Computer Science and Application 计算机科学与应用, 2026, 16(3), 638-653
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163091

Hans汉斯

基于区块链和语义增强的科研诚信智能管控平台

陈丽丽，李永忠，邹倩瑜*，周宏虹，邱舟强

广东省科技创新监测研究中心，广东 广州

收稿日期：2026年2月4日；录用日期：2026年3月3日；发布日期：2026年3月16日

摘　要

针对科研诚信共享“不敢、不愿、不能共享”的问题，及现有技术在架构-算法协同、检索深度、安全管控等方面的不足，本文提出并实现基于区块链与领域自适应语义增强检索的科研诚信平台。平台立足广东省科研诚信管理实际，构建五级协同架构，通过混合存储、动态权重混合共识与精细化智能合约，保障信息共享安全高效；设计专属检索算法，融合领域预训练语言模型与动态注意力机制，实现多源异构科研诚信信息的深度关联与精准检索；依托智能预警预测模块，结合时序演化特征挖掘与科研主体关联结构解析，构建多维度失信风险预测模型，整合历史行为与主体关联特征，实现科研失信行为预警与全链条精准管控。实验验证，平台核心指标显著优于传统数据库与现有区块链平台，大规模数据场景下仍高效稳定，为科研诚信信息跨域互联互通与智能化管控提供完整解决方案。

关键词

区块链，语义增强检索，智能预警，领域自适应算法

Intelligent Management and Control Platform for Scientific Research Integrity Based on Blockchain and Semantic Enhancement

Lili Chen, Yongzhong Li, Qianyu Zou*, Honghong Zhou, Zhouqiang Qiu

Guangdong Science and Technology Innovation Monitoring and Research Center, Guangzhou Guangdong

Received: February 4, 2026; accepted: March 3, 2026; published: March 16, 2026

*通讯作者。

文章引用: 陈丽丽, 李永忠, 邹倩瑜, 周宏虹, 邱舟强. 基于区块链和语义增强的科研诚信智能管控平台[J]. 计算机科学与应用, 2026, 16(3): 638-653. DOI: 10.12677/csa.2026.163091

Abstract

Addressing the issues of "unwillingness, reluctance, and inability to share" in scientific research integrity data sharing, and the shortcomings of existing technologies in architecture-algorithm collaboration, retrieval depth, and security control, this paper proposes and implements a scientific research integrity platform based on blockchain and domain-adaptive semantic-enhanced retrieval. Based on the actual scientific research integrity management practices in Guangdong Province, the platform constructs a five-level collaborative architecture. Through hybrid storage, dynamic weighted hybrid consensus, and refined smart contracts, it ensures secure and efficient information sharing. A dedicated retrieval algorithm is designed, integrating a domain-pretrained language model and a dynamic attention mechanism, to achieve deep association and precise retrieval of multi-source heterogeneous scientific research integrity information. Leveraging an intelligent early warning and prediction module, combined with temporal evolution feature mining and scientific subject association structure analysis, a multi-dimensional misconduct risk prediction model is constructed, integrating historical behavior and subject association features to achieve early warning and precise control of scientific misconduct throughout the entire chain. Experimental results show that the platform's core indicators significantly outperform traditional databases and existing blockchain platforms, remaining efficient and stable even in large-scale data scenarios, providing a complete solution for cross-domain interconnection and intelligent management of scientific research integrity information.

Keywords

Blockchain, Semantic-Enhanced Retrieval, Intelligent Early Warning, Domain Adaptation Algorithm

1. 引言

科研诚信是科技创新的基石，是规范科研活动秩序、维护科研生态健康发展的核心保障[1]。随着我国科技创新投入持续加大，广东省研究与试验发展经费投入已连续 3 年位居全国首位，2022 年全省跨地市申报科研项目达 15.6 万项，占全省项目总量的 29%。然而，伴随科研活动的持续扩张，科研失信行为呈现多发、复杂化态势，且因信息共享机制缺失导致的“一地失信、多地获资助”问题尤为突出[2]，严重扰乱科研管理秩序。为遏制科研失信行为，我国先后印发一系列政策文件，明确要求构建全域覆盖、互联互通的科研诚信信息共享体系[3]。但现有科研诚信管理模式仍面临三大短板：1) 信息孤岛问题突出，跨层级互通受阻，全省累计 375 万条科研诚信信息无法实现跨层级、跨区域高效流通[4]，形成典型的“数据烟囱”现象；2) 敏感信息安全风险高，存储与传输存在隐患，传统集中式存储架构面临严重的篡改风险与数据泄露风险[5]；3) 检索效率与精度不足，难以满足实时监管需求，无法实现同义术语的有效关联，且复杂查询响应时间超 8 秒，远不能满足科研诚信实时监管、快速核查的业务需求[6]。4) 监管模式被动滞后，缺乏前瞻性预警能力。现有管理多依赖事后核查与处罚，未能基于历史数据挖掘失信行为演化规律，无法提前识别高风险主体与项目，导致监管成本高、防控效果有限。

现有研究为科研诚信信息共享提供了思路，但仍存在明显局限性：区块链技术的去中心化特性为解决信任问题提供了可能，但现有研究未明确不同敏感等级信息的分级上链流程与加密标准[7]，且仅依赖

权限控制与加密存储，未结合联邦学习等隐私计算技术，难以从根本上消除“不敢共享”的顾虑；部分研究尝试构建区块链科研诚信平台，但未设计适配多类型数据的共识机制与可落地的智能合约体系，且未解决架构与检索算法的协同优化问题[8]；语义检索技术虽在政务数据检索中有所应用，但现有方案多为通用型设计，未结合科研诚信领域术语特点进行定制化优化[9]；在失信风险预测方面，现有研究多采用单一机器学习模型，未能充分挖掘科研主体间关联关系与失信行为时间演化特征，预测精度与泛化能力不足，且未与区块链共享平台深度融合，难以实现预警结果的可信溯源与联动管控。

针对上述问题，本文立足广东省科研诚信管理实际需求，设计并实现区块链与领域自适应语义增强检索、智能预警预测深度协同的科研诚信信息共享平台。一方面，通过区块链五级架构的精细化设计，明确核心问题；另一方面，创新设计领域自适应语义增强检索算法，实现多源异构数据的深度关联与高效检索；同时，利用智能预警与预测分析模块，挖掘历史数据中失信行为的时间规律与关联特征，实现科研失信风险的提前预警。

