Exploration of Large Language Model Applications in Classification and Indexing: Postprint

Authors: Jiang Peng, Ren Yan, Zhu Beilin

Date: 2024-10-06T00:00:00+00:00

Abstract

[Purpose/Significance] Document classification and indexing is one of the fundamental tasks in libraries and other information institutions. Currently, limited human resources are insufficient to categorize the vast volume of documents. Large language models, with their exceptional natural language understanding and processing capabilities, have been employed to accomplish various natural language processing tasks such as text generation, automatic summarization, and text classification. Their integration with the entire document indexing process can help alleviate the pressure of classification indexing. [Method/Process] Drawing upon the long-term practical experience of the National Index to Chinese Newspapers and Periodicals, this study explores the integration of large language models into the classification indexing workflow through three entry points: reducing the reading burden on indexers, direct application of large language models for classification, and their combination with automatic indexing models, aiming to enhance indexing efficiency. [Results/Conclusion] Through a series of comparative tests and analyses, we designed a Prompt-assisted thematic classification model and an ACBKSY automatic indexing model. The Prompt-assisted thematic classification model helps indexers quickly grasp the key points of documents, thereby reducing reading pressure. The ACBKSY model achieved a 2.16% improvement in overall classification accuracy and a 3.77% improvement in non-rejection accuracy. Building upon these results, we optimized the actual indexing workflow, which has currently been implemented in the indexing of R and F category documents. The optimized workflow can improve indexing efficiency by 1.1 to 1.4 times.

Full Text

Preamble

Journal of Library and Information Science in Agriculture
Exploring the Application of Large Language Models in Classification

and Indexing Work

Authors: Jiang Peng, Ren Yan, Zhu Beiling (Shanghai Library, Shanghai 200030)

Abstract:
[**Purpose/Significance**] Document classification and indexing is one of the fundamental tasks of information institutions such as libraries. Currently, limited human resources struggle to categorize the massive volume of documents. Large language models (LLMs), with their exceptional natural language understanding and processing capabilities, have been applied to natural language tasks such as text generation, automatic summarization, and text classification. These capabilities can be integrated throughout the entire document indexing process, helping to alleviate the pressure of classification and indexing. [**Method/Process**] Drawing upon the long-term practical experience of the *National Newspaper Index*, this study explores how to introduce LLMs into the classification and indexing workflow from three perspectives: reducing the reading burden on indexers, directly applying LLMs for classification, and combining them with automatic indexing models to improve indexing efficiency. [**Results/Conclusions**] Through a series of comparative tests and analyses, we designed a Prompt-assisted topic classification model and the ACBKSY automatic indexing model. The Prompt-assisted topic classification model enables indexers to quickly grasp the key points of documents, reducing reading pressure. The ACBKSY model improved overall classification accuracy by 2.16% and non-rejection accuracy by 3.77%. Building on this, we optimized the actual indexing workflow, which has now been deployed for documents in the R and F categories. The optimized workflow can increase indexing efficiency by 1.1 to 1.4 times.

Keywords: Classification and indexing; Large language model; ERNIE Bot; GPT-4
CLC Number: G250.7
Document Code: A
Article ID: 1002-1248

Citation: Jiang P, Ren Y, Zhu B. Research on the algorithmic discrimination risks of AI embedded in government data governance and their prevention strategies[J]. Journal of Library and Information Science in Agriculture, 2024, 36(5): 32-42.

0 Introduction

Document classification and indexing involves the logical division and systematic arrangement of documents based on their content characteristics and certain external features to facilitate document management and utilization. It is a fundamental and critical task for libraries and other information institutions. Currently, document classification and indexing still relies primarily on manual work. However, in the information age, an ever-increasing volume of literature

continues to emerge, making it difficult for manual labor to categorize such massive quantities of documents.

The development of automatic indexing technology has witnessed an evolution from traditional machine learning methods to modern deep learning techniques. Early approaches included Decision Trees, Naïve Bayes, and Support Vector Machines, which gradually gave way to deep learning algorithms such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Bidirectional Encoder Representations from Transformers (BERT). Against this backdrop, automatic indexing technology emerged to better classify and utilize the growing body of literature through these advanced algorithms.

However, automatic indexing cannot fully replace document classification and indexing. The latter involves multiple stages including document review, classification rule determination, and final assignment of classification numbers. For indexers, the automatic indexing process is a black box, providing only results without truly participating in the entire indexing workflow. When accuracy has not yet reached a certain level, the usability of results becomes questionable, limiting improvements to indexing efficiency.

