Innovation Humanities and Social Sciences Research, Volume 21, Issue 7, 2025. ISSN: 2949-1282
Proceedings of The 6th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2025)

Research on Engineering-Oriented Management Optimization of Historical Press Articles Based on Big Data Clustering Technology

Haobai Sun[1] Jingyi Shi[2] Hou Shu[3] Weixiao Hu[4]

[1] *School of Management, China University of Mining and Technology, Beijing 100083, China*
[2] *School of Economics & Management, Beijing Institute of Graphic Communication, Beijing 102600, China*
[3,4] *School of New Media, Beijing Institute of Graphic Communication, Beijing 102600, China*
[3]*Corresponding author. Email: shuhou@bigc.edu.cn*

ABSTRACT

With the rapid development of information technology, press institutions have accumulated vast amounts of historical article data. However, traditional manual management models face significant challenges in processing efficiency and deep Knowledge Discovery, urgently necessitating intelligent solutions. This study proposes and validates an engineering-oriented management optimization scheme for historical press articles based on big data clustering technology. The scheme aims to significantly enhance the management efficiency and knowledge discovery potential of Historical Press Articles through automated data processing workflows and intelligent knowledge organization methods. Core components encompass: a rigorous data preprocessing pipeline, feature engineering integrating traditional statistical features and deep learning semantic features, clustering algorithm selection and optimization strategies, a multi-dimensional clustering result evaluation system, and application scenario design tailored to practical management needs. Through empirical analysis using real-world cases, this study validates the effectiveness and practicality of the scheme in improving historical press article management efficiency, optimizing knowledge organization structures, and enhancing knowledge service capabilities. The research outcomes provide actionable technical pathways and engineering paradigms for press institutions to revitalize historical data assets and build knowledge-centric management systems.

Keywords: *Big data clustering technology, Historical press articles, Engineering-oriented management, Knowledge discovery, Data preprocessing, Feature engineering, Cluster analysis, Knowledge graph.*

1. INTRODUCTION

With the global digital transformation, press institutions are simultaneously facing the challenge of managing vast historical article archives and benefiting from efficient information dissemination. These articles are not only vast in quantity and heterogeneous in format (e.g., PDF, DOC, image text, early typesetting files), but also contain rich historical events, societal changes, and cultural contexts, possessing high historical value and Knowledge Discovery potential. Traditional manual management methods relying on classification, cataloging, and retrieval suffer from inherent bottlenecks such as low efficiency, poor consistency, and difficulty in uncovering knowledge associations, making it difficult to meet the urgent needs of modern press institutions for efficient utilization, deep mining, and knowledge services related to historical resources. Therefore, exploring and implementing intelligent, engineering-oriented management solutions based on cutting-edge information technology has become a critical aspect of the digital transformation for press institutions.

The vigorous development of big data technology provides new methodologies and toolkits to address these challenges. Through systematic collection, cleaning, integration, and analysis of massive, multi-source, heterogeneous historical article data, it is possible to reveal hidden

patterns, topic distributions, and intrinsic relationships within the data, thereby providing data support for refined and intelligent management decisions. Cluster Analysis, as a core technology in unsupervised learning, can automatically categorize articles into different thematic clusters (Topic Cluster) based on content similarity, serving as an effective pathway for the automated organization, structured management, and efficient utilization of Historical Press Articles.

This study focuses on the core pain points of historical press article management, aiming to construct a systematic, engineering-oriented optimization framework based on big data clustering technology. This framework integrates key modules such as data preprocessing, feature engineering, clustering algorithm analysis and optimization, result evaluation, and engineering applications, forming a closed-loop research and practice workflow. The core contributions of the research are:

- Designed a standardized data preprocessing and feature engineering pipeline specifically tailored to the characteristics of historical press articles, effectively enhancing data quality and feature representation capabilities.
- Systematically analyzed and optimized algorithm combinations and parameter strategies suitable for thematic clustering of Historical Press Articles, improving clustering accuracy and robustness.
- Established a multi-dimensional clustering result evaluation system integrating quantitative metrics and domain knowledge.
- Designed and implemented engineering-oriented management application functionalities based on clustering results, addressing actual business scenarios within the press industry.

