

Design and Implementation of an Intelligent Classification and Copyright Protection System for Academic Papers in Economic and Financial Fields Based on Blockchain

Tongxin Ma ^{1, a}, Haiyang Zhang ^{1, b}, Minze Ge ^{1, c*}, and Meimei Xia ¹

¹ School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China.

^a 23241199@bjtu.edu.cn, ^b 23241066@bjtu.edu.cn, ^c 23241076@bjtu.edu.cn.

Abstract. Aiming at the pain points of difficult confirmation of academic copyright, insufficient refined management, and low efficiency of manual classification in the economic and financial domain, this paper develops an academic paper management system that integrates blockchain technology and machine learning-enabled intelligent classification. Supported by Hyperledger Fabric, IPFS distributed storage and hash encryption, the system leverages blockchain to accomplish copyright depository, differentiated permission management and full-link traceability. The XGBoost model is adopted to conduct three-level intelligent classification of papers (Grades A, B, and C), which supplies automated rule input for the permission system and automates the whole workflow encompassing classification, depository and permission configuration. The layered collaborative architecture of off-chain rule computation and on-chain data depository resolves the performance bottlenecks and iteration dilemmas triggered by uploading business rules onto the chain in conventional solutions. The system caters to two user roles: end-users and auditors. Experimental results indicate that the system achieves a classification accuracy of 82.1%, a depository latency of less than 200 ms, and a 100% success rate of permission verification, thereby realizing the trustworthy and intelligent management of academic copyright.

Keywords: Blockchain; Machine Learning; Hyperledger Fabric; IPFS; Intelligent Classification; Copyright Protection

1. Introduction

With the rapid development of academic research in the economic and financial domain, the global number of papers published in this field exceeded 1.2 million in 2024. The massive volume of research outputs poses dual challenges for copyright protection and quality management: traditional centralized systems face the risks of data tampering and long cycles for copyright ownership confirmation, while manual classification suffers from low efficiency and strong subjectivity, making it difficult to support refined management.

In existing research, blockchain technology has provided a new path for copyright protection by virtue of its immutability feature. Zheng et al. [1] and Wang et al. [2] respectively proposed smart contract-based copyright transaction schemes and IPFS hybrid storage solutions. The MedRec system by Azaria et al. [3] also offered a reference for blockchain-enabled permission management. However, most existing schemes directly upload business rules onto the blockchain, ignoring the frequent iteration requirements of academic copyright rules and thus causing performance and iteration dilemmas [4].

On the other hand, machine learning-based automatic classification technology has been applied to paper quality assessment. Chowdhury et al. [5] verified the advantages of the XGBoost model in this task; Dien et al. [6] validated the effectiveness of textual features in classification; and Müller et al. [7] compared the performance of different machine learning methods in textual quality evaluation. Nevertheless, existing research solely focuses on improving classification accuracy, failing to achieve deep integration between classification results and copyright management, which cannot support automatic permission configuration.

To address the above issues, this paper designs an integrated system that integrates the two technologies. It leverages blockchain to realize trusted copyright protection and intelligent classification to enable automatic permission configuration. Meanwhile, the layered collaborative

mode of "off-chain rule computation + on-chain data depository" is adopted to resolve the performance bottlenecks of traditional schemes. At present, the core function verification of the system has been completed.

2. Overall System Design

2.1 Layered Architecture

2.1.1 Data Layer

A hybrid storage scheme integrating IPFS and blockchain is adopted. Large paper files are stored in IPFS, while core data including hashes, copyright metadata and classification results are anchored on the blockchain, balancing storage efficiency and data credibility.

2.1.2 Algorithm Layer:

This layer realizes three-level intelligent classification of papers (Grades A, B, and C) via machine learning, which automatically generates rule inputs for the permission system.

2.1.3 Blockchain Layer:

Built on Hyperledger Fabric, this layer completes copyright depository, smart contract execution and on-chain recording of operation logs, guaranteeing the trustworthiness and traceability of core operations.

2.1.4 Application Layer:

It provides interactive interfaces for users and auditors, supporting functions such as paper upload, access application and approval, while shielding the underlying technical details.

2.2 Classification-Permission Mapping Rules

The system formulates an in-depth mapping mechanism between paper classification and access permissions to facilitate differentiated copyright protection, with customized access regulations configured for papers of diverse quality levels:

Grade A: Top international journal papers; access needs dual approval for copyright security.

Grade B: Core journal papers; access requires only author authorization.

Grade C: General papers; open access without approval.

