

大数据在非遗保护中的应用

——基于深度学习与云计算的融合架构

肖子涵, 贾庆元, 谢欣语

青岛黄海学院信息科学与工程学院, 山东 青岛

收稿日期: 2025年8月10日; 录用日期: 2025年9月9日; 发布日期: 2025年9月17日

摘要

非物质文化遗产(非遗)作为人类文明的活态载体, 正面临技艺断层、传播受限等挑战。本研究立足大数据时代背景, 构建基于深度学习与云计算的融合架构。先通过TF-IDF和知识图谱提取特征, 再利用CNN、BiLSTM-CRF完成非遗图像识别、实体解析, 云计算全程提供弹性存储与分布式计算资源, 为非遗保护提供数字化支持。

关键词

非物质文化遗产, 特征提取, 云计算, BiLSTM-CRF

The Application of Big Data in Intangible Cultural Heritage Protection

—Based on the Integrated Architecture of Deep Learning and Cloud Computing

Zihan Xiao, Qingyuan Jia, Xinyu Xie

School of Information Science and Engineering, Qingdao Huanghai University, Qingdao Shandong

Received: Aug. 10th, 2025; accepted: Sep. 9th, 2025; published: Sep. 17th, 2025

Abstract

As a living carrier of human civilization, Intangible Cultural Heritage (ICH) is facing challenges such as technical discontinuity and limited dissemination. Against the backdrop of the big data era, this study constructs an integrated architecture based on deep learning and cloud computing. Firstly, features are extracted using TF-IDF and knowledge graphs; subsequently, CNN and BiLSTM-CRF are employed

to accomplish ICH image recognition and entity parsing. Cloud computing provides elastic storage and distributed computing resources throughout the process, offering digital support for ICH protection.

Keywords

Intangible Cultural Heritage, Feature Extraction, Cloud Computing, BiLSTM-CRF

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

非物质文化遗产作为承载民族记忆与人类文明的活态传承，正面临着技艺流失、传播受限、保护手段滞后等多重挑战。传统非遗保护模式在海量数据采集、深度特征挖掘及跨域资源共享方面的局限性日益凸显。与此同时，信息科学正经历从计算模拟的“第三范式”向数据密集型科学发现的“第四范式”跃迁。大数据在非遗保护中应用广泛，主要体现在数据的采集、存储、处理和分析等环节；深度学习有强大的特征提取与模式识别能力，为非遗数据处理与分析提供核心动力；云计算则凭借弹性扩展的存储能力与高效协同的分布式计算架构。基于此，本文提出一套大数据 + 深度学习 + 云计算的融合架构(见图1)，重构非遗资源数字化采集、智能化分析、网络化传播的全链条体系为非遗的可持续发展提供新思路。

2. 概述

2.1. 大数据——非遗保护的基石

大数据是由规模巨大、类型多样、处理速度快且具有潜在价值的数据集组成，有 5 种特点：Volume (大量)、Velocity (高速)、Variety (多样)、Value (低价值密度)、Veracity (真实性)。

大数据下的“非遗”信息资源，以数字化处理为前提，将抽象知识关系体系与实体资源关系体系组成形成综合信息库[1]。这种数据关系体系也是大数据时代的数据存储模式，一方面能充分挖掘“非遗”的特征及价值，另一方面也能发挥各维度信息资源的交互性。通过对非遗数据的处理与特征提取，为后续处理与分析提供数据支撑。本文在大数据分析的基础上，提出了利用 TF-IDF [2]动态加权模型进行特征提取和时空融合的知识图谱，精准发现濒危非遗技艺。首先，通过文本向量化与主题建模构建基础分析框架，其次，利用深度神经网络实现结构化知识抽取，最后，建立多模态知识演化系统。

2.2. 深度学习——智能分析引擎

深度学习是机器学习的一个重要分支，它以人工神经网络为基础，通过模拟人脑的神经系统结构和工作机制，实现对数据的深度分析和模式识别。深度学习具有强大的特征学习能力，能够自动从海量数据中提取复杂的特征信息，无需人工进行特征工程设计，大大提高了数据处理效率和准确性。

在非遗保护中主要承担特征自动提取、模式智能识别与知识深度挖掘三大核心功能。在非遗图像识别方面，深度学习算法可以对非遗传统图案、服饰纹样、古建筑装饰等图像数据进行精准识别和分类。例如，通过训练深度卷积神经网络，可以让计算机自动识别不同非遗项目中的独特图案元素，实现非遗图案的数字化归档和检索。此外，深度学习在非遗技艺传承方面也能发挥重要作用。通过对非遗传承人的技艺展示视频进行动作捕捉和分析，利用深度学习算法可以提取出技艺动作的关键特征和运动轨迹，构建起非遗技艺的数字化模型。同时，基于深度学习的虚拟仿真技术还可以实现非遗技艺的模拟展示，