In November 2022, OpenAI launched the ChatGPT large language model, which gained widespread societal attention for its superior natural language understanding and processing capabilities. LLMs have been applied to natural language tasks such as text generation and automatic summarization. This study, based on the long-term indexing practice of the *National Newspaper Index*, explores how to leverage the strong semantic understanding and analysis capabilities of LLMs to integrate them into the existing indexing workflow, using document reading and automatic indexing as entry points to improve work efficiency.

1 Literature Review

Research on large language models primarily concentrates on three aspects: First, comparative analyses of core LLM technologies and their differences to identify application scenarios, such as proposing how to leverage LLMs in libraries for information consultation and other services. Second, utilizing the natural language processing capabilities of LLMs for specific business applications, including retrieval systems. Third, evaluating LLM performance across dimensions such as generated content quality, security risks, price, and time consumption.

Studies have shown that through structured information prompt templates containing example prompts and output format information, LLMs can achieve optimal classification performance in terms of precision, recall, and F1-score. Other research has explored using ChatGPT for entity recognition, pseudo-label generation, and training data generation, analyzing its usability from multiple perspectives.

2 Large Language Model-Assisted Document Reading

Classification and indexing essentially involves extracting key concepts that reflect the main content of a document and assigning them to corresponding categories according to classification rules. Document reading is the first and most fundamental step in this work. Document titles provide a brief summary of the theme and innovation points, generally containing background, research objectives, and content. The subject matter in titles and abstracts can be used to represent the full text' s research content.

With increasing document volumes and advancing technology, indexers face growing reading pressure. The primary way to improve work efficiency is to enhance reading efficiency. To address challenges faced by indexers, particularly new staff when reading documents outside their expertise—especially scientific literature—we designed a Prompt-assisted topic classification model using Baidu' s ERNIE Bot. This model intelligently analyzes and extracts document content, guiding the model to output concise information summaries that help indexers quickly grasp core points.

The model extracts research objectives and content, uses one sentence to summarize the current document, identifies professional terminology, and determines the document' s disciplinary category. Specific examples are shown in [Figure 1: see original paper] and [Figure 2: see original paper]. The quality of LLM outputs was manually evaluated by experienced indexers based on dimensions shown in , including grammatical correctness, content readability, and alignment with the original text. Evaluation results demonstrated that model-generated content has high readability and effectively extracts and summarizes core points from original documents.

Through the Prompt-assisted topic classification model, indexers can quickly understand the research theme, grasp the essence of key concepts and their interrelationships, and thus rapidly and accurately determine document classification positions. However, since LLM outputs may alter semantic relationships (such as causality or parallelism) present in the original documents, this content should only be used to reduce reading pressure and not as direct basis for classification.

3 Direct Application of Large Language Models for Classification

The ultimate goal of document classification and indexing is to assign appropriate category information. To investigate whether LLMs can be directly applied to text classification tasks based on the *Chinese Library Classification* (CLC), we conducted experiments with multiple models including ERNIE Bot and ChatGPT. We directly input document titles, abstracts, and other indexing information into the LLMs.

Test results were unsatisfactory. Without prior training, the optimal classifica-

tion accuracy was only around 30%. Three main issues emerged: First, output results varied significantly with different prompts. As shown in , even with identical prompts, the classification numbers provided by the model showed substantial differences. Second, LLMs cannot fully understand the CLC classification rules. For example, when directly asked "What category does R287 represent in the CLC?" , the model responded that it might be a more detailed subcategory not directly listed in the classification table. Third, LLMs have difficulty understanding explicit classification rules, such as those governing relationships between categories, leading to incorrect classification numbers.

4 Combining Large Language Models with Automatic Indexing Systems

Since LLMs cannot be directly used for CLC-based text classification tasks, we explored how to combine them with existing automatic indexing systems.

4.1 The *National Newspaper Index* Automatic Indexing Model (CBKSY)

The *National Newspaper Index* has long maintained a specialized team for document classification, primarily conducting classification and indexing of Chinese academic literature. This work has accumulated a large volume of high-quality indexing data, providing valuable information resources for research and a solid foundation of training data for automatic indexing technology.

Since the early 21st century, the *National Newspaper Index* has been exploring and utilizing automatic indexing tools. The main workflow is shown in [Figure 3: see original paper]. With technological development, the *National Newspaper Index* continues to optimize its automatic indexing model to improve accuracy and reliability.

4.1.1 Training Data CBKSY training data originates from manually indexed modern academic literature from the *National Newspaper Index*, primarily containing document titles, keywords, and classification numbers. To ensure dataset quality, the data underwent three rounds of cleaning, mainly focusing on standardizing newly added categories and unifying content. Simultaneously, K-fold cross-validation was employed to minimize bias potentially introduced by human subjective factors.