2. 相关工作

2.1. 科研诚信信息共享平台研究

国内外学者围绕科研诚信信息共享展开了一系列探索，但仍存在显著不足。国外方面，构建集中式科研失信案例数据库，虽实现部分信息整合，但存在单点故障与数据篡改风险[10]；还有某些科研诚信系统采用分布式架构实现跨机构信息共享，但跨机构查询响应时间长达 48 小时，效率低下，无法满足实时监管需求[11]。国内方面，孙晶等[3]分析了广东省科研信用管理体系建设现状与“数据烟囱”问题，但未提出具体的技术解决方案；鲍锋等[4]设计了基于区块链的科研信息共享平台框架，但仅验证了存储安全性，未涉及检索效率、跨层级协同机制与风险预测功能；胡伏湘等[11]从高校科研诚信管理视角提出了区块链应用思路，但仅覆盖单一场景，无法满足省市跨层级、跨区域的共享需求与智能化预警需求。

2.2. 区块链在政务信息共享中的应用

区块链技术因其去中心化、不可篡改、可追溯等特性，在政务信息共享领域的应用逐步成熟。郑荣等[10]基于区块链构建了政府数据开放共享平台，但未设计适配多类型政务数据的分层共识机制，导致敏感数据与非敏感数据的安全管控缺乏差异化；张雪媛等[12]将区块链用于科学实验数据协同管理，提升了数据共享的信任度，但未给出架构各层的详细技术实现方案与性能优化策略；Gao 等[13]提出了基于区块链的建筑项目信息完整性保障系统，但未涉及信息检索与跨域协同问题。部分研究尝试引入联邦学习实现跨机构隐私保护模型训练，但尚未将其与区块链科研诚信平台深度融合，未能实现“数据可用不可见”与“全程可追溯”的协同。现有应用表明，区块链技术在政务信息共享中的落地，需针对数据的特点，设计差异化的共识机制和智能合约体系，同时需融合人工智能技术实现智能化管控[14]。

2.3. 信息检索技术在政务数据中的发展

传统政务数据检索以关键词匹配为主，面对海量、非结构化的政务数据时，存在检索精度低、漏检率高的问题[15]。语义检索技术通过捕捉文本深层语义关联，显著提升了检索性能，但在科研诚信领域的应用仍处于起步阶段。Liu 等[16]提出面向政府开放数据的领域增强语义检索算法，但未结合区块链架构进行协同设计，存在数据安全与检索效率的平衡问题；覃俊等[17]基于 BERT 与主题模型联合增强的长文档检索模型，提升了长文本检索精度，但未实现多源数据的关联检索；Zhen 等[18]提出了基于语义分片的区块链信任检索方案，但未针对科研诚信领域的术语特点进行定制化优化。

3. 算法模型与平台架构设计

3.1. 领域增强语义检索算法设计

本研究创新设计领域自适应语义增强检索算法，以“文本预处理－领域自适应词嵌入编码－动态双因子注意力加权－链上链下协同向量检索”四步核心流程，实现多源科研诚信信息的深度关联与精准检索。算法四大模块形成闭环协同，与区块链架构层级深度耦合，依托总损失函数反向优化参数，实现检索精度、效率的协同最优，整体框架如图 1 所示。

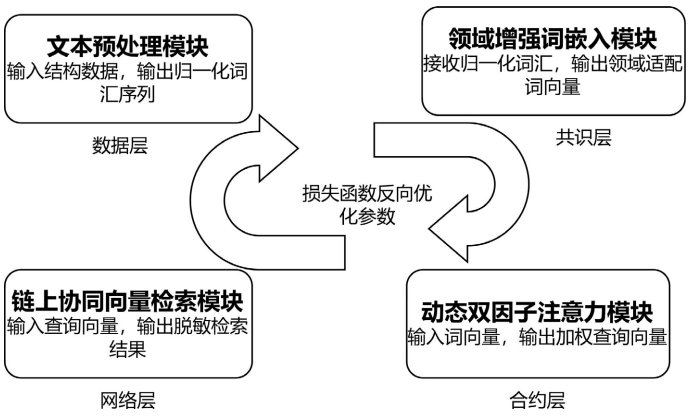

Figure 1. Algorithm framework diagram
图 1. 算法框架图

3.1.1. 模块关联逻辑

1) 数据流关联：文本预处理模块的归一化词汇序列输入领域自适应词嵌入模块，生成的文本向量与用户查询向量共同输入动态双因子注意力模块，该模块的加权查询向量作为链上链下协同模块的检索条件。

2) 控制流关联：检索权限的校验结果，反向控制注意力的权重分配，避免越权检索与敏感信息泄露。

3) 反馈流关联：检索模块校验结果反馈至总损失函数，不一致则微调词嵌入模块参数、调整注意力权重，提升向量抗篡改能力与检索精度。

3.1.2. 各模块详细设计

(1) 文本预处理模块

输入数据来自区块链数据层的结构化与半结构化存储，预处理流程如下：

1) 数据读取与解密：应用层调用数据层接口，获取加密存储的科研诚信文本数据，通过合约层的解密权限校验后，采用 AES-256 算法解密数据；

2) 领域适配分词：采用结巴分词工具，结合数据层预存储的科研诚信领域专用词典，将文本分割为词汇序列 $W=\{w_1,w_2,\cdots,w_n\}$ (n 为词汇数量)，如“2022 年省级项目经费滥用”分词为{2022 年，省级项目，经费滥用}；

3) 去停用词：基于合约层统一管理的科研诚信领域停用词表，移除词汇序列中的停用词，保留核心词汇 $W'=\{w_1',w_2',\cdots,w_m'\}(m\le n)$，停用词表的更新需经核心节点 PBFT 共识通过，确保全网一致性；

4) 术语归一化：通过合约层的术语映射合约，调用领域术语映射表，消除同义术语歧义，确保语义一致性。术语映射表的更新需经核心节点与属地共识节点的混合共识通过。

(2) 领域增强词嵌入编码

为精准捕捉科研诚信领域术语的语义关联，采用基于中文科技文本微调的 BERT 预训练语言模型(PLMs)将核心词汇转化为高维向量，具体流程如下：

1) 基于数据层 3.2 万例科研失信案例、48 万条项目信用记录，构建包含 3200 个领域核心术语、1200 组同义术语映射的科研诚信领域词典 D；对中文科技文本微调的 BERT 模型二次预训练，设指定超参数，经 Masked Language Model 优化参数，生成适配该领域的词嵌入模型。