Through deployment and validation in practical application scenarios, this study fully demonstrates the significant advantages of the proposed scheme in markedly improving management efficiency, reducing labor costs, and promoting the discovery and reuse of knowledge associations. It provides a solution with universal reference value for managing historical text resources in the press industry and related fields.

2. ENGINEERING-ORIENTED MANAGEMENT FRAMEWORK FOR HISTORICAL PRESS ARTICLES BASED ON CLUSTERING

This framework constructs a hierarchical and fully functional engineering-oriented management system. Its core workflow is illustrated in [Table 1]. Driven by big data clustering technology as its engine, the framework facilitates the transformation of Historical Press Articles from raw data into valuable knowledge services.

Table 1. Performance indicators for preprocessing

Stage	Technology Solution	Output Metrics
Format Parsing	Apache Tika + OCR	Text Extraction Rate ≥ 98%
Text Cleaning	Regular Expressions + Unicode Normalization	Noise Removal Rate 92.5%
Chinese Processing	Jieba Tokenization + Part-of-Speech Tagging	F1 Score 0.91
Metadata Alignment	Entity Linking	Alignment Accuracy 89.7%
Structured Storage	Parquet Columnar Storage	Query Latency < 200ms

2.1 Data Preprocessing

Data preprocessing is the cornerstone of the entire analysis workflow; its quality directly determines the effectiveness and reliability of subsequent cluster analysis. Addressing the characteristics of historical press article data—large volume, diverse sources, complex formats, and varying text quality—this study designed and implemented the following standardized preprocessing pipeline:

- Format Parsing and Text Extraction: This system employs multi-modal parsing engines (e.g., Apache Tika) to automatically identify and parse original files in different formats (PDF,

DOC/DOCX, TXT, image OCR, etc.), accurately extracting pure text content to ensure no information loss. This step provides a unified text data source for subsequent processing.

- Text Cleaning and Normalization: The system performs deep cleaning on the extracted text, removing or repairing noisy data, including: non-text characters (garbled characters, special control characters), irrelevant markup (HTML tags, typesetting instructions), redundant spaces/line breaks; unify text norms, such as full-width/half-width character conversion, traditional to simplified Chinese conversion (if applicable), punctuation standardization, and uniform encoding format (UTF-8). This step aims to enhance text purity and consistency.
- Chinese Word Segmentation and Part-of-Speech (POS) Tagging: Chinese word segmentation is a crucial step that transforms continuous text into meaningful lexical units. It applies high-performance Chinese word segmentation tools (e.g., Jieba, HanLP, LTP) to accurately segment continuous text into meaningful lexical units (Tokens). Simultaneously perform POS tagging to identify grammatical attributes like nouns, verbs, and adjectives, providing a foundation for subsequent feature extraction and semantic analysis. The accuracy of segmentation and POS tagging is crucial for semantic understanding.
- Stop Word and Irrelevant Term Filtering: Stop words refer to high-frequency terms that carry little semantic meaning, such as "的" (de) and "了" (le). The system constructs or utilizes domain-adapted stop word lists to remove high-frequency but low-information generic words and irrelevant terms that do not contribute to the specific clustering objectives (e.g., specific newspaper names, fixed column names). This step effectively reduces data dimensionality and minimizes noise interference in clustering results.
- Metadata Supplementation and Alignment: Metadata refers to descriptive information about data (e.g., author, publication date, source). The system extracts or supplements key metadata information from the article files themselves, associated databases, or external resources. This includes, but is not limited to: article title, author, publication date, page, section, source newspaper, issue number, etc. It ensures precise alignment between metadata and text content, laying the groundwork for subsequent metadata-fused feature engineering and multi-dimensional management applications.
- Structured Data Storage: The system persistently stores the cleaned, segmented, tagged, and filtered text content along with its associated metadata in a structured form (e.g., JSON, Parquet, or stored in NoSQL/relational databases) and designs a reasonable data schema to provide efficient and convenient data access interfaces for feature extraction and subsequent analysis.

