Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 1, 2025

Algorithm Optimization and Performance Improvement of Debt Enterprise Information Retrieval System in the Big Data Environment

Sheng Xu *

Hangzhou Yuancheng Technology Co., Ltd, Hangzhou, Zhejiang, 310013, China
*** Corresponding author Email:** xusheng133@126.com

Abstract: Against the backdrop of the big data era, the debt enterprise information retrieval system, as the core tool for financial risk management, is confronted with the challenge of processing massive heterogeneous data. The multi-source heterogeneity, high-frequency dynamics and concealed correlations of debt information lead to high data integration costs, difficult timeliness guarantee and insufficient penetration of deep risks, causing deviations in risk assessment and errors in the prediction of innovation potential. This paper reviews the existing technical solutions, deeply analyzes the characteristics of debt information and industry problems, including the shortcomings of insufficient capture of debt dynamics and lack of adaptability, discusses the combined optimization of distributed architectures such as Hadoop-ElasticSearch for massive data processing, and the hybrid BM25- neural embedding framework to improve the accuracy and efficiency of retrieval. As well as the application effectiveness and limitations of the RAG model in a multilingual environment, and evaluate key constraints such as insufficient intention understanding, weak cross-modal fusion, and lack of time sensitivity. The significance of this article lies in providing a comprehensive reference basis for the improvement of system performance, supporting real-time and accurate risk assessment and policy response, and indicating future research directions, including deepening the intention understanding model, developing cross-modal fusion mechanisms, and optimizing temporal dimension filtering.

Keywords: Big Data Environment; Debt Enterprise Information Retrieval; Algorithm Optimization; Performance Improvement; A Review of Information Retrieval Technology.

1. Introduction

Under the background of the big data era, the debt enterprise information retrieval system, as the core tool of financial risk management, is confronted with the challenge of processing massive heterogeneous data. With the implementation of China's "deleveraging" policy and the emergence of the pressure transmission mechanism of local government implicit debts, the demand for real-time and accurate enterprise debt risk assessment is increasingly urgent to support policy response and enterprise innovation decisions. However, the multi-source heterogeneity, high-frequency dynamics and covert correlation of debt information have led to significant bottlenecks in data integration, timeliness guarantee and risk penetration of the retrieval system, causing deviations in risk assessment and errors in the prediction of innovation potential. To this end, this paper systematically reviews the existing technical solutions, deeply analyzes the characteristics of debt information and industry issues, and discusses the application effects and limitations of optimization methods such as distributed architectures like Hadoop-ElasticSearch combinations, hybrid BM25- neural embedding frameworks, and RAG models. And evaluate key restrictive factors such as insufficient understanding of intentions, weak cross-modal fusion and lack of time sensitivity, aiming to provide a comprehensive reference basis for the improvement of system performance.

2. Industry Issues Analysis of Debt Enterprise Information Retrieval System

2.1. Characteristics and Retrieval Difficulties of Debt Enterprise Information

The information of debt enterprises presents significant characteristics such as multi-source heterogeneity, dynamic evolution and covert correlation. These characteristics jointly constitute the core difficulties of the retrieval system. From the perspective of data composition, debt information not only covers structured financial data (such as balance sheets and cash flow statements), but also contains a large amount of unstructured or semi-structured texts (such as contract terms, legal documents, credit rating reports, and news reports). The research on the structure of corporate debt capital by N.A. Dalisova et al. [1] indicates that evaluating debt efficiency requires integrating multi-dimensional and multi-format business operation data, which directly increases the complexity of data cleaning, standardization and correlation. Secondly, debt information is highly dynamic. The scale of enterprise debt, financing costs, guarantee status and risk ratings fluctuate frequently over time. The research on debt financing decisions of private listed companies by ZhiShan Xie et al. [2] confirmed this point. The Bayesian dynamic panel model they established emphasized the "inertia" feature of debt levels, highlighting the limitations of static data models and the urgent need for real-time or quasi-real-time data updates and indexing. Finally, the strong concealment of information correlation is particularly prominent. The complex guarantee chain among enterprises,

the transmission mechanism of local government implicit debt pressure to enterprises (such as the transfer of fiscal resources and the credit squeeze effect revealed by Lei Wang et al. [3]), and related party transactions make it extremely difficult to identify the real debt subjects and their risk exposures. It is required that the retrieval system has strong capabilities in deep relationship mining and graph computing.

In summary, the three major characteristics of multi-source heterogeneity, high-frequency dynamics and concealed correlations make the retrieval of debt enterprise information face key challenges such as high data integration costs, difficult timeliness guarantee and insufficient penetration of deep risks.