2) 词向量生成：对核心词汇 w_i'，调用数据层预训练模型，生成维度 $d=512$ 的领域增强词向量：

$$v(w_i') = PLMs(w_i'; \theta, D) \tag{1}$$

其中，θ 为模型参数，D 为领域词典，$v(w_i') \in R^{512}$ 为词汇 w_i' 的词向量。

3) 文本向量生成：对核心词汇序列 w'，采用加权均值池化生成文本整体语义向量 V，向量哈希值同步至共识层，确保后续检索时向量未被篡改，反映文本全局语义信息：

$$V = \frac{\sum_{i=1}^{m} \gamma_i \cdot v(w_i')}{\sum_{i=1}^{m} \gamma_i} \tag{2}$$

其中，若 $w_i' \in D$，$\gamma_i = 1.2$，否则 $\gamma_i = 1.0$，$V \in R^{512}$ 为文本向量。

(3) 动态双因子注意力加权

为精准捕捉用户检索意图，强化领域术语与核心信息的权重，设计动态双因子注意力加权机制，具体流程如下：

1) 查询语句输入与加密传输：用户通过应用层共享查询模块输入查询语句(如“2022 年省级项目经费滥用企业”)，查询请求经网络层 TLS 1.3 协议加密后，传输至共识节点；

2) 领域术语重要性因子：合约层自动判断查询词是否属于领域词典 D，分配重要性因子 β_j（$q_j \in D$ 则 $\beta_j = 1.5$），否则 $\beta_j = 1.0$），规则不可篡改；

3) 语义相似度计算因子：计算查询词与文本核心词汇的余弦相似度，衡量语义关联程度：

$$sim(q_j, w_i') = \frac{v(q_j) \cdot v(w_i')}{\|v(q_j)\| \cdot \|v(w_i')\|} \tag{3}$$

其中，$sim(q_j, w_i') \in [-1, 1]$，值越大表示语义关联越强；

4) 动态权重分配：融合语义相似度与术语重要性，生成查询词权重 α_j，计算过程日志上链存数据层，基于相似度分配查询词权重，突出“2022 年”、“省级项目”、“经费滥用”等关键信息：

$$\alpha_j = \frac{\beta_j \cdot \sum_{i=1}^{m} sim(q_j, w_i')}{\sum_{t=1}^{k} \beta_t \cdot \sum_{i=1}^{m} sim(q_j, w_i')} \tag{4}$$

其中，α_j 为查询词 q_j 的权重，满足 $\sum_{j=1}^{k} \alpha_j = 1$；

5) 加权查询向量生成：应用层调用合约层接口，计算查询向量 V_Q，向量哈希上链存数据层，融合权重与词向量，生成查询语句的语义向量：

$$V_Q = \sum_{j=1}^{k} \alpha_j \cdot v(q_j) \tag{5}$$

其中，$V_Q \in R^{512}$ 为加权查询向量，可精准反映用户检索意图。

(4) 向量检索

采用“链上索引 + 链下向量库”模式，数据层存向量索引哈希与关联 ID，链下部署 FAISS 向量数据库，将文本向量存储于 FAISS 向量数据库，计算 V_Q 与文本向量 V 的余弦相似度，筛选 Top-N 结果：

1) 检索权限校验：合约层调用权限管控合约，验证用户角色，通过后生成临时检索令牌；

2) 向量库访问：应用层携带令牌，经网络层 P2P 通道访问链下向量库；

3) 相似度匹配：计算 V_Q 与向量库中文本向量 V 的余弦相似度，筛选相似度 >0.65 的结果：

$$sim\left(V_Q,V\right)=\frac{V_Q\cdot V}{\left\|V_Q\right\|\cdot\left\|V\right\|} \tag{6}$$

其中设定相似度阈值 $\tau=0.6$，筛选相似度大于 τ 的文本作为检索结果，实现语义精准匹配。

4) 结果溯源：通过数据层存储的向量关联 ID，查询链上原始数据哈希，验证检索结果的一致性；

5) 日志上链：合约层自动记录检索日志，日志哈希上链存数据层，实现全流程可溯源。

(5) 总损失函数设计

提出三重损失加权求和的总损失函数，反向优化词嵌入参数 θ、注意力权重 α_j，实现模块协同优化：

$$L_{total}=\lambda_1\cdot L_{sim}+\lambda_2\cdot L_{term}+\lambda_3\cdot L_{hash} \tag{7}$$

其中，$\lambda_1=0.4$、$\lambda_2=0.3$、$\lambda_3=0.3$ (通过交叉验证确定权重)，各损失项定义如下：

1) 语义相似度损失 L_{sim}：最小化查询向量与相关向量距离，最大化与无关向量的距离，确保检索精度：

$$L_{sim}=\frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0,sim\left(V_Q,V_{neg,i}\right)-sim\left(V_Q,V_{pos,i}\right)+b\right)\right] \tag{8}$$

其中 $V_{neg,i}$ 为相关文本向量，$V_{neg,i}$ 为无关文本向量，b 为防止过拟合的参数。

2) 领域术语匹配损失 L_{term}：确保领域术语的权重与匹配精度，避免术语被低估：

$$L_{term}=1-\frac{1}{M}\sum_{j=1}^{M}\alpha_j\cdot sim\left(q_j,D_{sc,j}\right) \tag{9}$$

其中 M 为查询中的领域术语数量，$D_{sc,j}$ 为领域词典中对应术语的标准向量，α_j 为注意力权重。

3) 链上哈希一致性损失 L_{hash}：最小化链下向量哈希与链上共识哈希的差异，确保向量未被篡改：

$$L_{hash}=\frac{1}{K}\sum_{k=1}^{K}\left|H\left(V_k\right)-H_{chain}\left(V_k\right)\right| \tag{10}$$

