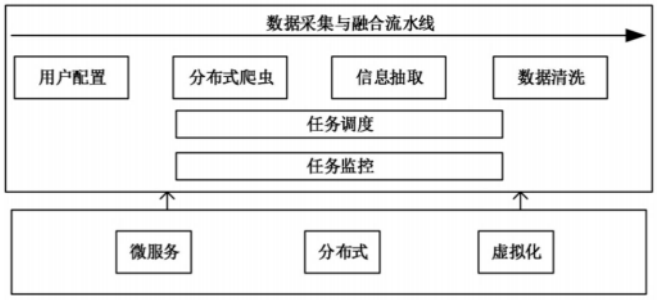

图 1　数据采集架构

对采集的数据进行整理，建立电网档案数据库。数据库作为电子文件管理系统的一环，承载着大量的信息。然而，鉴于电网档案数据（包括文档、图片、音频等）在数据结构上的独特性，这些非结构化数据难以直接通过传统的电子文件管理系统无缝整合至结构化数据库中。为了应对这一挑战，本文提出索引数据库的概念。索引数据库不仅能够将原本杂乱无章的电网档案数据进行系统性的整理，还能确保每一份档案都遵循统一的格式和标准进行存储。

假设存在一组电网档案数据，其中某些字段存在缺失值，使用均值填补法进行填补。如果某个数值型字段有 n 个观测值，其中 m 个是缺失的，那么填补后的值 $\hat{x}_i$（对于缺失的观测）可以表示为：

$$\hat{x}_i = \frac{1}{n-m} \sum_{j \in \text{非缺失索引}} x_j \tag{1}$$

通过这一步骤，确保电网档案数据的准确性和完整性。此外，值得注意的是，随着电力企业规模的扩大，对文件库的要求也越来越高。为了适应企业的长远发展，这就要求电力企业要在原有的基础上，对现有的数据库进行扩充，使其总体结构更加完善。为了量化数据库结构完善度，以各类档案的占比 p_i 的均匀性为目标。为了评估 p_i 的均匀性，使用熵（Entropy）的概念，在信息论中，熵越高，表示信息越不确定，也即分布越均匀[3]。对于离散概率分布 $\{p_i\}$（其中 $\sum_i p_i=1$），熵 H 定义为：

$$H = -\sum_{i=1}^{N} p_i \log 2^{p_i} \tag{2}$$

优化目标是使熵最大化，即 $H \to \max$。这表示各类档案的占比更加均匀。这样做不仅有助于及时更新数据索引库，提升数据库的综合性能，还能为企业提供更全面、更深入的决策支持。

2 利用深度学习算法提取档案特征

在索引数据库建立的基础上，深度学习算法能够自动从海量的电网档案数据中提取出关键信息和特征。这些特征不仅包括文本内容中的关键词、短语，还可能涉及图像的识别、时间序列的分析等多个维度。通过深度学习算法的高效特征提取，系统能够更准确地理解档案内容，为后续的信息匹配提供了强有力的支持。

利用深度学习算法，特别是卷积神经网络（CNN），层次结构严谨，由输入层起，经过层层卷积与池化，最终汇聚于全连接层，描绘 CNN 处理电网档案数据的内在逻辑。将电网档案映射为一系列隐含主题的概率图谱（即电网档案 - 主题模型），进而，每个主题又被细致地剖析为词汇概率的微观宇宙（主题 - 词汇）[4]。其基础构造如图 2 所示。