2.2. Main Problems in the Current Industry

The main problems in the current industry are mainly reflected in the insufficient capture and lack of adaptability of the debt enterprise information retrieval system to complex debt dynamics. Specifically, the research conducted by Lei Wang et al. [4] based on the data of Chinese listed enterprises from 2012 to 2018 indicates that the implicit debt pressure of local governments significantly reduces the total factor productivity of non-local government financing platform enterprises by transferring fiscal resources, strengthening tax collection and administration, and transferring credit resources. However, the existing retrieval systems often fail to effectively integrate such macro policy factors, resulting in deviations in the assessment of enterprise debt risks, especially in non-state-owned and small enterprises where the distortion effect is more prominent. Yu Du et al. [5] conducted a quasi-natural experiment based on China's "deleveraging" policy and found that the innovation level and achievements of enterprises with high default risk significantly improved after the implementation of the policy. However, the retrieval system lacked a real-time update mechanism, making it difficult to adapt to the rapid changes in debt default risk and affecting the accurate prediction of the innovation potential of enterprises. Furthermore, Hongqin Tang et al. [6] analyzed from the perspective of supply chain digitalization and pointed out that different digital flows have different impacts on debt costs. However, retrieval algorithms often ignore these subtle differences, resulting in an incomplete assessment of debt costs. These defects seriously restrict the reliability and practicability of the system and urgently need to be solved through algorithm optimization.

3. Review of Existing Technical Solutions

3.1. Information Retrieval Technology in the Big Data Environment

The core of information retrieval technology in the big data environment lies in addressing the challenges of rapid processing and precise query of massive and heterogeneous data. Distributed architecture has become the mainstream solution. Pan Gao et al. proposed an efficient construction scheme for an unstructured data retrieval system. The core of this scheme is to integrate multiple technologies based on the Hadoop framework: using HBase to achieve distributed storage of structured and unstructured data; Build distributed index technology with ElasticSearch to provide powerful full-text search capabilities; Combine IKAnalyzer for efficient Chinese word segmentation processing to enhance the semantic understanding of queries; And introduce the Redis in-memory database to cache hot data, significantly accelerating the real-time query response. This hierarchical collaborative architecture effectively solves the performance bottleneck of large-scale data retrieval. At the computational model level, for complex queries and correlation analysis, researchers are committed to optimizing parallel processing capabilities. Junchen Guo et al. [7] explored the use of MapReduce technology to analyze the correlation characteristics of large-scale data and achieve automatic query expansion and distributed computing. This method shows good scalability and computing efficiency in a multi-source heterogeneous data environment. These technologies, through the combination of distributed storage, parallel computing and intelligent indexing, provide key support for efficient information retrieval in the big data environment and lay the foundation for the subsequent exploration of optimization in the debt field.

3.2. Optimization Methods for Information Retrieval of Debt Enterprises

For the optimization of information retrieval of debt enterprises, researchers have proposed a variety of innovative methods to address the unique challenges in the big data environment. Anant Manish Singh et al. [8] proposed a hybrid information retrieval framework, which effectively combines the efficiency of the traditional BM25 algorithm with the semantic understanding ability of neural embedding technology. This framework utilizes the Transformer model for context weighting. Tests on the MS-MARCO and TREC-CAR datasets show that its recall rate and average accuracy have significantly improved by 25-30% and 12% respectively, and user satisfaction has also increased by 15-20%. It is particularly suitable for handling enterprise-level complex queries. Meanwhile, Kunying Li et al. [9] focused on the processing bottleneck of unstructured data and developed an optimization algorithm based on periodic data popularity and predefined category labels. This algorithm achieves effective filtering and precise sorting of massive unstructured data by intelligently associating the user's retrieval historical behavior patterns with the inherent category labels of files, providing key support for subsequent efficient analysis and edge computing. Furthermore, the multilingual RAG model optimization scheme proposed by Syed Rameel Ahmad et al. [10] effectively reduces hallucinations and erroneous responses by improving the data input strategy and real-time update mechanism, and enhances the retrieval accuracy and response speed in the multilingual enterprise environment. These methods jointly point to the core path for improving the accuracy and efficiency of the debt enterprise information retrieval system.

3.3. Limitations of the Technical Solution

The existing technical solutions still have significant limitations when dealing with the complexity of information retrieval for debt enterprises. First of all, insufficient understanding of intentions is the core bottleneck. The research of Hanseok Oh et al. [11] shows that traditional retrieval models often overly focus on literal queries while ignoring the deep intentions of users, resulting in limited accuracy of the results. Although the instruction aware model attempts to solve this problem, the experiments conducted by Oh et al. reveal that it may overfit and perform worse than the basic model in real and diverse search scenarios (such as multi-dimensional exploration of corporate debt risks).