其中 $H(\cdot)$ 为 SHA-256 哈希函数，$H_{chain}\left(V_k\right)$ 为链上共识后的向量哈希，K 为向量数量。

3.2. 智能预警与预测分析模块设计

本模块挖掘科研失信时间演化规律与主体关联特征，构建失信风险预测模型并深度耦合区块链五级架构，各层协同形成闭环流程，模块整体框架如图 2 所示。

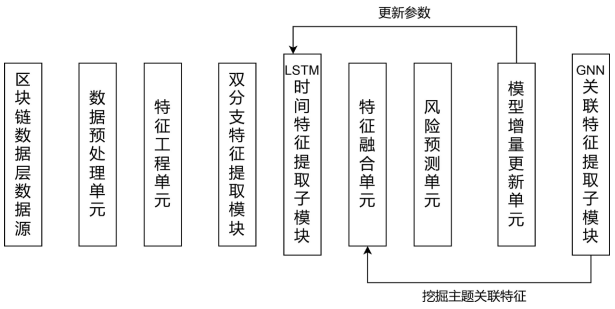

Figure 2. Intelligent early warning and analysis module
图 2. 智能预警与分析模块

3.2.1. 模块关联逻辑

1) 数据关联：模型训练数据来源于区块链数据层存储的历史检索日志、科研人员档案、项目信用记录、失信案例等多源数据，经数据预处理单元清洗后输入特征工程单元；

2) 参数关联：模型训练完成后的最优参数经 PBFT 共识后上链存储于合约层，确保模型参数不可篡改，每次预测时从合约层调用参数加载模型；

3) 预警关联：预测结果经合约层预警合约校验后，根据风险等级触发不同预警机制，高风险结果自动同步至应用层管控模块，触发核查流程；

4) 反馈关联：预警结果作为反馈数据回传至模型，驱动模型更新，更新后的参数重新经共识后上链。

3.2.2. 模块关联逻辑

1) 数据预处理单元

针对数据层多源异构数据的特点，进行标准化预处理：

a) 数据集成：整合链上存储的 127 万份科研人员档案、48 万条项目记录、3.2 万例失信案例及 500 万条历史检索日志，建立“人员–项目–案例”关联数据视图；

b) 数据清洗：采用缺失值填充、异常值剔除、数据去重等操作，确保数据质量；

c) 数据标准化：对数值型特征进行 Z-score 标准化：

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{11}$$

其中，μ 为特征均值，σ 为特征标准差，x_{norm} 为标准化后的特征值；

d) 数据时序化：以时间戳为索引，将每个科研主体的历史数据形成长度为 T = 12 (月)的时序样本。

2) 特征工程单元

构建多维度特征体系，涵盖基础特征、时间特征、关联特征三大类：

a) 基础特征 X_{base}：包括科研主体类型、项目类型、经费规模、历史失信次数等 16 维结构化特征；

b) 时间特征 X_{time}：基于历史数据提取的时序统计特征，包括近 12 个月项目申报频率、经费使用波动系数、检索关键词敏感程度变化等 8 维特征；

c) 关联特征 X_{rel}：构建科研主体关联图 $G = (V, E)$，其中节点 V 代表科研人员、项目、合作单位，边 E 代表合作关系、项目参与关系等，基于图结构提取节点度数、聚类系数、关联主体失信率等特征。

3) LSTM 时间特征提取子模块

采用 LSTM 网络捕捉科研失信行为的时间演化特征，其核心公式如下：

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t &= \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \\ \tilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \\ h_t &= o_t \odot \tanh\left(C_t\right) \end{aligned} \tag{12}$$

其中，σ 为 sigmoid 激活函数，$\odot$ 为元素-wise 乘积，W_f, W_i, W_C, W_o 为权重矩阵，b_f, b_i, b_C, b_o 为偏置项，h_{t-1} 为上一时刻隐藏状态，C_{t-1} 为上一时刻细胞状态，x_t 为 t 时刻输入特征，h_t 为 t 时刻输出的时间特征向量。

将时序化的特征序列 $X_{seq} = \{x_1, x_2, \cdots, x_T\}$ 输入 LSTM 网络，取最后一个时刻的隐藏状态 h_T 作为时间

特征提取结果 $F_{time}=h_T$。

4) GNN 关联特征提取子模块

采用 GCN 提取科研主体间的关联特征，通过聚合邻居节点信息更新当前节点特征，核心公式如下：

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}+b^{(l)}\right) \tag{13}$$

其中，$H^{(l)}$ 为第 l 层隐藏特征矩阵，$\tilde{A}=A+I$ 为添加自环的邻接矩阵(A 为原始邻接矩阵，I 为单位矩阵)，$\tilde{D}$ 为 $\tilde{A}$ 的度矩阵，$W^{(l)}$ 为第 l 层权重矩阵，$b^{(l)}$ 为第 l 层偏置项，σ 为 ReLU 激活函数。

将科研主体关联图 G 的特征矩阵 $X_{node}=\left[X_{base},X_{time}\right]$ 输入网络，输出节点关联特征向量 $F_{rel}\in R^{256}$。

5) 特征融合单元

采用注意力机制加权融合时间特征与关联特征，自适应调整两类特征的重要性：

$$\begin{aligned}F_{att}&=\mathrm{Softmax}\left(W_a\cdot\left[F_{time},F_{rel}\right]+b_a\right)\\F_{fusion}&=F_{att}\cdot\left[F_{time};F_{rel}\right]\end{aligned} \tag{14}$$

其中，W_a 为注意力权重矩阵，b_a 为偏置项，$\left[F_{time},F_{rel}\right]$ 为特征拼接，$\left[F_{time};F_{rel}\right]$ 为特征堆叠，F_{fusion} 为融合后的特征向量。

6) 风险预测单元

采用两层全连接网络实现风险等级预测，输出科研主体的失信风险概率：

$$\begin{aligned}F_1&=\tanh\left(W_1F_{fusion}+b_1\right)\\p&=\sigma\left(W_2F_1+b_2\right)\end{aligned} \tag{15}$$

其中，W_1,W_2 为权重矩阵，b_1,b_2 为偏置项，p 为失信风险概率($p\in\left[0,1\right]$)。根据风险概率划分 3 个等级：低风险($p<0.3$)、中风险($0.3\le p<0.7$)、高风险($p\ge0.7$)。

7) 损失函数设计

采用交叉熵损失函数优化模型参数，考虑到失信案例样本不平衡问题，引入加权因子：

$$L_{pred}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right]\cdot w_i \tag{16}$$

其中，N 为样本数量，y_i 为样本真实标签(1 表示失信，0 表示守信)，p_i 为模型预测概率，w_i 为加权因子，通过 Adam 优化器最小化 L_{pred}。

8) 预警合约触发单元

基于合约层的预警智能合约，实现预警规则的固化与自动化执行：

① 预警规则存储：合约中预设风险等级与预警措施的映射关系；② 预测结果校验：将风险预测结果与链上历史数据进行一致性校验，确保预测结果基于可信数据；③ 预警发布：根据风险等级触发对应预警机制，通过应用层向相关管理部门推送预警信息；④ 日志上链：预警触发日志、核查结果日志经 PBFT 共识后上链存储，实现预警全流程可追溯。

3.3. 区块链五层架构设计

为支撑领域自适应语义增强检索算法与智能预警预测模块的高效运行，设计“数据层－网络层－共识层－合约层－应用层”五级深度协同架构，明确各层的技术实现方案、功能模块与与算法的衔接逻辑，确保架构与算法的深度耦合。

3.3.1. 数据层

作为算法输入输出的核心存储载体，采用链上核心 + 链下扩展的混合存储模式，按科研诚信信息敏感等级设计分层存储策略，保障敏感信息加密存储、公开信息可查可追溯，如表 1 所示：

