Advancements in Information Retrieval: Enhancing Query Expansion and Ranking with CORAG

Abjal Hussain Shaik
11643584
abjalhussainshaik@my.unt.edu

Hari Krishna Sai Rachuri
11682376
harikrishnasairachuri@my.unt.edu

Tharun Ramula
11706360
tharunramula@my.unt.edu

***Abstract*—Effective information retrieval is necessary for managing massive academic datasets but traditional information retrieval systems have trouble with retrieval accuracy, query expansion, and ranking optimization In order to improve document retrieval in scholarly research this research introduces Chain of Retrieval Augmented Generation (CORAG), a novel method that mix vector based semantic retrieval (FAISS) adaptive query expansion and BM25 based keyword search. We assess CORAG's ability to retrieve pertinent research publications using text-based searches using the ArXiv Research publications dataset. To improve interpretability and relevance, our method makes use of hybrid retrieval models, hierarchical indexing, and effective ranking algorithms. Here we evaluate the retrieval performance of CORAG against the conventional BM25 and TF IDF models, showing improvements in precision, recall. Overall the optimization of vector based indexing lightweight query expansion and quick document ranking here our approach helps to lowers search latency and improves retrieval accuracy This work will offer a workable answer to contemporary research information retrieval problems by bridging the gap between scalable IR approaches and effective academic article search**

I. Introduction

Because of the huge expansion of digital publications, the efficient and precise IR systems are crucial in fields such as scholarly information access, document retrieval and academic research. Mainly in extensive academic databases, retrieval performance and domain specific query interpretation are problems wth traditional Information retrieval approaches.

- **Computational Inefficiency**: Static term frequency models these are the foundations of standard retrieval techniques such as BM25 and TF IDF, which cause performance issues when working with massive academic datasets like ArXiv. Even neural retrieval models experience O(N) complexity problems while working with large document embeddings, which reduces search efficiency.
- **Challenges in Contextual Understanding**: Traditional Information retrieval models like BM25 and TFIDF, are not able to capture the context and understanding of research queries beacause they rely on keyword matching. This frequently leads to low retrieval accuracy when dealing with the intricate scholarly jargon or changing research patterns.
- **Domain-Specific Limitations**: Knowledge of domain specific terminology and the semantic connections between research concepts are nedded for effective retrieval in academic databases. In scholarly domains like physics and ML and medicine conventional models produce less than ideal search results due to their inability to handle scientific jargon, synonyms, and lengthy research inquiries.

To address these issues, we need CORAG, a dynamic query reformulation and retrieval framework that helps to improve IR by utilizing:

- Context aware transformations are used to refine search queries through adaptive query expansion.
- 2. BM25 (keyword based retrieval) and FAISS (vector based retrieval) are mixed/combined in hybrid retrieval models to improve ranking accuracy
- 3. Semantic reranking and hierarchical indexing these are used in efficient document ranking to order tio increase retrieval speed and relevance.

Enhancing retrieval accuracy optimizing ranking effectiveness and improving scholarly information access in large scale research databases are our goals when integrating CORAG in academic IR pipelines..

II. Literature Review

Domain specific retrieval flexibility, and query efficiency have all been enhanced by recent developments in IR. But retrieval accuracy, scalability, and dynamic query reformulation remain significant shortcomings of current methods. Here, we provides an overview of five important research that constitute the basis of our suggested CORAG framework.

- Seismic: Efficient Indexes for Approximate Retrieval (SIGIR 2024) Contribution: This paper introduces a technique called seismic an inverted index which is optimized for learned sparse representations. It achieves a latency of sub millisecond retrieval. Limitations: It is ineffective for complex search tasks because of lacking query reformulation mechanisms.
  Furthermore, its sparse indexing design restricts academic search's ability to retrieve semantic information.