让人们能够更直观地感受非遗技艺的魅力。深度学习算法不仅有助于非遗技艺的记录和保存，还可以为非遗传承教学提供直观的参考资料，降低传承难度，提高传承效率。

2.3. 云计算——基础设施支撑

云计算是一种基于互联网的计算方式，通过将计算资源、存储资源、网络资源等进行集中管理和动态分配，为用户提供按需使用的服务。云计算具有高可靠性、高扩展性、高灵活性和低成本等特点，它将计算任务分布在大量的计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和信息服务。

正如王曙光在《数智时代云计算在文化遗产活化中的应用研究》[3]提到云计算作为新型基础设施，需协同大数据、人工智能、虚拟现实等前沿技术，驱动文化遗产数据价值生成。在实现路径上，依托云计算体系分层响应需求：基础资源层(IaaS)提供海量存储与弹性算力，实现文化遗产物理形态向数字形态的精准映射；平台服务层(PaaS)融合大数据、人工智能等技术，建立异构数字资源的高效治理机制；智能应用层(SaaS 或 MaaS)聚焦文化价值挖掘与沉浸式展示，推动传统元素在数字空间的创新延续。

3. 融合架构的关键技术

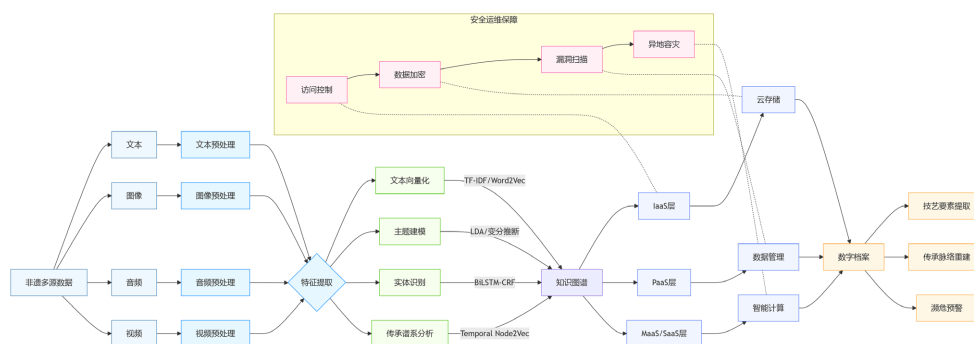


Figure 1. Overall technical architecture diagram
图 1. 总体技术架构图

3.1. 文本数据分析技术架构

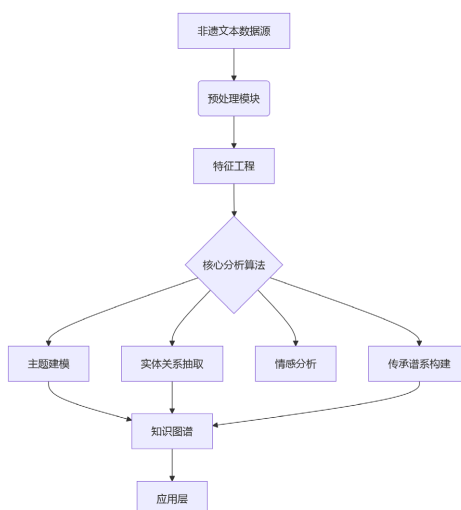


Figure 2. Flowchart of the technical architecture for text data analysis
图 2. 文本数据分析技术架构流程图

为实现非遗文本数据的深度挖掘，本研究构建了一套文本数据分析技术架构。首先经过预处理模块进行数据清洗、去噪与标准化。然后，进行特征提取与多算法融合，将非遗文本转化为结构化的知识体系，最终赋能于知识查询、传承决策与数字人文研究(见图 2)。

3.1.1. 数据来源与清洗

(1) 数据来源

本研究的数据以权威性和代表性为首要原则，来源于中国非物质文化遗产文化官网、各省市非物质文化遗产保护中心官网及官方媒体发布的非遗相关新闻报道(见表 1)，用 Python 编写爬虫程序使用 requests + BeautifulSoup 抓取静态网页。