Table 1. Comparison table of hybrid storage scheme and algorithm interface logic at data layer
表 1. 数据层混合存储方案与算法衔接逻辑对照表

存储类型	存储内容	技术选型	与算法的衔接逻辑
链上核心数据	1) 科研诚信数据哈希 2) 语义向量哈希 3) 数据关联 ID 4) 检索日志哈希 5) 领域词典哈希 6) 预警日志哈希 7) 核查结果哈希	Hyperledger Fabric	1) 算法预处理读取原始数据前，校验链上哈希一致性 2) 算法生成向量后，同步向量哈希上链 3) 算法检索时，通过关联 ID 匹配原始数据 4) 算法分词与词嵌入时，校验领域词典哈希一致性 5) 预警触发后，同步预警日志与核查结果
链下扩展数据	1) 结构化科研诚信数据 2) FAISS 向量库 3) 领域词典 4) 领域词典与术语映射表 5) 模型中间训练结果 6) 预警规则配置文件	① MongoDB ② FAISS 1.7.4	1) 算法预处理直接读取 MongoDB 数据 2) 算法向量检索访问 FAISS 库 3) 算法分词调用领域词典 4) 算法术语归一化调用术语映射表 5) 预警合约从配置文件读取预警规则

技术细节：全省 375 万条科研诚信数据按地域、类型划分为 20 个数据分片，配套 FAISS 向量库设 20 个向量分片，两类分片均在 3 个节点实现副本存储。向量库每小时增量更新，更新后向量哈希经混合共识上链，检索前自动校验链上一致性；模型训练数据需经合约层权限校验，仅授权节点可获取。链下核心数据异地多活备份并哈希上链，保障数据不丢失。针对敏感数据跨机构共享，引入联邦学习联合训练，核心节点部署可信执行环境处理协同计算，杜绝原始信息泄露，破解数据“不敢共享”痛点。

3.3.2. 网络层

基于电子政务外网构建 P2P 分布式网络，为算法请求提供安全、高效的传输通道，支撑算法的实时性需求。网络层主要包含节点部署、通信安全与负载均衡三大模块，具体设计如表 2 所示：

Table 2. Comparison table of network layer module technical design and algorithm interface logic
表 2. 网络层模块技术设计与算法衔接逻辑对照表

模块	技术设计	与算法的衔接逻辑
节点部署	1) 核心节点：3 台高性能服务器 2) 共识节点：地市各 1 台服务器 3) 向量节点：与共识节点对应，部署向量库 4) 接入节点：科研单位、管理部门接入终端	1) 算法查询请求优先路由至属地共识节点 2) 向量同步请求在核心节点与共识节点间传输 3) 用户检索请求经接入节点转发至共识节点
通信安全	1) 传输加密：TLS 1.3 协议 2) 身份认证：基于 X.509 数字证书 3) 路由机制：分布式哈希表 4) 数据完整性校验：CRC32 校验	1) 算法查询请求经 TLS 加密，防止传输过程被窃取 2) 向量节点通过证书认证后，方可参与向量同步 3) 传输数据经 CRC32 校验，确保完整性
负载均衡	1)请求分流：基于地域的负载均衡 2) 重试机制：请求失败时自动重试 3) 熔断机制：节点负载超过阈值	1) 算法高并发检索时，避免单节点拥堵 2) 确保算法检索请求成功率 ＞99.9% 3) 避免节点过载导致的检索延迟增加

技术细节：1) 检索请求传输路径：用户→接入节点→属地共识节点→向量节点→共识节点→接入节点→应用层→用户；2) 向量同步：核心节点每小时发哈希校验包，共识节点比对后，不一致则触发增量同步。

3.3.3. 共识层

针对科研诚信信息敏感程度差异，设计“PBFT + dBFT”动态权重混合共识机制，根据数据敏感等级自动调整共识节点参与范围与共识阈值，保障算法依赖的数据不可篡改与全网一致性。具体设计如表 3 所示：

Table 3. Comparison table of consensus mechanism design and algorithm interface logic
表 3. 共识机制设计与算法衔接逻辑对照表

共识类型	适用数据类型	参与节点	共识流程	与算法的衔接逻辑
PBFT	1) 非敏感数据哈希 2) 领域词典更新 3) 检索日志哈希 4) 非敏感向量哈希 5) 预警日志哈希	核心节点 + 共识节点	1) 提议节点发起请求 2) 验证节点校验合法性 3) 节点同意后，生效并上链	1) 算法检索的非敏感向量哈希经 PBFT 共识，确保一致性 2) 领域词典更新，防止算法分词错误 3) 检索日志经 PBFT 共识，确保可溯源
dBFT	1) 敏感数据哈希 2) 向量库增量更新哈希 3) 术语映射表更新	核心节点 + 属地共识节点	1) 提议节点发起请求 2) 所有参与节点验证 3) 共识生效并上链	1) 检索敏感向量哈希，提升安全性 2) 敏感数据向量更新，防止篡改 3) 术语映射表更新经 dBFT 共识，确保算法术语归一化准确性

技术细节：1) 向量库每小时增量更新、领域词典与术语映射表更新、算法检索日志生成，依数据敏感等级或场景分别触发 PBFT/dBFT 共识，预警触发后实时上链共识；2) 恶意节点篡改向量库数据会触发合约层篡改预警合约，自动切换备用向量节点保障检索准确；3) 通过动态权重分配，有效缩短共识延迟。

3.3.4. 合约层

基于 Hyperledger Fabric 智能合约框架，设计 6 类核心智能合约，将算法的“权限校验 - 流程执行 - 结果脱敏 - 日志记录 - 预警预测”全流程固化为智能合约逻辑，确保算法合规落地与安全可控。设计如表 4 所示：

Table 4. Comparison table of contract function design and algorithm interface logic
表 4. 合约功能设计与算法衔接逻辑对照表

合约类型	核心功能	与算法的衔接逻辑	执行耗时
权限管控合约	1) 定义三级角色权限 2) 校验检索请求权限 3) 控制向量库访问权限	1) 应用层调用该合约校验权限，通过则生成临时令牌 2) 向量节点凭令牌允许算法访问向量库 3) 读取敏感数据前，调用该合约校验解密权限	<0.1 s/次