4.1.2 Category Level Selection Regarding category levels, CBKSY differs from previous approaches that set the deepest level at the fourth layer. Instead, based on practical work requirements, it adopts a "classify to the end" strategy, where the model directly outputs what it determines to be the most appropriate classification number.

4.1.3 Model Framework CBKSY employs deep classification algorithms and knowledge graph-enhanced graph neural network classification algorithms. The deep classification algorithm includes MacBERT, RobertaBERT, and BERT-Chinese variants. The graph neural network classification algorithm uses a knowledge graph as its core, representing entities appearing in academic literature (e.g., "Severe Acute Respiratory Syndrome" for R512.93) and their relationships. Through deep weighted ensemble algorithms, these approaches are fused together, adjusting voting weights for different models to enhance classification accuracy and robustness.

The model incorporates an uncertainty value evaluation strategy to accurately assess uncertainty during document indexing and determine whether human intervention is required. Uncertainty comprises two components: epistemic uncertainty (due to insufficient knowledge or unseen samples) and aleatory uncertainty (due to ambiguous category meanings or indexer subjective bias). The weighted combination of these uncertainties enables comprehensive judgment of document uncertainty, with high-uncertainty documents routed for manual classification.

4.2 Using Large Language Models for Classification Result Optimization

Analysis of CBKSY output results revealed that in cases where NUM1 (the highest probability classification number) was incorrect but NUM2 or NUM3 was correct—accounting for approximately 8% of total data—selecting the correct classification from non-NUM1 options could significantly improve automatic indexing accuracy.

4.3 Combining Large Language Models with CBKSY

4.3.1 Integration Approach The greatest advantage of LLMs lies in their information organization capabilities. While indexers subjectively believe that titles and abstracts can represent main document content, the input still contains considerable classification-irrelevant information or insufficiently precise information. Ideal automatic indexing models should input as many key classification-related elements as possible to reduce irrelevant factors.

This study tested using GPT-4 and ERNIE Bot 4.0 to regenerate titles and one-sentence abstracts, then input these generated results into the automatic indexing model. As shown in , both ERNIE Bot and GPT-4 generated content achieved slightly higher overall accuracy than the original documents, suggesting they could complement the original documents to some extent. Considering stability and performance, we selected ERNIE Bot to integrate with CBKSY, forming the new ACBKSY model.

The integration method involves inputting the document to be indexed into the LLM to generate new text En(i), then inputting both the original document

E(i) and En(i) into the indexing model. The classification weights are primarily determined by performance using the formula:

$$w_i = \frac{\text{num}_{E(i)}}{\text{num}_{E(i)} + \text{num}_{En(i)}}$$

where $\text{num}_{E(i)}$ and $\text{num}_{En(i)}$ represent the number of samples misclassified by the classifier under two different inputs, and num represents the total number of samples. The final probability of En(i) belonging to a certain category is determined by weighted voting.

4.3.2 Results and Analysis As shown in , with the default threshold of 0.8, all metrics of the automatic indexing results improved to varying degrees. Overall accuracy increased by 2.16%, and non-rejection accuracy improved by 3.77%, demonstrating the effectiveness of ACBKSY. When the threshold was set to 0.9, data accounting for 26.15% of the total volume could directly enter the verification stage for sampling inspection, balancing indexing quality and work efficiency.

In terms of stability, testing was conducted on medical category data collected by the *National Newspaper Index* from various months in 2023. The accuracy fluctuation range was controlled between 1% and 4%. Considering factors such as data distribution diversity and inter-indexer consistency, the model demonstrates good stability.

5 Conclusion and Outlook

This study, relying on large language models, designed a Prompt-assisted topic classification model to help indexers quickly understand document key points and reduce reading pressure. Through a series of experiments, we verified how LLMs can be combined with automatic indexing systems, resulting in the ACBKSY model. This model improved overall accuracy by 2.16% and non-rejection accuracy by 3.77%. Building on this, we optimized the actual indexing workflow, as shown in [Figure 5: see original paper], enhancing systematicity and coherence to ensure every step from document input to final classification is more efficient and accurate.

The optimized workflow has been deployed for documents in the R and F categories. Manual testing indicates it can increase indexing efficiency by 1.1 to 1.4 times. For secondary categories, classification accuracy improved from 85.79% to 90.53% after introducing LLMs.

However, this study has certain limitations. For instance, the LLM was not provided with sufficient learning to fully understand CLC category settings and some simple rule divisions. Since CLC classification is essentially hierarchical, guiding LLMs to gradually output classification results through multi-round dialogue requires further research. Subsequent work will continuously validate

and refine the workflow through practical application and gradually extend it to other categories to transform the classification and indexing ecosystem.