该程序依赖于 Requests 库模拟 HTTP 请求，高效获取网页 HTML 源码。然后，利用 BeautifulSoup 库对源码进行解析，通过 CSS 选择器或 XPath 路径精准定位并提取目标数据字段，从而实现静态网页中非结构化信息的自动化抓取与结构化存储。

Table 1. Data source table

表 1. 数据来源表

地区	网站	新闻数量(篇)	总计(篇)
苏州	人民网	523	2785
	苏州新闻网	1620	
	名城苏州	642	
陕西	人民网	486	1890
	陕西非物质文化遗产网	720	
	中国非物质文化遗产文化网	684	
北京	人民网	4297	6627
	北京新闻网	1050	
	新京报	920	
	中国非物质文化遗产文化网	360	

(2) 数据清洗

针对原始数据中存在的噪声、不一致性与缺失值等问题，本研究进行了数据清洗与评估清洗后的数据质量。通过计算字段填充率、信息熵等指标，客观衡量了数据集的完备性、一致性及信息丰富度。所有经过清洗及评估的高质量数据均以结构化的 JSON 格式进行持久化存储，该格式明确定义了每个字段的名称、类型与语义描述，为后续的特征提取、知识图谱构建以及深度学习模型训练提供数据基础。

3.1.2. 文本向量化表示

(1) TF-IDF 加权模型

$$tf-idf(t, d) = tf(t, d) \times \log \frac{N}{df(t)}$$

其中 N 为总文档数， $df(t)$ 是包含词项 t 的文档数。

(2) 词嵌入模型(Word2Vec)

目标函数：

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c} \log p(w_{t+j} | w_t)$$

3.1.3. 主题建模

(1) LDA 算法

采用潜在狄利克雷分布(Latent Dirichlet Allocation)建模非遗文本的隐式主题结构[4]。假设每篇文档为话题的概率混合($\theta_d \sim Dir(\alpha)$)，每个话题为词语的概率分布($\phi_k \sim Dir(\beta)$)，通过变分 EM 算法求解文档 - 主题分布(γ_{di})与主题 - 词语分布(β_{iv})。

(2) 变分推断更新公式

$$\gamma_{di} = \alpha i + \sum_{n=1}^{N_d} \phi_{dni}$$

$$\phi_{dni} \propto \exp(\Psi(\gamma_{di}) + \Psi(\beta_{i w_n}))$$

3.1.4. 实体关系抽取

BiLSTM-CRF 模型

由双向长短期记忆网络(BiLSTM)与条件随机场(CRF)构成级联架构[5]。BiLSTM 层捕获上下文语义表征，CRF 层通过转移概率矩阵约束标签序列合法性，解码最优标注路径。

双向 LSTM 输出

$$\vec{h}_t = \overrightarrow{\text{LSTM}}(x_t, \vec{h}_{t-1})$$

$$\overleftarrow{h}_t = \overleftarrow{\text{LSTM}}(x_t, \overleftarrow{h}_{t-1})$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t]$$

CRF 解码层

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^T W_{y_{t-1}, y_t} + \sum_{t=1}^T V_{y_t}^T h_t\right)$$

3.1.5. 传承谱系分析

(1) 时序网络模型

定义传承关系图

$$G = (V, E, \tau)$$

(2) 节点嵌入算法(Temporal Node2Vec)

$$\max_f \sum_{u \in V} \log P(N_T(u) | f(u))$$

其中 $N_T(u)$ 为时间约束的邻居节点。

(3) 谱系稳定性度量

$$S = \frac{1}{|D|} \sum_{d \in D} \frac{|\{e \in E | \tau(e) < t_d\}|}{|\{e \in E | \tau(e) \geq t_d\}|}$$

其中 D 为传承人集合， t_d 为传承人活跃时间段。

3.2. 深度学习算法层

3.2.1. 视觉特征解析模型

(1) 卷积神经网络(CNN)架构

$$F_{i,j,f} = \sigma \left(\sum_{c=1}^C \sum_{u=-k/2}^{k/2} \sum_{v=-k/2}^{k/2} X_{i+u,j+v,c} \cdot W_{u,v,c,f} + bf \right)$$

$F_{i,j,f}$: 输出特征图第 f 个通道在位置 (i,j) 的激活值;

$\sigma(\cdot)$: ReLU 激活函数 ($\sigma(z) = \max(0, z)$), 引入非线性;

$W_{u,v,c,f}$: 卷积核参数;

bf : 第 f 个通道的偏置项。

(2) Vision Transformer 模型(见图 3)