续表

术语映射合约	1) 存储领域术语映射表 2) 提供术语归一化接口 3) 管理术语映射表更新	1) 预处理的术语归一化步骤，调用该合约接口 2) 词嵌入时，调用该合约获取术语标准向量	<0.05 s/次
检索流程合约	1) 自动化检索流程 2) 检索结果脱敏 3) 管理检索结果导出权限	1) 应用层发起检索请求后，自动驱动算法执行 2) 算法返回结果后，合约自动脱敏并记录日志 3) 结果导出前，调用该合约校验导出权限	<0.3 s/次
篡改预警合约	1) 实时比对向量哈希 2) 异常时触发预警 3) 隔离恶意向量节点	1) 算法检索前，该合约校验向量哈希一致性 2) 发现篡改时，算法自动切换至备用节点	<0.08 s/次
隐私计算合约	1) 联邦学习训练任务调度与权限管控 2) TEE 执行环境的准入与安全校验 3) 隐私计算结果的哈希上链与溯源 4) 跨机构协同计算的规则固化	1) 调用该合约调度联邦学习任务，确保数据不出本地 2) 敏感数据计算前，通过合约校验环境安全性 3) 计算结果经合约记录并上链，确保可追溯	<0.2 s/次

3.3.5. 应用层

基于 B/S 架构，将语义增强检索算法封装为可视化模块，与区块链其他功能协同，设计如表 5 所示：

Table 5. Comparison table of application layer functional module design and algorithm interface logic
表 5. 应用层功能模块设计与算法衔接逻辑对照表

功能模块	核心功能	与算法的衔接逻辑
共享查询模块	1) 提供多条件检索界面 2) 展示检索结果 3) 支持结果导出	1) 用户输入查询词后，模块调用检索流程合约 2) 合约驱动算法执行，返回结果后模块可视化展示 3) 导出前调用合约校验导出权限
信用管理模块	1) 科研诚信数据录入/修改/审核 2) 自动触发数据上链与向量生成 3) 显示数据上链状态与向量哈希	1) 录入数据，模块调用数据层接口存储，并触发算法生成向量 2) 向量生成后，模块显示链上哈希，供用户校验
系统管理模块	1) 用户角色配置 2) 向量节点状态监控 3) 检索日志查询	1) 配置角色后，模块同步更新合约权限 2) 监控向量节点状态，确保算法检索可用 3) 查询日志时，模块读取链上记录
统计分析模块	1) 基于算法检索结果生成报表 2) 多源数据关联分析	1) 模块调用算法进行批量检索，获取分析数据 2) 关联分析时，算法联动多数据分片向量库

交互流程：以“2022 年省级科技计划项目经费滥用案例”为例，应用层与算法、架构的交互闭环如下：

1) 数据层每季度自动同步新增的科研诚信数据(含失信案例、检索日志等)，生成增量训练数据集；

2) 模型训练节点经合约层权限校验后，从链下 HDFS 读取增量数据集，加载链上存储的历史模型参数；

3) 模型训练节点执行增量训练，优化模型参数，训练完成后计算参数哈希，发起 PBFT 共识；

4) 共识通过后，模型参数与哈希上链存储于合约层，更新链下模型文件；
5) 用户通过应用层风险管控模块发起风险预测请求；
6) 应用层调用领域自适应语义增强检索算法，获取该单位的历史数据与关联数据；
7) 智能预警预测模块加载链上可信模型参数，输入预处理后的特征数据，输出风险等级与概率；
8) 合约层调用预警规则合约，校验预测结果合理性，触发对应预警机制；
9) 应用层预警可视化模块展示预警结果、预测依据与演化趋势，风险管控模块分配核查任务。

4. 实验过程与结果分析

4.1. 实验环境与数据

4.1.1. 实验环境

Table 6. Comparison table of application layer functional module design and algorithm interface logic
表 6. 实验环境硬件/软件配置及支撑模块对照表

硬件/软件	配置详情	支撑的架构/算法模块
CPU	Intel Xeon Gold 6348	共识层、算法
RAM	512GB DDR4	数据层、算法
GPU	NVIDIA GeForce GTX 2080 Ti	算法、智能预警与预测分析
OS	Ubuntu 22.04 64bit	全架构
软件栈	Hyperledger Fabric 2.5 FAISS 1.7.4 MongoDB 6.0 PyTorch 2.0	全架构、算法向量检索、算法词嵌入、数据层存储、智能预警与预测分析

4.1.2. 实验数据

1) 基础数据：广东省 2020~2022 年科研诚信数据 375 万条，包括 127 万份科研人员档案、48 万条科研项目记录、3.2 万例科研失信案例；
2) 模型测试数据：广东省 2023 年 1~6 月科研诚信数据，包括 5000 例失信案例、20 万例守信样本；
3) 领域词典：构建科研诚信领域词典，包含 3200 个领域核心术语、1200 组同义术语映射关系。

4.1.3. 对比对象与核心指标

(1) 领域增强语义检索算法对比对象：
① 传统方案：MySQL 数据库存储 + 通用语义检索算法；
② 基础区块链方案：区块链五层架构 + 通用语义检索算法；
③ 本文方案：区块链五层架构 + 领域自适应语义增强检索算法。
(2) 智能预警与预测分析对比对象：
① 单一 LSTM 模型；② 单一 GCN 模型；③ 随机森林模型；④ 逻辑回归模型；⑤ 本文模型。
(3) 核心指标：
1) 架构–算法协同性；2) 安全性；3) 效率；4) 精度；5) 预警有效性。

4.2. 领域增强语义检索算法有效性验证

4.2.1. 总损失函数收敛曲线

如图 3 所示，总损失函数收敛曲线数据趋势：迭代 0~20 轮：L_{total} 从 1.8 快速降至 0.6，因 L_{sim} (语义

损失)显著降低；迭代 20~80 轮：L_{total} 缓慢降至 0.3，主要因 L_{hash} (哈希损失)优化(向量哈希一致性从 95%→99.9%)；迭代 80~100 轮：L_{total} 稳定在 0.3，收敛完成，此时检索准确率达 92%。总损失函数有效整合三大模块，实现参数协同优化。