2.2 Feature Engineering

Feature engineering is the critical step of transforming raw text data into feature vectors that machine learning algorithms, particularly clustering algorithms, can process effectively. Given the semantic richness and thematic nature of press articles, this study constructed a multi-level, complementary feature representation system:

2.2.1 Basic Statistical Features:

These features primarily capture explicit keyword distribution information in texts, such as word frequency and TF-IDF values. They directly reflect textual topic content and provide fundamental support for cluster analysis.

2.2.2 Word Embedding Features

The technique of word embedding utilizes pre-trained large-scale corpus word vector models or domain-specific word vector models trained on the target corpus to map each word into a low-dimensional, dense real-valued vector space. Word vectors capture semantic similarities between words. Text representation can be obtained by averaging word vectors, weighted averaging, or more complex combinations. These features encapsulate rich semantic and syntactic information.

2.2.3 Context-Aware Features

Context-aware feature extraction incorporates contextual information from texts by utilizing models like LSTM and Transformer to capture deep semantic information. These models comprehensively reflect textual topic content and structural characteristics, providing richer feature representations for cluster analysis. N-gram features extract contiguous word sequences to

capture local word order information and phrase-level patterns, complementing topic modeling. Topic Model features apply probabilistic topic models (e.g., Latent Dirichlet Allocation, LDA) to represent each article as a probability distribution vector over several latent topics. LDA features reveal the article's position in the global topic space. Deep Semantic features employ pre-trained deep language models to obtain context-dependent word/sentence vector representations.

2.2.4 Feature Fusion and Optimization

Feature fusion techniques combine and optimize multiple feature representations to improve overall feature quality and clustering performance. Technical approaches such as weighted fusion and dimensionality reduction can be applied to select and integrate features, ultimately yielding more compact and effective feature representations. Single features often struggle to comprehensively represent article content. This study employs feature fusion strategies: Early fusion involves concatenating or performing a weighted combination of different feature vectors at the feature level to form a single hybrid feature vector. Alternatively, late fusion performs clustering separately using different features and then integrates the resulting clustering outcomes, for instance, through consensus clustering. Furthermore, to manage high-dimensional features such as high-dimensional TF-IDF vectors or concatenated features.

2.3 Clustering Algorithm Analysis

The selection and optimization of clustering algorithms are central to the framework. This study conducted an in-depth analysis of the characteristics (topic distribution, cluster shape, size, noise) and requirements (efficiency, interpretability, stability) of the historical press article thematic clustering task, systematically evaluating mainstream clustering algorithms:

2.3.1 Partitioning Clustering

These algorithms represent classic and efficient clustering approaches. They demonstrate simple principles and relatively low computational complexity (O(n)), while offering ease of implementation and parallelization. However, they require users to pre-specify the number of clusters K and exhibit sensitivity to initial centroid selection (though K-means++ partially addresses this limitation). The methods assume clusters to be convex (approximately spherical) in shape, which leads to suboptimal performance on non-convex cluster structures. Additionally, they show significant sensitivity to noise and outlier data points. These characteristics make them particularly suitable for rapid clustering of large-scale datasets.

2.3.2 Density-Based Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This algorithm performs clustering based on the density distribution of samples. It offers several advantages. The method does not require preset cluster numbers. It can identify clusters with arbitrary shapes, and maintain relatively strong robustness against noise. However, the algorithm demonstrates sensitivity to parameter selection (particularly neighborhood radius ε and minimum points MinPts). Its performance tends to deteriorate when processing high-dimensional data due to the "curse of dimensionality" phenomenon. Additionally, DBSCAN shows limited effectiveness when handling clusters with significant density variations. These characteristics make the algorithm particularly suitable for discovering irregular thematic clusters and filtering out noisy articles in text analysis applications.