References

[1] Editorial Committee of the "Chinese Library Classification" of the National Library. Manual of Chinese Library Classification (5th Edition)[M]. Beijing: National Library of China Publishing House, 2012: 13.

[2] SHEN L L, JIANG P, WANG J. A study on the automatic classification of Chinese literature in periodicals based on BERT model[J]. Library Journal, 2022, 41(5): 109-118, 135.

[3] ZHANG Y H. A study of automated deep classification of literature based on Chinese Library Classification[J]. Library Journal, 2024, 43(3): 61-74.

[4] ZHAO X, DOU Z C, WEN J W. The development of information retrieval in the era of large language model[J]. Bulletin of National Natural Science Foundation of China, 2023, 37(5): 786-792.

[5] FU R X, YANG X H. Analysis of AIGC language models and application scenarios in university libraries[J]. Journal of Library and Information Science in Agriculture, 2023(7): 27-38.

[6] LIU X M, LI C Z X, WU S C, et al. A survey of text classification algorithms and application scenarios[J/OL]. Chinese Journal of Computers, 2024: 1-44. http://kns.cnki.net/kcms/detail/11.1826.TP.20240229.1608.002.html.

[7] SHI Y L, HE H Y. Research and practice progress of automatic indexing in China from 2003 to 2023[J]. Information Research, 2024(4): 120-127.

[8] WANG J J, YE Y, WANG W R. A prospective analysis on ChatGPT-type AI-GPT technical applications for changing library information processing[J]. Library Theory and Practice, 2024(1): 122-127, 136.

[9] WEI X, CUI X Y, CHENG N, et al. ChatIE: Zero-shot information extraction via chatting with ChatGPT[J/OL]. arXiv Preprint, arXiv: 2302.10205, 2023.

[10] MENG X Y, CHEN Y, BAI H Y. Research on intelligent generation of structured review for retrieval result set[J]. Library and Information Service, 2024, 68(6): 129-141.

[11] XU Z W, LI H L, LI B, et al. A survey of AIGC model evaluation: Enabling technologies, vulnerabilities and mitigation[J/OL]. Journal of Frontiers of Computer Science and Technology, 2024: 1-34. http://kns.cnki.net/kcms/detail/11.5602.tp.20240523.1947.002.html.

[12] RONG L. A research on prompt learning of large language models for automated book classification[J]. Research on Library Science, 2024(1): 86-103.

[13] ZHANG H P, LI L H, LI C J. ChatGPT performance evaluation on Chinese language and risk measures[J]. Data Analysis and Knowledge Discovery, 2023,

7(3): 16-25.

[14] LI J C, XIAO C L, QIN X T, et al. Text-relation-extraction algorithm based on large-language model and semantic enhancement[J]. Computer Engineering, 2024, 50(4): 87-94.

[15] ZHANG Y Y, ZHANG C Z, ZHOU Y, et al. ChatGPT-based scientific paper entity recognition: Performance measurement and availability research[J]. Data Analysis and Knowledge Discovery, 2023, 7(9): 12-24.

[16] LI C X, WANG Z Q, ZHOU G D. LLM enhanced cross domain aspect-based sentiment analysis[J/OL]. Journal of Software, 2024: 1-16. https://doi.org/10.13328/j.cnki.jos.007156.

[17] YANG D J, HUANG J T. A Chinese scientific literature annotation method based on large language model[J/OL]. Computer Engineering, 2024: 1-7. https://doi.org/10.19678/j.issn.1000-3428.0068400.

[18] ZHAO L, ZHANG C Z. Difference analysis of research topics in a specific domain based on different content levels[J]. Journal of Library and Information Science in Agriculture, 2021, 33(5): 14-27.

[19] JIANG P. A case study of the BERT model based on Chinese Library Classification and influence factors[J]. Library Science Research & Work, 2022(5): 43-48.

[20] HE L, LIU J, HOU H Q. An analysis of the impact factors in the multi-layer automatic classification based on CLC[J]. Journal of Library Science in China, 2009, 35(6): 49-55.

[21] LUO H Y, LIU W. Research on massive literature indexing based on semantic hierarchy granularity[J]. Information Studies (Theory & Application), 2024, 47(5): 194-203, 193.

[22] ZHANG Z X, ZENG J X, XIA C J, et al. Information resource management researchers thinking about the opportunities and challenges of AIGC[J]. Journal of Library and Information Science in Agriculture, 2023, 35(1): 4-28.

Note: Figure translations are in progress. See original paper for figures.