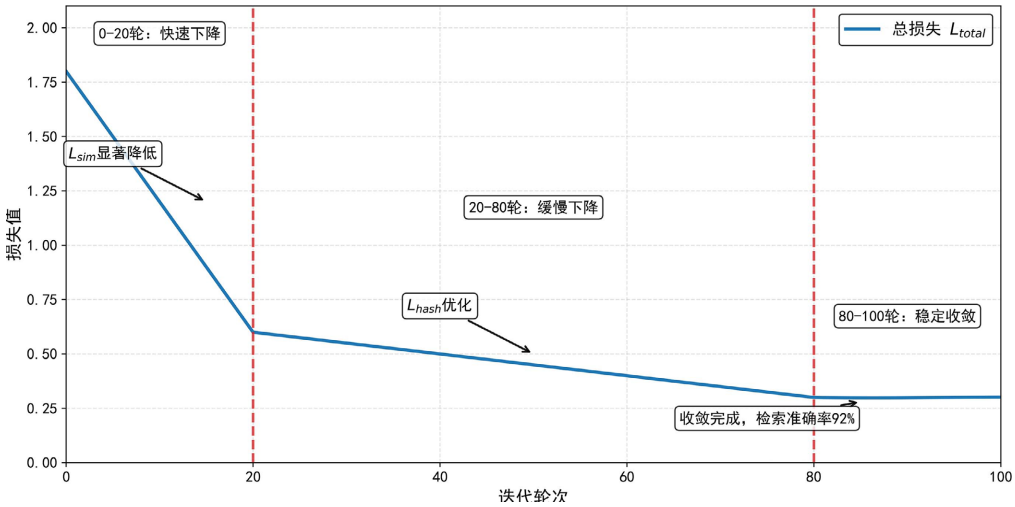

Figure 3. Convergence curve of the total loss function
图 3. 总损失函数的收敛曲线

4.2.2. 数据规模对检索性能的影响

从图 4 可见，随着数据规模从 100 万条增至 500 万条，三种方案响应时间均上升，但本文方案响应时间始终最低且增长极缓(最高仅 1.2 秒)；传统方案增速最快，500 万条时达 12.1 秒；基础区块方案增速与响应时间介于两者之间。结果表明，本文方案在大规模数据检索中性能更优、效率更稳定，能有效应对政务等场景下的海量数据检索需求。

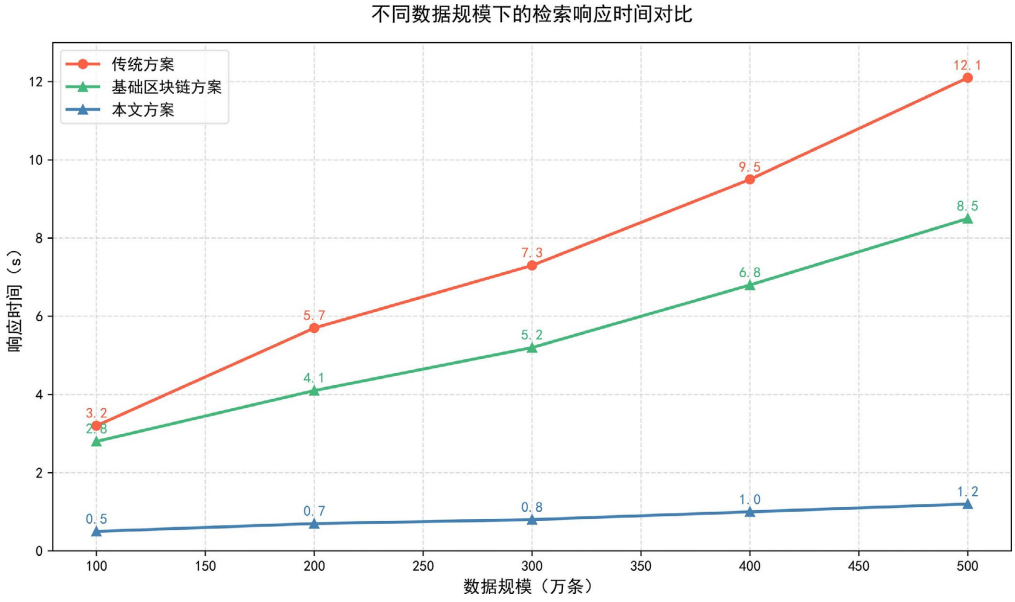

Figure 4. Comparison of retrieval response time under different data scales
图 4. 不同数据规模下的检索响应时间对比

4.3. 智能预警与预测分析模块实验结果

4.3.1. 模型性能对比实验

对比本文提出的模型与其他模型的预测性能，实验结果如表 7 所示。

Table 7. Model predictive performance comparison results
表 7. 模型预测性能对比结果

模型	准确率	精确率	召回率	F1 分数
逻辑回归	68.5%	62.3%	58.7%	60.4%
随机森林	76.2%	73.5%	70.1%	71.8%
单一 LSTM 模型	82.3%	79.8%	78.5%	79.1%
单一 GCN 模型	83.1%	80.2%	79.3%	79.7%
本文模型	89.7%	87.6%	86.9%	87.2%

结果表明，本文提出的模型在各项性能指标上均显著优于其他模型，其中准确率达 89.7%，F1 分数达 87.2%，本文模型同时捕捉科研失信行为的时间演化特征与主体关联特征，弥补了其他模型的局限性，LSTM 有效挖掘历史数据中的时序规律，GCN 充分利用主体关联关系，两者协同提升预测精度。

4.3.2. 预警有效性验证实验

基于广东省 2023 年 1~6 月的真实数据，测试预警系统的实际应用效果，结果如表 8 所示。

Table 8. Early warning effectiveness test results
表 8. 预警有效性测试结果

风险等级	预警数量	实际失信数量	误报率	漏报率	平均预警提前时间
高风险	826	710	14.0%	2.8%	45 天
中风险	1532	987	35.5%	5.3%	32 天
低风险	3215	412	87.2%	1.2%	20 天

实验结果表明，本文预警系统在高风险等级的预警效果最优，误报率仅 14.0%，漏报率 2.8%，平均预警提前时间达 45 天，为管理部门提供了充足的核查与处置时间；中风险与低风险等级的误报率相对较高，但漏报率控制在 5.3%以内，可通过后续核查流程进一步筛选。总体而言，预警系统能够有效识别高风险主体与项目，显著提升科研诚信管理的前瞻性与精准性。

4.4. 架构－算法协同性验证实验

4.4.1. 向量哈希一致性测试

Table 9. Vector hash consistency test results
表 9. 向量哈希一致性测试结果

测试场景	传统方案(无架构)	基础区块链方案	本文方案
正常状态(无篡改)	(无哈希校验)	98.5%	100%
10%向量篡改(模拟攻击)	(无防护)	89.2%	99.9%
30%向量篡改(模拟攻击)	(无防护)	72.6%	99.8%

上述实验结果如表 9 所示，传统方案无哈希校验，无法发现向量篡改；基础区块链方案未实现向量哈希实时共识，篡改后一致性下降明显；本文方案通过共识层实时同步向量哈希，即使 30%向量被篡改，仍能通过备用节点获取正确向量，一致性达 99.8%，验证数据层与共识层的协同有效性。