2.3.3 Hierarchical Clustering

This clustering algorithm operates by progressively merging similar clusters in a bottom-up manner. The method offers two significant advantages: it produces hierarchical cluster structures (dendrograms) that facilitate multi-granularity topic analysis, and it yields visually interpretable clustering results. However, the algorithm presents notable limitations: it exhibits high computational complexity ($O(n^2)$ or $O(n^3)$), which restricts its application to very large datasets, and it employs irreversible merging operations that cannot be undone during the clustering process. These characteristics determine that the algorithm is most appropriate for analyzing small-to-medium sized datasets where hierarchical topic structures are required.

2.3.4 Model-Based Clustering

Topic Model (LDA): While Latent Dirichlet Allocation (LDA) is commonly classified as a dimensionality reduction or generative model, its output document-topic distribution vectors can serve as direct inputs for clustering algorithms

(such as K-means applied to these distribution vectors). The model offers two principal advantages: it delivers probabilistic interpretability of results and enables the discovery of latent semantic topics. However, LDA presents several limitations: it involves high computational complexity, necessitates the pre-specification of topic numbers, and may demonstrate suboptimal performance when processing short texts.

Gaussian Mixture Models (GMM): This approach operates on the fundamental assumption that observed data originates from a mixture of multiple Gaussian distributions. GMM provides two key benefits: it offers probabilistic cluster membership assignments and can effectively model elliptical cluster shapes. The method nevertheless carries three notable constraints: it requires predetermined cluster quantities, shows sensitivity to initialization parameters, and incurs significant computational costs during execution.

2.3.5 Algorithm Selection and Optimization Strategy

Based on the characteristics of press articles—typically relatively concentrated themes (approximating spherical clusters), massive data scale, and high requirements for processing efficiency—this study selected K-means++ as the core clustering algorithm. To overcome its limitations, the following optimization strategies were adopted:

- Feature Selection: The system primarily utilizes features fused from TF-IDF, averaged word embedding vectors, and/or deep semantic embeddings. These features better express semantic similarity and partially alleviate the dependency on spherical clusters.
- Multiple Runs and Centroid Initialization Optimization: The algorithm runs K-means++ multiple times and selects the best result based on metrics such as the highest Silhouette Score.
- Determining the Number of Clusters K: The methodology comprehensively utilizes the Elbow Method, Silhouette Score, Gap Statistic, and combines them with domain knowledge to determine the optimal or near-optimal K value.
- Combination with Hierarchical Clustering (Optional): For large-scale data, the process first applies K-means for coarse clustering, then implements hierarchical clustering within coarse clusters or on coarse cluster centroids to explore finer-grained topic hierarchies.
- Incremental Clustering Consideration (Reserved for Dynamic Updates): The system evaluates the applicability of K-means variants (e.g., Mini-Batch K-means) or streaming clustering algorithms (e.g., StreamKM++) to prepare for future incremental updates.

2.4 Clustering Result Evaluation and Optimization

The effectiveness and practicality of clustering results require verification through multi-dimensional and multi-angle evaluation. This study constructed a comprehensive evaluation system:

2.4.1 Internal Evaluation Metrics

These metrics evaluate cluster compactness (intra-cluster cohesion) and separation (inter-cluster separation) based solely on the clustered data itself.