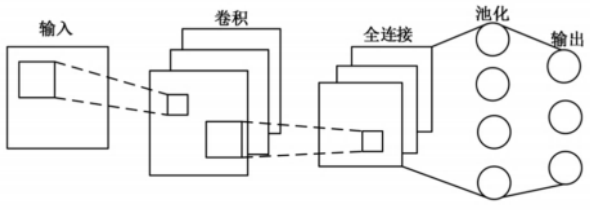

图 2　卷积神经网络的基础构造图

1）输入层：接收电网档案的数值化表示 X，其中 X 的维度取决于具体的数据表示方式。

2）卷积层：应用多个卷积核 K 对输入进行卷积操作，提取局部特征。数学上，这可以表示为：

$$Y=\sigma(X*K+b) \tag{3}$$

式中，$*$ 为卷积操作；σ 为激活函数（如 ReLU）；b 为偏置项。

3）池化层：对卷积层的输出进行最大池化操作，以减少数据维度并保留重要特征。

4）全连接层：将池化层的输出展平后，通过全连接层进行分类或回归等任务。全连接层的操作可以表示为：

$$h=\sigma(W\cdot \text{flatten}(Y)+b') \tag{4}$$

式中，W 为权重矩阵；b' 为偏置项；flatten(Y) 为将 Y 展平为向量的操作。

5）输出层：根据任务需求（如主题分类、关键词提取等），输出相应的概率分布或预测结果。

同时，增加局部连接机制，确保卷积神经网络能够聚焦于数据的局部特征，减少计算负担，同时保留关键的空间结构信息，为智能电网的智能化管理提供坚实的技术支撑。

3 档案信息匹配输出检索结果

在深度学习算法完成特征提取后，根据用户输入的查询条件，在索引数据库中进行快速的信息匹配。这一过程充分利用之前提取的特征信息，实现了对电网档案的高效检索。最终将匹配到的档案信息以用户友好的方式呈现出来，如列表、图表或详细报告等，以满足用户的实际需求。

空间索引将庞大的地理空间数据集分割成多个逻辑上紧密相连的子空间区域。这种划分策略极大地优化搜索过程，使得当用户进行位置查询时，系统能够自动锁定并探索那些与查询点最为接近的子空间，有效避免全数据集扫描的冗长与低效，从而实现搜索效率的飞跃。接着利用 K-means 算法，将用户输入的关键词与数据库中的相匹配。K-means 算法以 K 个点为初始簇核来进行聚类。然后，利用欧几里德距离公式，对各簇中心进行逐个计算，然后按照最小距离准则，将数据点按最近的簇进行赋值 [5]。当初始指派结束时，演算法会重新计算每一个群集的中心点，也就是这个群集中的全部数据点的平均值，这一过程将持续迭代，直至聚类结果趋于稳定。算法表达式如（5）所示：

$$\text{SSE}=\sum_{i=1}^{K}\sum_{x\in C_i}\left\|x-m_i\right\|^2 \tag{5}$$

式中，K 为聚类的总数；x 为数据点；$\left\|x-m_i\right\|^2$ 为数据点 x 与其所属聚类中心 m_i 之间的欧几里得距离的平方。在 K-means 算法的迭代过程中，不断尝试重新分配数据点和更新聚类中心，以最小化 SSE。当 SSE 的变化量达到预设的迭代次数时，算法停止，认为聚类结果已经收敛，找到最终的电网档案检索目标。

在数字档案与 Web 服务相结合的今天，通过此查询服务，服务器能够迅速响应用户的检索请求，根据索引款目中的主题标签、资源属性及存储路径等信息，智能筛选、排序并展示最符合用户需求的档案资料。

4 实验

4.1 实验准备

某大型电力公司拥有数十年的电网运行和维护记录，包括设备参数、故障报告、维修记录等多种类型的档案，总量超过百万份。为了满足公司日益增长的信息需求，优化其庞大的电网档案管理系统，引入本文基于深度学习的智能检索技术。

在引入这一智能检索技术之前，首先需要搭建一个稳定、高效的软件环境，以确保深度学习模型能够顺利运行并发挥出最佳性能。表 1 是软件环境搭建所需组件。

表 1　软件环境搭建

序号	设备	型号
1	CPU	Intel Core i9-10900K
2	内存	64GB DDR4 3200MHz
3	操作系统	Windows 10 Pro 64 位
4	图形处理器（GPU）	NVIDIA GeForce RTX 3090
5	存储	1TB NVMe SSD