The aforementioned mapping rules can be formally depicted as the piecewise function below, defining the correlation between the classification result $f(p)$ and the access permission $\text{Permission}(p)$ of paper p :

$$\text{Permission}(p) = \begin{cases} \text{Dual Approval (Author + Auditor),} & \text{if } f(p) = A \\ \text{Single - level Approval (Author) ,} & \text{if } f(p) = B \\ \text{Open Access,} & \text{if } f(p) = C \end{cases} \quad (1)$$

where $f(p) \in \{A, B, C\}$ represents the output generated by the XGBoost classification model for paper p . The mapping rules are pre-configured in the backend service. Upon completion of automatic classification, the system can initialize access rules directly based on $f(p)$ without manual intervention, realizing end-to-end automated integration of classification and permission configuration.

3. Core Module Design

3.1 Machine Learning-Based Intelligent Classification Module

3.1.1 XGBoost Classification Model

To achieve automatic configuration of access permission rules, the system designs a lightweight XGBoost-based classification algorithm that strikes a balance between classification accuracy and

operational efficiency [5]. XGBoost eliminates the need for GPU acceleration and features low inference latency, which fully meets the lightweight design requirements of the classification module.

The XGBoost ensemble learning framework is employed to implement the three-class classification task. Its optimization objective function takes into account both the predictive loss and the regularization constraint on model complexity, with the specific mathematical expression provided as follows:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k). \quad (2)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 + \alpha \|w\|_1.$$

Where $l(y_i, \hat{y}_i)$ denotes the *multi-class cross-entropy loss*, T represents the number of leaf nodes, w is the leaf node weight vector, γ is the minimum gain threshold for node splitting, and λ and α are the $L2$ and $L1$ regularization coefficients, respectively. The explicit regularization terms constrain the model complexity and mitigate the risk of overfitting.

3.1.2 Model Input

The model input consists of two components: TF-IDF-based text vectors (with a maximum feature count of 8000 and an n-gram range of 1–3) and handcrafted linguistic and metadata features. These two feature sets are concatenated and fed into the XGBoost classifier [8]. Specifically, the handcrafted features cover eight categories including text statistics and lexical richness, all extracted from the paper's title and abstract. No external information directly correlated with the labels is introduced, ensuring feature independence. The metadata features include indicators such as the number of authors and the number of affiliated institutions, which reflect the academic background of the paper. The specific input model is as follows:

$$x = [x_{tfidf} \oplus x_{hand}] \in \mathbb{R}^{8000+d}. \quad (3)$$

Where x_{tfidf} denotes the text vector weighted by sublinear TF-IDF. The weighted formula is expressed as follows:

$$\{\text{TF} - \text{IDF}\}(t, d) = \log(1 + \text{tf}(t, d)) \times \log \frac{1+n}{1+\text{df}(t)} + 1.$$

And x_{hand} represents the handcrafted feature vector consisting of linguistic and metadata features.

To improve the discrimination capability between Grade A and Grade B, the system designs a combined score of academic rigor as the core discriminative feature, which integrates and weights six categories of rigor indicators. The formula is presented as follows:

$$S_{\text{rigor}} = 2.0 \cdot c_{\text{causal}} + 1.5 \cdot c_{\text{robust}} + 1.5 \cdot c_{\text{theory}} + 1.0 \cdot c_{\text{data}} + 1.0 \cdot c_{\text{precision}} + 1.5 \cdot c_{\text{discourse}}. \quad (4)$$

Where each component corresponds to the density of causal identification strategies, the density of robustness checks, the depth of theoretical modeling, the data authority signal, argumentation precision and academic discourse markers. The weights are determined according to the discriminatory contribution of each type of feature to distinguishing top journal papers.

3.2 Blockchain and Layered Collaboration Module

The system implements core blockchain capabilities based on the Hyperledger Fabric consortium blockchain [9]. Core smart contracts are developed in the Go programming language to guarantee execution efficiency, providing trustworthy support for copyright depository and access control. Two categories of core contracts are designed: the copyright depository contract writes paper hashes, classification results and author information onto the blockchain to complete copyright confirmation; the access approval depository contract stores approval logs and processes the depository of access applications. All operations are recorded on-chain throughout the process, featuring full traceability and immutability.

To tackle the pain point of embedding business rules on-chain in traditional architectures, this work designs a layered collaboration mode of off-chain rule computation and on-chain data depository, realizing the decoupling of business logic and data depository. The off-chain service layer deploys volatile business rules, including classification-permission mapping, access control and approval verification. Rule iteration only requires backend configuration updates without modifying on-chain contracts, drastically lowering iteration costs. The on-chain depository layer only undertakes the storage of core data to ensure data immutability.

The system adopts a unified backend scheduling architecture. Once classification is completed, the system can initialize access rules immediately and then invoke depository contracts to write data on-chain. The whole process of classification, depository and permission configuration is fully automated with no manual intervention.