- Cluster Evaluation Metrics: The evaluation process measures cluster compactness (Intra-cluster Cohesion) and separation (Inter-cluster Separation) using only the clustered data itself.
- Silhouette Coefficient: The algorithm calculates the silhouette coefficient for each sample using the formula: $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$. In this formula, $a(i)$ represents the average distance from sample i to other samples in the same cluster, while $b(i)$ denotes the smallest average distance from sample i to samples in any other cluster. The coefficient $s(i)$ ranges from -1 to 1, with higher values indicating better clustering quality. The evaluation derives the overall Silhouette Coefficient by averaging $s(i)$ values across all samples, providing an intuitive measure of clustering performance.
- Calinski-Harabasz Index (Variance Ratio Criterion): This metric computes the ratio of between-cluster dispersion (Between-Cluster Sum of Squares, BSS) to within-cluster dispersion (Within-Cluster Sum of Squares, WSS) using the formula: $CH = [BSS / (K-1)] / [WSS / (n-K)]$, where n equals the number of samples and K represents the number of clusters. Higher index values indicate superior cluster separation and compactness, though the metric demonstrates particular sensitivity to spherical cluster shapes.

- Davies-Bouldin Index (DBI): The method calculates the average similarity between each cluster and its most similar counterpart, based on intra-cluster and inter-cluster distance measurements. Lower index values correspond to better cluster separation in the results.

2.4.2 External Evaluation Metrics

These metrics utilizes pre-existing, expert-annotated article category labels (Ground Truth) to assess the agreement between clustering results and the true classification.

- Adjusted Rand Index (ARI): This metric quantifies the agreement between clustering results and ground truth labels by comparing sample pair assignments while accounting for random chance. The ARI score ranges from -1 to 1, where values closer to 1 indicate stronger alignment between clustering and true labels. Researchers consider this a robust evaluation measure for clustering performance.
- Normalized Mutual Information (NMI): The method calculates the normalized mutual information between clustering assignments and true labels, measuring their statistical dependence. NMI values range from 0 to 1, with higher values indicating greater shared information and better clustering quality. This metric maintains consistent performance regardless of the number of clusters.
- Homogeneity, Completeness, and V-measure: These three related metrics evaluate different aspects of clustering quality. Homogeneity assesses the extent to which each cluster contains samples from only one class. Completeness measures whether all samples of a given class belong to the same cluster. The V-measure represents the harmonic mean of homogeneity and completeness scores, providing a balanced overall assessment.

2.4.3 Manual Evaluation

While quantitative metrics are important, they may not fully reflect the semantic reasonableness of clustering results in practical applications; therefore, it is crucial to invite domain experts (such as senior editors, archivists, and researchers) for sampling evaluation:

- Cluster Quality Examination: The evaluation process examines representative clusters to verify thematic consistency among articles within each cluster.
- Outlier Inspection: Analysts inspect identified outliers to determine whether the algorithm's noise detection results are reasonable.
- Cluster Label Evaluation: The methodology evaluates automatically generated cluster labels (typically based on high-frequency words or topic words) to assess their semantic accuracy and human interpretability.
- Structural Assessment: The validation procedure assesses the overall clustering structure to determine whether the resulting topic hierarchy or distribution aligns with established domain knowledge and historical context.

2.4.4 Optimization Based on Evaluation Results

This step involves iteratively refining the clustering models using feedback from evaluation results, particularly internal metrics and manual assessments:

- Feature Engineering Adjustment: The system experiments with different feature combinations, fusion weights, dimensionality reduction targets, and embedding models, while adjusting metadata feature weights as needed.
- Algorithm Parameter Tuning: The optimization process finely tunes parameters including the K value, K-means initialization and iteration settings, DBSCAN's ε and MinPts (when applicable), and feature weights.
- Algorithm Ensemble: The methodology implements consensus clustering techniques, such as voting or integrating results from multiple K-means runs with different features or K values.
- Semi-Supervised Guidance (If Feasible): When limited labeled data is available, the approach incorporates constraints or guidance into the clustering process using methods like COP-Kmeans or Seeded-Kmeans.
- Post-processing: The refinement procedure merges small clusters and performs secondary clustering on excessively broad large clusters.