Secondly, the ability of cross-modal information fusion is weak. Yan Gong et al. [12] 's evaluation of visual semantic embedding networks indicates that even in relatively mature tasks such as image-text retrieval, models struggle to capture complex scenes, latent objects, and deep semantics, with an average Recall@5 performance that is uneven and has obvious weak categories. This limitation, when mapped to debt enterprise retrieval, will seriously affect the precise correlation analysis of multi-source heterogeneous information containing unstructured data (such as scanned copies of financial reports, contract images). Furthermore, the contradiction of supervised learning in specific scenarios and the handling of time sensitivity are insufficient. Ilias Chalkidis et al. [13] found in the retrieval of regulatory documents that the neural resorter performed poorly due to the contradictory supervisory signals of the training data; Meanwhile, its research also highlights the key role of time-dimension filtering in improving the validity of the results, which is often not fully optimized by existing schemes. These limitations jointly restrict the performance of the system in key tasks such as dynamic risk assessment of debt enterprises and correlation mining of multi-source heterogeneous information.

4. Conclusion

This paper discusses the algorithm optimization and performance improvement of the debt enterprise information retrieval system in the big data environment, focusing on the core difficulties such as the multi-source heterogeneity, high-frequency dynamics and covert correlation of debt information. These characteristics lead to high data integration costs, difficult timeliness guarantee and insufficient penetration of deep risks. The existing industry problems include insufficient capture of debt dynamics and lack of adaptability, which lead to deviations in risk assessment and mispredictions of innovation potential. In terms of technical solutions, distributed architectures such as Hadoop combined with ElasticSearch optimize the processing of massive data, while innovative methods such as the hybrid BM25- neural embedding framework and RAG model enhance the accuracy and efficiency of retrieval. However, the existing schemes still have significant limitations, such as the overfitting risk caused by insufficient intention understanding, the weak cross-modal information fusion affecting unstructured data analysis, and the insufficient processing of time sensitivity restricting dynamic risk assessment.

The future technological outlook should focus on deepening the intention understanding model to avoid overfitting, developing a more powerful cross-modal fusion mechanism to integrate heterogeneous data such as text and images, and optimizing the temporal dimension filtering to capture the dynamic changes of debt in real time. Meanwhile, it is necessary to solve the contradictory signals in supervised learning, enhance the robustness and adaptability of the system in complex scenarios, and promote the evolution of debt enterprise information retrieval towards higher accuracy and real-time performance.

References

[1] Dalisova N, Martynova T, Eremeev D. EVALUATION OF THE EFFECTIVENESS OF THE DEBT CAPITAL STRUCTURE BASED ON COMMERCIAL ENTERPRISE DATA[J]. Socio-economic and humanitarian magazine, 2025, .

[2] Xie Z. Research on Bayesian Financial Panel Data Model Based on BP Neural Network[C]//2023 International Conference on Electronics and Devices, Computational Science (ICEDCS). , 2023: 645-648.

[3] Wang L, Wang C, Chen J, et al. The hidden crowding out effect: How does local government implicit debt pressure influence enterprise productivity in China?[J]. Managerial and Decision Economics, 2024, .

[4] Wang L, Wang C, Chen J, et al. The hidden crowding out effect: How does local government implicit debt pressure influence enterprise productivity in China?[J]. Managerial and Decision Economics, 2024, .

[5] Du Y, He Y. Deleveraging, Debt Default Risk, and Enterprise Innovation: Evidence from China[J]. Transactions on Economics, Business and Management Research, 2024, .

[6] Tang H, Zhu J, Li N, et al. Impact of Enterprise Supply Chain Digitalization on Cost of Debt: A Four-Flows Perspective Analysis Using Explainable Machine Learning Methodology [J]. Sustainability, 2024, .

[7] Guo J, Cui Y. Research on Big Data Retrieval System of Intelligent Computer AI in Post-Moore Era[C]//2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). , 2024: 1231-1235.

[8] Singh A M, Singh D B, Pandey A R, et al. Integrating Deep Learning Techniques in Information Retrieval: A Hybrid Approach to Relevance Optimization[J]. Journal of Information Systems Engineering and Management, 2025, .

[9] Li K, Qiao D, Li X, et al. Analysis and Optimization of Information Retrieval Algorithms for Unstructured Data[C]//International Conference on Computer and Information Application. , 2019.

[10] Ahmad S R. Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise[J]. ArXiv, 2024.

[11] Oh H, Lee H, Ye S, et al. INSTRUCTIR: A Benchmark for Instruction Following of Information Retrieval Models[J]. ArXiv, 2024.

[12] Gong Y, Cosma G, Fang H. On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval[J]. Journal of Imaging, 2021, 7.

[13] Chalkidis I, Fergadiotis M, Manginas N, et al. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations [J], 2021,: 3498-3511.