通过以上步骤，可以搭建起一个稳定、高效的软件环境，为基于深度学习的电网档案智能检索技术的引入和实施奠定坚实的基础。

4.2 实验说明

公司组织专业人员对电网档案进行数字化处理，包括扫描纸质文档、转换电子文件格式等，确保所有档案均可被计算机识别和处理。随后，建立电网档案索引数据库，对档案进行结构化存储。

为了全面评估基于深度学习的电网档案智能检索技术的性能，选取以下五个具有代表性的电力档案数据类型作为实验对象：设备参数、故障报告、维修记录、巡检记录以及设备改造方案。这些档案数据不仅涵盖电网运行与维护的各个方面，而且其多样性和复杂性能够充分考验检索系统的效能。

在实验过程中，记录每一种档案类型从提交检索请求到正确返回所需档案数据的具体时间。这一时间指标直接反映了检索系统的响应速度和效率，是评估实验效果的关键依据。

同时应用本文方法、网络档案信息检索可视化内容研究以及面向数字记忆开发利用的档案检索模型构建方法于该公司电网档案，以验证本文方法的效果。对比不同档案类型的检索时间，深入分析智能检索技术在处理不同类型数据时的表现差异，进而识别出潜在的优化空间。

4.3 实验结果与分析

在本次电网档案智能检索效率对比实验中，通过对设备参数、故障报告、维修记录、巡检记录以及设备改造方案等多种类型的档案数据进行实验，结果如表 2 所示。

表 2　电网档案智能检索效率

档案类型	数据	方法 1	方法 2	本文方法
设备参数	1500 份	120	90	60
故障报告	1200 份	95	75	50
维修记录	1000 份	80	65	40
巡检记录	800 份	65	50	30
设备改造方案	500 份	45	35	20

本文方法在所有测试档案类型上均表现出显著的检索速度优势。相比于方法 1 和方法 2，本文方法的检索时间大幅度减少，尤其是在处理大规模档案数据（如设备参数和故障报告）时，检索速度的提升更为明显，本文方法的检索时间平均减少了约 50%~70%，这证明本文方法在优化检索算法和提升系统处理能力方面取得了显著成效。考虑检索速度这一指标，本文方法在所有测试档案类型上均优于方法 1 和方法 2。这表明本文方法在电网档案智能检索领域具有更高的实用价值和推广前景。

5 结束语

基于深度学习算法的电网档案智能检索方法通过自动化、智能化的方式，极大地提升了电网档案的检索效率与准确性，使得管理人员能够迅速获取所需信息，为电网的日常运营、故障排查及战略规划提供了强有力的数据支撑。但深度学习模型对训练数据的质量与数量高度依赖，数据的不完整或偏差可能影响检索结果的准确性。可以期待更高效的模型训练算法、更优化的数据处理技术以及更智能的人机交互界面的出现，从而进一步提升系统的性能与用户体验。总之，基于深度学习算法的电网档案智能检索系统将在不断完善与创新中，为电网行业的智能化发展贡献更大力量。

（下转第 91 页）

（上接第100页）

参考文献

[1] 赵屹．网络档案信息检索可视化内容研究[J]．档案学研究，2023（5）：124-130.

[2] 房小可．面向数字记忆开发利用的档案检索模型构建研究[J]．数字图书馆论坛，2021（11）：21-27.

[3] 高渝，宋若鹏，王明惠，等．电网建设项目档案在线归档实践[J]．中国档案，2022（9）：62-63.

[4] 彭玉芳，陈将浩，何志强．基于机器学习和深度学习的南海证据性数据抽取算法比较与应用[J]．现代情报，2022，42（2）：55-69.

[5] 刘伟，樊海玮．人工智能赋能高校档案检索技术研究[J]．档案天地，2023（6）：28-31.

（收稿日期：2024-09-05）