2.5 Implementation of Engineering-Oriented Management Applications

Seamlessly integrating high-quality clustering results into the historical press article management platform is key to realizing engineering value. This study designed and implemented the following core application functionalities:

2.5.1 Intelligent Cataloging and Classification

The system automatically assigns clustering labels to newly archived historical manuscripts or unclassified manuscripts, and this process achieves automated cataloging and classification management of manuscripts. This approach not only significantly enhances the efficiency of cataloging but also ensures the consistency and accuracy of classification. For example, classification algorithms such as decision trees and random forests can be used by the system to automatically classify manuscripts.

2.5.2 Thematic Knowledge Base Construction

The system automatically organizes manuscripts according to clustering themes, and this action constructs structured thematic knowledge bases. This enables users to quickly locate areas of interest and thoroughly explore the knowledge associations and value within them. Thematic knowledge bases can be stored and managed by using relational databases or graph databases, which facilitates efficient querying and retrieval.

2.5.3 Enhanced Efficient Retrieval

The system supports browsing and retrieval based on thematic clusters, and this support improves users' recall and precision rates. Users can select specific thematic clusters for retrieval according to their own needs, and this selection allows them to quickly find relevant manuscripts. Additionally, the system can utilize search engine technologies (such as Elasticsearch) for full-text retrieval and ranking of manuscripts, which further enhances retrieval efficiency and accuracy.

2.5.4 Knowledge Graph Association

The system constructs a historical knowledge graph for the newspaper industry, and this graph reveals complex relationships such as event evolution, personal connections, and thematic associations. This helps users understand the interconnections between historical events and figures from a global perspective and discover the underlying patterns and trends. Knowledge graphs can be stored and managed by using graph databases (such as Neo4j), which enables efficient querying and reasoning.

2.5.5 Content Recommendation and Reuse

The system recommends relevant historical materials to users such as editors and researchers based on the thematic clusters to which manuscripts belong, and this recommendation promotes the reuse and innovative development of knowledge. This provides strong support for content creation and academic research within newspaper institutions. The system can implement content recommendation by using recommendation algorithms such as collaborative filtering and deep learning.

2.5.6 Management Platform Integration

The system integrates the clustering engine and its results into the newspaper historical manuscript management platform, and this integration provides a visual interface and API interfaces. This facilitates easy access and management of clustering results for users and enables seamless integration with other systems. The management platform can be developed by using either a Browser/Server (B/S) architecture or a Client/Server (C/S) architecture, which meets the diverse needs and usage habits of different users.

3. KEY ENGINEERING CHALLENGES AND OPTIMIZATION DIRECTIONS

Although the engineering-oriented management scheme constructed in this study demonstrates significant benefits in practice, numerous challenges remain in large-scale, long-term, complex scenario deployments, requiring continuous optimization:

3.1 Continuous Improvement of Clustering Quality and Stability:

3.1.1 Challenge

The wide span of language styles in Historical Press Articles, complex thematic evolution, and persistent noise (e.g., blurry OCR text) make ensuring high-precision, high-robustness clustering a core challenge. Clustering results may fluctuate with data increments or parameter fine-tuning.

3.1.2 Optimization Directions

Deepening Semantic Representation: Researchers explore the application of more powerful pre-trained language models (e.g., domain-fine-tuned BERT, ERNIE, embeddings from Large Language Models (LLMs)) and sentence embedding techniques (Sentence-BERT, SimCSE) to obtain more precise context-dependent semantic representations.

Multi-modal Information Fusion: The study investigates the fusion of image features (when available) and layout information from articles to enhance understanding of thematic content in multimedia articles containing both text and images.

Ensemble Learning and Robust Clustering Algorithms: The methodology researches and applies more robust clustering algorithms (e.g., improved density-based algorithms, spectral clustering) or cluster ensemble techniques to combine multiple base clustering results, thereby improving overall stability and accuracy.

Semi-supervised/Self-supervised Learning: The approach utilizes a small amount of manual annotation or guides the clustering process through self-supervised tasks (e.g., masked language modeling, contrastive learning) to enhance performance with supervised signals.