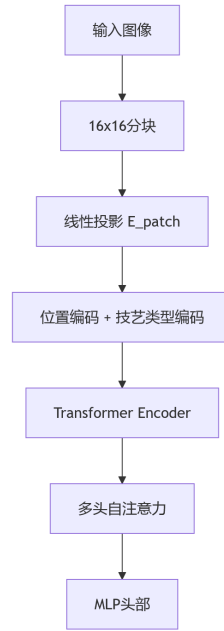


Figure 3. Flowchart of the ViT model

图 3. ViT 模型流程图

图像分块编码:

$$z_0 = [x_{class}; x_1^p E; \dots; x_n^p E] + E_{pos}$$

E_{pos} : 是位置嵌入矩阵, 其每一行对应序列中一个位置的位置编码。

序列长度: 分类标记 1 个 + N 个图像块 = N + 1。

自注意力机制增强:

$$Attention_{nh}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{dk}} + \log(P_{heritage}) \right) V$$

其中 $P_{heritage} \in \mathbb{R}^{n \times n}$ 为非遗传拓扑关联矩阵。

3.2.2. 生成式保护模型

(1) 生成对抗网络(GAN)

目标函数:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

(2) 变分自编码器(VAE)

证据下界(ELBO)

$$L$$

参数设置:

隐变量分布: $p(z) = N(0, I) z \in \mathbb{R}^d$;

推理网络(编码器): $q_\phi(z|x) = N(u_\phi(x), \sigma_\phi^2(x)I)$;

生成网络(解码器): $p_\theta(x|z) = \begin{cases} N(\mu_\theta(z), I) & \text{连续数据} \\ \prod_i \text{Bern}(x_i | u_{\theta,i}(z)) & \text{离散数据} \end{cases}$ 。

3.3. 云计算支撑层

3.3.1. IaaS 层

(1) 技术原理

采用资源虚拟化模型实现物理资源池化, 计算资源分配满足

$$V_{Malloc} = \left\lceil \frac{\sum_{i=1}^n (C_i^{CPU} \times t_i^{task})}{T_{slot}} \right\rceil \times \eta_{util}$$

其中 C_i^{CPU} 为任务计算强度系数 t_i^{task} 为任务时长, η_{util} 为集群利用率。

(2) 存储资源保障:

多副本机制: $R_{copies} = 3 + \lfloor \log_2(S_{data}/1TB) \rfloor$;

久性指标: $Durability = 1 - 10^{-12}$ 。

3.3.2. PaaS 层

弹性伸缩算法:

$$\Delta N_{worker} = k \cdot \frac{\partial Q_{queue}}{\partial t} + \beta \cdot \text{sgn}(Q_{queue} - Q_{threshold})$$

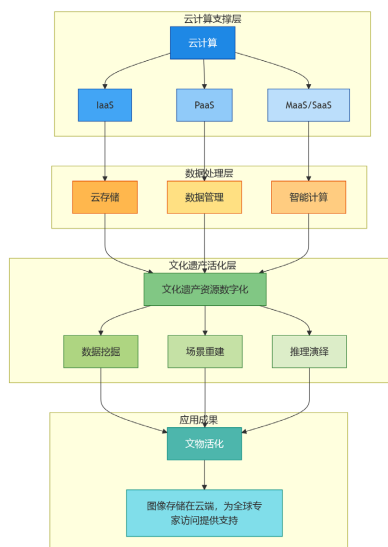


Figure 4. Flowchart of cloud computing empowerment
图 4. 云计算赋能流程图

3.3.3. MaaS/SaaS 层

服务化封装：

$$API_{service}(x)=g\theta\circ f_{\phi}\circ En_{che}(x)$$

f_{ϕ} ：文化遗产特征提取函数 $g\theta$ ：领域适配层；

En_{che} ：同态加密传输。

4. 融合架构在非遗保护中的应用

在非遗资源数字化归档方面，该融合架构发挥着关键效用首先通过大数据采集技术，对非遗项目的各类多维资源进行全面采集。然后，利用云计算平台的存储能力(图 4)，对这些数据进行安全存储。最后，运用深度学习技术对图像和视频数据进行处理，结合 Adobe Illustrator 的图像描摹功能完成纹样矢量转换(见图 5)，实现非遗图案的自动识别与分类、技艺动作的提取与分析，为非遗资源的数字化归档提供精准标签与索引，便于后续查询与管理。



Figure 5. Adobe Illustrator pattern extraction (image source: internet) [6]
图 5. Adobe Illustrator 纹样提取(图片来源：网络) [6]