4. System Testing and Evaluation

The experimental dataset comprises 6,000 annotated papers in the economic and financial domain, with 2,000 papers allocated to each of the three grades, covering diverse subfields. The dataset is partitioned into a training set (4,200 papers), a validation set (900 papers) and a test set (900 papers) at a ratio of 7:1.5:1.5. The model hyperparameters are set as follows: the number of trees is 800, the maximum depth is 6, and the learning rate is 0.03. Overfitting is restrained through L1 ($\text{reg_alpha} = 0.3$) and L2 ($\text{reg_lambda} = 2.0$) regularization, combined with subsampling ($\text{subsample} = 0.75$, $\text{colsample_bytree} = 0.7$). Training is monitored by the multi-class log loss of the validation set, with an early stopping patience of 50 epochs. The training procedure halts if the validation loss does not improve for 50 consecutive epochs.

Evaluation results indicate that the model achieves an overall accuracy of 82.1%, a Cohen's Kappa coefficient [10] of 0.732, and a macro-F1 score of 0.821, reflecting satisfactory classification consistency. The F1-scores of the three grades are 0.734 for Level-A, 0.732 for Level-B, and 0.997 for Level-C. The recognition performance of Level-C is nearly perfect, and the main confusion between Level-A and Level-B originates from boundary samples, which conforms to the expectation that adjacent grades present a natural transition interval in quality characteristics.

The latency of copyright depository is below 200 ms, and the correct execution rate of the classification-permission mapping rules reaches 100%, demonstrating the high efficiency of the system in copyright depository and permission configuration.

Table 1. Core Performance Test Results of the System

Test Indicators	Test Results	Description
Overall Classification Accuracy	82.1%	Classification performance of the machine learning-based classification module
F1-Score of Level-A Identification	0.734	Recognition performance of top journal papers
F1-Score of Level-B Identification	0.732	Recognition performance of core journal papers
F1-Score of Level-C Identification	0.997	Recognition performance of general papers
Cohen's Kappa Coefficient	0.732	Consistency between model predictions and annotated labels
Copyright Depository Latency	187 ms	Time consumption for on-chain copyright storage of a single paper
Rule Execution Success Rate	100%	Correct execution rate of classification-permission mapping

5. Summary

This paper proposes an academic paper management system integrating blockchain and machine learning-based intelligent classification. Blockchain provides a trusted foundation for copyright protection, addressing the pain points of tampering and inefficient copyright confirmation in traditional systems. The machine learning-based intelligent classification offers automatic rule inputs for the backend permission system, solving the low efficiency issue caused by manual configuration. Meanwhile, the layered collaborative architecture of "off-chain rule computation + on-chain data depository" resolves the dilemmas of excessive on-chain load and impeded rule iteration resulting from embedding business rules on-chain in traditional architectures. The system enables trusted, efficient and refined academic copyright management. The core functional test results verify the effectiveness of the system, indicating that the proposed solution can be directly applied to fields such as academic publishing and scientific research management.

However, the current research has limitations: the existing classification model and rule configuration are mainly targeted at academic papers in the economic and financial domain. In future work, the system can be extended to digital rights management scenarios in other fields to enhance its versatility.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.72271017).

References

- [1] Zheng Z, Xie S, Dai H N, et al. An overview of blockchain technology: Architecture, consensus, and future trends[C]//2017 IEEE International Congress on Big Data. IEEE, 2017: 557-564.
- [2] Wang S, Zhang Y. A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems[J]. IEEE Access, 2018, 6: 38437-38450.
- [3] Azaria A, Ekblaw A, Vieira T, et al. MedRec: Using Blockchain for Medical Data Access and Permission Management[C]//2016 2nd International Conference on Open and Big Data (OBD). IEEE, 2016: 25-30.
- [4] Zhang Y, Li X. A Survey of Blockchain-based Copyright Protection Research[J]. Computer Engineering, 2024,50(1): 1-12.
- [5] Chowdhury S, Schoen M P. Research Paper Classification using Supervised Machine Learning Techniques[C]//2020 Intermountain Engineering, Technology and Computing (IETC). IEEE, 2020: 1-6.
- [6] Dien T T, Loc B H, Thai-Nghe N. Article Classification using Natural Language Processing and Machine Learning[C]//2020 International Conference on Machine Learning and Data Engineering. 2020: 112-117.
- [7] Müller S, et al. Neural Networks or Linguistic Features? Comparing Different Machine-Learning Approaches for Automated Assessment of Text Quality Traits[J]. Journal of Educational Computing Research, 2024, 62(12): 2345-2372.
- [8] Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.
- [9] Androulaki E, Barger A, Bortnikov V, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains[C]//Proceedings of the 13th EuroSys Conference. 2018: 1-15.
- [10] Cohen J. A Coefficient of Agreement for Nominal Scales[J]. Educational and Psychological Measurement, 1960,20(1):37-46.