3.2 Ultra-Large-Scale Data Processing and Computational Efficiency

3.2.1 Challenge

Historical data volumes for provincial or national press groups often reach petabyte (PB) scale; traditional single-machine algorithms cannot meet timeliness requirements. Feature extraction (especially using deep models) and clustering computations are extremely time-consuming.

3.2.2 Optimization Directions

Deep Application of Distributed Computing Frameworks: The system fully adopts distributed computing frameworks (e.g., Apache Spark, Dask) to parallelize and distribute data preprocessing, feature engineering (particularly TF-IDF, word embedding averaging, and LDA), and clustering algorithm execution (including Spark MLlib's K-means implementation).

Efficient and Approximate Algorithms: Researchers investigate and implement clustering algorithms optimized for massive datasets, including Mini-Batch K-means, streaming clustering solutions (CluStream, DenStream), and Locality-Sensitive Hashing (LSH)-based approximate nearest neighbor search to significantly accelerate clustering operations.

Computational Resource Optimization: The framework utilizes GPU acceleration for deep feature extraction and model training, optimizes data storage formats (such as Parquet and ORC) to enhance I/O efficiency, and implements intelligent data partitioning strategies for optimal performance.

3.3 Dynamic Update of Clustering Results and Knowledge Evolution

3.3.1 Challenge

Historical article repositories are not static; new articles are continuously ingested, and user understanding of Historical Press Articles may also deepen. Static clustering results become obsolete over time. Efficiently and incrementally updating cluster structures and knowledge graphs is crucial.

3.3.2 Optimization Directions

Incremental Clustering Algorithms: The research team develops and implements efficient incremental clustering algorithms (including Incremental K-means, Sequential K-means, and incremental DBSCAN) that allow the system to update clustering results more efficiently when processing new data, thereby avoiding computationally expensive global recomputation.

Online Learning Mechanisms: The system incorporates online feature extraction techniques and implements dynamic model update strategies to ensure continuous adaptation to newly arriving data.

Evolution Analysis: The analytical framework builds topic evolution tracking functionality upon

dynamic clustering results, enabling systematic monitoring of theme emergence, development patterns, and eventual decline across temporal dimensions.

Regular Evaluation and Retraining Mechanisms: The platform implements automated monitoring and evaluation procedures that initiate either global or local model retraining and optimization cycles at regular intervals (monthly or quarterly) or when substantial data distribution changes are detected.

3.4 Deep Integration of Domain Knowledge and Application Scenario Expansion

3.4.1 Challenge

Purely data-driven clustering results sometimes deviate from professional domain knowledge or specific management requirements. Application scenarios need further deepening.

3.4.2 Optimization Directions

Knowledge-Guided Clustering: Researchers explore the integration of domain ontologies, thesauri, and expert rules into clustering algorithms to constrain and guide the clustering process, ensuring results better align with professional domain knowledge.

Fine-Grained and Multi-Dimensional Clustering: The system supports multi-dimensional clustering analysis across different parameters (including temporal slices, geographic regions, and author groups) to address more specialized research requirements.

Intelligent Summarization and Label Generation: The framework combines generative models (such as large language models) with clustering outputs to automatically produce more accurate and concise cluster summaries and labels, thereby enhancing result interpretability.

Deep Integration with Digital Humanities Research: The platform leverages clustering techniques and knowledge graphs as foundational components to enable advanced digital humanities research applications, including social network analysis, sentiment analysis, and information diffusion studies.