4.4.2. 权限管控准确率测试

测试不同角色用户的检索权限管控效果，验证合约层对算法合规性的支撑，实验结果如表 10 所示。

Table 10. Permission control accuracy test results
表 10. 权限管控准确率测试结果

用户角色	测试用例数	传统方案	基础区块链方案	本文方案
核心节点	200	(无角色区分)	97.5%	100%
地市共识节点	300	(无角色区分)	92.0%	100%
科研单位	500	(无角色区分)	88.6%	100%

传统方案无权限管控，所有用户可访问全量数据；基础区块链方案权限规则未完全固化，存在越权检索(如科研单位访问其他单位数据)；本文方案通过合约层固化权限规则，应用层严格调用合约校验，权限管控准确率 100%，验证合约层与应用层的协同有效性。

4.5. 消融实验

移除架构某一层的支撑功能，测试算法精度变化，量化各层对算法的贡献，如表 11 所示。

Table 11. Results of architectural layer ablation experiments
表 11. 架构层消融实验结果

实验组	检索准确率	较本文方案精度下降
本文方案(全架构)	92%	-
移除数据层领域词典	88.7%	3.3%
移除合约层术语映射	87.2%	4.8%
移除共识层向量校验	85.5%	6.5%
同时移除三层	76.1%	15.9%

数据层领域词典、合约层术语映射、共识层向量校验均对算法精度有显著贡献，其中共识层向量校验贡献最大(6.5%)，因向量篡改直接影响检索结果准确性；同时移除三层后，算法精度下降 15.9%，证明架构各层与算法的深度耦合是精度提升的关键。

5. 结论

本文针对科研诚信信息共享核心痛点，及现有技术在架构－算法协同、检索深度与安全管控上的不足，构建融合区块链与领域自适应语义增强检索的科研诚信智能管控平台。平台设计五级协同架构，结合混合存储、动态权重混合共识机制及联邦学习与可信执行环境保障信息共享安全高效；提出专属检索算法并融合预训练语言模型，实现多源异构数据深度关联与精准检索；智能预警与预测分析挖掘时序、关联特征，提升失信风险预测的前瞻性与准确性。实验表明，平台核心指标显著优于传统及基础区块链方案，大规模数据场景下仍高效稳定，为科研诚信信息跨域互联互通提供可落地解决方案。未来将探索

多链互联技术，深化相关技术融合，结合同态加密推动科研诚信管控向全域化、智能化升级。

基金项目

广东省重点领域研发计划项目“面向政务数据跨部门协同的区块链技术研究与应用”(2020B0101090004)。

参考文献

[1] 江利红, 罗仙凤. 论政府在科研诚信管理中的查处责任[J]. 中国软科学, 2019(11): 1-8.

[2] 邢文明, 陈继丽, 王张华. 面向科研诚信的科研数据管理保存: 逻辑关联、作用机制与实现策略[J]. 图书情报知识, 2021, 38(6): 134-143.

[3] 孙晶, 高燕, 杨尔璞, 等. 广东省科研信用管理体系建设现状、问题及对策建议[J]. 科技管理研究, 2022, 42(20): 197-203.

[4] 鲍锋, 李羿. 基于区块链技术的科研信息共享平台构建与运行机制研究[J]. 情报科学, 2022, 40(11): 72-77.

[5] 闫晴. 区块链赋能科研诚信管理的理论证成与制度创新[J]. 科技进步与对策, 2021, 38(23): 113-120.

[6] 张宇, 李静. 基于机器学习的科研失信风险预测模型研究[J]. 科研管理, 2022, 43(5): 245-252.

[7] 李阳, 王磊. 基于图神经网络的科研主体关联风险分析[J]. 数据分析与知识发现, 2023, 7(3): 89-96.

[8] Li, X., Li, J., Yu, F., Fu, X., Yang, J. and Chen, Y. (2021) BEIR: A Blockchain-Based Encrypted Image Retrieval Scheme. 2021 *IEEE* 24*th International Conference on Computer Supported Cooperative Work in Design* (*CSCWD*), Dalian, 5-7 May 2021, 452-457. https://doi.org/10.1109/cscwd49262.2021.9437677

[9] 王振江, 陈立松. 基于区块链技术的科研诚信平台架构设计[J]. 环渤海经济瞭望, 2020(11): 128-130.

[10] 郑荣, 张薇, 高志豪. 基于区块链技术的政府数据开放共享平台构建与运行机制研究[J]. 情报科学, 2022, 40(5): 137-143.

[11] 胡伏湘, 陈超群. 高校科研诚信存在问题的改进探讨——基于区块链技术的视角[J]. 中国高校科技, 2022(9): 23-27.

[12] 张雪媛, 都平平, 雷镭. 基于区块链技术的科学实验数据协同管理研究[J]. 情报杂志, 2022, 41(8): 149-155.

[13] Gao, Y., Xu, P., Yu, H. and Xu, X. (2024) A Novel Blockchain-Based System for Improving Information Integrity in Building Projects from the Perspective of Building Energy Performance. *Environmental Impact Assessment Review*, **109**, Article 107637. https://doi.org/10.1016/j.eiar.2024.107637

[14] Tan, J., Huang, Y., Huang, S., Hu, B., Zhang, W. and Dong, Y. (2024) Enhancing the Credibility of Data Trading through Blockchain-Enforced Semantic Analysis. 2024 4*th International Conference on Blockchain Technology and Information Security* (*ICBCTIS*), Wuhan, 17-19 August 2024, 289-294. https://doi.org/10.1109/icbctis64495.2024.00052

[15] Yao, C., Jiang, R., Long, L., Dong, J. and Wang, C. (2024) Blockchain-Based Secure Storage and Cross-Domain Sharing Mechanism for Medical Image Data. *Proceedings of the* 2024 8*th International Conference on Cloud and Big Data Computing*, Oxford, 15-17 August 2024, 33-40. https://doi.org/10.1145/3694860.3694865

[16] Liu, Y., Wang, X. and Zhang, H. (2023) Domain-Enhanced Semantic Retrieval for Government Open Data. *Journal of Information Science*, **49**, 456-472.

[17] 覃俊, 刘璐, 刘晶, 等. 基于 BERT 与主题模型联合增强的长文档检索模型[J]. 中南民族大学学报(自然科学版), 2023, 42(4): 469-476.

[18] Zhen, Z., Wang, X., Yang, X., Shu, J., Hu, J., Lin, H., *et al*. (2024) SemantiChain: A Trust Retrieval Blockchain Based on Semantic Sharding. *IEEE Transactions on Information Forensics and Security*, **19**, 10339-10354. https://doi.org/10.1109/tifs.2024.3488501