在非遗传承与教学方面，融合架构提供了创新方式与方法。通过深度学习技术对传承人的技艺动作进行分析与建模，构建非遗技艺的数字化模型。利用云计算平台的共享功能，将这些模型及相关教学资源上传至云端，供传承人教学与学习者学习使用。同时，融合架构还能根据学习者的学习数据，通过深度学习算法分析学习情况，为其提供个性化学习建议与指导。

在非遗文化传播方面，融合架构能够拓展非遗的传播广度与影响力。通过大数据分析技术对受众群体进行画像分析，了解不同受众的兴趣爱好与需求特点。基于这些分析结果，利用云计算平台的分发能力，将非遗数字资源精准推送给目标受众。同时，结合深度学习技术生成个性化非遗文化内容，提高受众参与度与满意度。此外，融合架构还可支持非遗文化的线上展示活动，如虚拟非遗展览等，让更多人能便捷地感受非遗文化的魅力。

“AR 数字探寻龟兹文化之美”项目(见图 6)采用视觉 - 惯性 SLAM 技术，融合 LiDAR 点云数据与视觉特征匹配，确保虚拟内容与真实石窟环境的精准叠加。用户可通过自然手势操作虚拟文物，通过语音命令获取导览信息，系统实时跟踪用户视线焦点提供智能信息提示。此外采用边缘计算与云计算协同

的工作流程,复杂的高精度壁画和文物模型由云端渲染,通过 5G 网络低延迟传输至终端设备,实现了从智能手机到 AR 眼镜多种终端的无缝体验。



Figure 6. “AR Digital Exploration of the Beauty of Kucha Culture” wearable devices interact with smart mobile terminals (image source: internet) [7]

图 6. “AR 数字探寻龟兹文化之美”穿戴设备与智能移动端互动(图片来源:网络) [7]

5. 结论

本研究验证了多技术融合应用于非遗保护的可行性及应用价值。数据处理方面,采用 TF-IDF 动态加权与知识图谱进行特征提取,解决了人工提取效率低、覆盖窄的问题;深度学习方面, CNN 与 Vision Transformer 用于图像识别分类, BiLSTM-CRF 实现文本实体抽取与谱系重建, GAN 和 VAE 支持技艺数字化建模与仿真,提升了处理智能化水平;云计算方面,依托 IaaS 弹性资源、PaaS 异构治理与 MaaS/SaaS 服务封装,构建分布式架构,为非遗数据存储、处理与共享提供稳定支撑。未来,本研究将聚焦于优化多模态数据融合算法与跨领域资源共享机制,推动技术赋能与非遗活态传承的深度耦合。

参考文献

- [1] 达妮莎, 王爱玲. 大数据环境中非物质文化遗产的信息分析[J]. 大连理工大学学报(社会科学版), 2015, 36(4): 132-136.
- [2] 李鹏程. 基于深度学习的文化遗产文本大数据分析研究[D]: [硕士学位论文]. 青岛: 青岛理工大学, 2024.
- [3] 王曙光, 范晖, 李展. 数智时代云计算在文化遗产活化中的应用研究[J/OL]. 计算机技术与发展, 1-9. https://kns.cnki.net/kcms2/article/abstract?v=bJ89IKU86K_bn1PnLvQ_h-apnCUBmOh-hqA4VF5SAkg2aa9AESEpt2Mp0qpRTHbMSpSMvtB8zcr3OJ1Z6etvitrQvcd_W14mi44bVtyWn-lQtmCfjt_PwXNTZWtk2cjlz9vP6goAolIrp_WnhBvL0_BUG3-xXD68Efr5zSinKLIAjZTZiNG_Jg==&uniplatform=NZKPT&language=CHS, 2025-08-05.
- [4] Pylop, P., Roman, M. and Andrey, P. (2023) The Latent Dirichlet Allocation (LDA) Generative Model for Automating Process of Rendering Judicial Decisions. *E3S Web of Conferences*, **431**, Article 05005.
- [5] Li, Y., Luo, L., Zeng, X. and Han, Z. (2025) Fine-Tuned BERT-BiLSTM-CRF Approach for Named Entity Recognition in Geological Disaster Texts. *Earth Science Informatics*, **18**, Article No. 368. <https://doi.org/10.1007/s12145-025-01870-5>
- [6] 毛德龙. 虚拟展馆在非物质文化遗产保护中的应用研究[D]: [硕士学位论文]. 沈阳: 鲁迅美术学院, 2025.
- [7] 蔡睿瑶. 数字影像在民族文化保护与传承中的应用研究[D]: [硕士学位论文]. 沈阳: 鲁迅美术学院, 2025.