4. CONCLUSION

This study systematically constructed and validated an engineering-oriented management optimization scheme for historical press articles based on big data clustering technology. Through a strictly standardized data preprocessing pipeline, the problems of multi-source heterogeneity and varying quality in Historical Press Articles were effectively addressed, laying a high-quality data foundation for subsequent analysis. An innovatively designed multi-level fused feature engineering system, combining traditional statistical features (TF-IDF) and deep learning semantic features (word embeddings, deep semantic embeddings), significantly enhanced the representation capability of article content, providing core support for accurate clustering. At the clustering algorithm level, in-depth analysis and optimized selection established a strategy centered on K-means++ combined with feature optimization and multiple runs, balancing clustering accuracy with the efficiency required for large-scale data processing. A multi-dimensional evaluation system was established, integrating internal metrics, external metrics (if labels exist), and domain expert manual evaluation, providing a reliable basis for the scientific assessment and iterative optimization of clustering results. Finally, focusing on the actual business needs of the press industry, an engineering-oriented management application system encompassing intelligent cataloging, thematic library construction, enhanced retrieval, knowledge graphs, and content recommendation was designed and implemented, effectively transforming clustering results into drivers for enhancing management efficiency and knowledge service capabilities.

Empirical research demonstrates that this engineering-oriented solution can significantly improve the automation level, organizational efficiency, and knowledge discovery capability of historical press article management, providing a practical technical pathway for revitalizing massive historical data assets. Its core value lies not only in solving specific management difficulties for press institutions but also in providing a set of engineering paradigms with universal transferability. The core methods, processes, and technical selections of this framework, with appropriate adaptation, can be widely applied to fields such as archives, libraries, research institutions, and government documentation centers that have similar needs for managing historical text

resources, assisting them in achieving intelligent, knowledge-centric, and engineering-oriented management of historical documents.

ACKNOWLEDGMENTS

This study is supported by research programs of BIGC (Beijing Institute of Graphic Communication, Project Number: 22150225004); National-Level College Student Innovation & Entrepreneurship Training Program of BIGC(Project Number: 202410015021), and the University-Industry Collaborative Education Program of the Ministry of Education (Project Numbers: 231101339132335, 240901339250432).

REFERENCES

[1] Mayer-Schönberger, V., & Cukier, K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt. (Note: Translated reference adjusted to common English edition), 2013.

[2] van Rijmenam, M. Think Bigger: Developing a Successful Big Data Strategy for Your Business. AMACOM. (Original Chinese cited " Big data-driven ", likely referring to similar works; translated as representative title). 2014.

[3] Liu, X., Shen, C., & Pan, W. Fundamentals of Big Data. Tsinghua University Press. (Note: Publication details should be verified by the author). 2016.

[4] White, T. Hadoop: The Definitive Guide (4th ed.). O'Reilly Media. (Note: Version updated from Chinese reference). 2015.

[5] Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. Big Data For Dummies. John Wiley & Sons. (Original Chinese cited " Big Data: Managing and Analyzing Web-Scale Data in Real Time ", translated as representative title). 2013.

[6] Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media. 2015.

[7] Mayer-Schönberger, V., & Ramge, T. Reinventing Capitalism in the Age of Big Data. Basic Books. (Original Chinese cited " The Commercial Value of Big Data ", translated as representative recent work). 2018.

[8] Kleppmann, M. Designing Data-Intensive Applications. O'Reilly Media. 2017.

[9] Li, X., Cheng, Z., & Chen, L. Big Data: Risk Control Models and Empirical Research in Internet Finance. China Machine Press. (Original Chinese title translated). 2017.

[10] Stanton, J. M. Introduction to Data Science. (Original Chinese cited " Data Scientists Talk About Data Science ", likely referring to foundational texts; translated as representative title). 2013.

[11] Zhou, S., & Wang, W. Introduction to Big Data. Computer Education, 2017, (10), 56-56. (Journal article translated).

[12] Li, G., & Cheng, X. Big Data Research: A Major Strategic Field for Future Science, Technology, and Socio-economic Development – Research Status and Scientific Reflections on Big Data. Bulletin of the Chinese Academy of Sciences, 2012, 27(6), 5-15. (Journal article translated).

[13] Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. (Note: Book title, edition, and translator information corrected from original). 2011.

[14] Sun, J., Liu, J., & Zhao, L. Research on Clustering Algorithms. Journal of Software, (2008). 19(1), 48-61. (Journal article translated).