- Inquire: A Natural World Text to Image Retrieval Benchmark (NeurIPS 2024) Contribution: A large scale multimodal dataset known as iNaturalist 2024 is developed for text to image retrieval it highlights problems in semantic search across modalities.
  Limitations: There is no adaptive query expansion, which lowers accuracy because queries are static. It does not have domain-specific retrieval enhancements (such as technical or legal searches).
- AD-DRL: Attribute-Driven Disentangled Representation Learning for Multimodal Recommendation (ACM MM 2024) Contribution: AD-DRL is proposed, which improves retrieval interpretability and robustness by decoupling semantic components in multimodal data.
  Limitations: Not generic IR, although recommendation systems are the main focus. lacks query reformulation, which reduces its usefulness for changing search requirements.
- CaseLink: Inductive Graph Learning for Legal Case Retrieval (SIGIR 2024) Contributions: Increases the accuracy of legal case retrieval by using graph neural networks (GNNs) to record case-to-case links.
  Limitations: Limited real-time applications due to high processing expense. Furthermore, its relevance to academic document retrieval is limited because it was created especially for legal case searches.
- GenQREnsemble: Zero-Shot LLM-Based Query Reformulation (JIR 2024) Contribution: Offers a zero-shot large language model (LLM) framework for query reformulation and enlargement called GenQREnsemble.
  Limitations: The inability to optimize for domain-specific retrieval (e.g., legal, medical) It has trouble with long-tail searches, which reduces the accuracy of retrieval.

A. Identifies Gaps in Existing IR Techniques

Although there are many advancements, existing IR methods still have major shortcomings:

- **Absence of Reformulation of Dynamic Queries**: In academic literature retrieval, the majority of retrieval algorithms do not iteratively refine queries, which results in less-than-ideal search results.
- **Large-Scale Retrieval Computational Inefficiencies**: While graph-based retrieval models (such as CaseLink) increase accuracy, they frequently have significant latency, which makes them unsuitable for real-time document retrieval.
- **Restricted Domain Adaptability**: A lot of current IR models have trouble with searches that are technical, scientific, or domain-specific (such as physics or AI research). Synonyms, contextual linkages, and domain-specific terminology are not captured by traditional models.

CORAG overcomes these gaps by combining query expansion, hybrid retrieval, and improved ranking strategies.

III. Proposed Methodology

This study focuses on creating and assessing the Chain-of-Retrieval Augmented Generation (CORAG) framework, which makes use of sophisticated ranking methods, hybrid retrieval models, and machine learning-based query expansion. Enhancing retrieval efficiency and accuracy in extensive academic search systems is the goal of the methodology.

A. Datasets

The datasets used for this research guarantee a thorough assessment of CORAG's retrieval capabilities in text-based academic paper searches..

- **ArXiv Research Papers Dataset**: The datasets used for this research guaranteeThe ArXiv Research Papers Dataset is perfect for assessing academic search performance because it includes scientific research papers with metadata (titles, abstracts, and full-text PDFs).
- It enables testing the efficacy of query expansion, ranking enhancements, and the accuracy of large-scale document retrieval. a thorough assessment of CORAG's retrieval capabilities in text-based academic paper searches.
- **Custom PDF and Text Data**: A dataset of academic documents and scientific papers in PDF format is used to evaluate CORAG's PDF-based retrieval capabilities.
- The system's capacity to extract, process, and retrieve data from full-text PDFs is the basis for its evaluation.

B. Tools and Technologies

To effectively process and retrieve research articles, CORAG combines a number of machine learning frameworks and information retrieval tools:

- Python: Used to preprocess textual input and integrate retrieval models.
- An effective vector-based retrieval package for quick nearest-neighbor searches in huge academic datasets is called FAISS (Facebook AI Similarity Search).
- Hugging Face Transformers: Large language models (BERT, T5) that have been pre-trained for semantic search refinement and query extension.

- A keyword-based retrieval model called BM25 (Whoosh) is utilized in conjunction with FAISS for hybrid IR.
- Text is transformed into dense embeddings for FAISS semantic search using sentence transformers.

C. Information Retrieval Techniques

- **Query Reformulation**:Produces more semantically relevant search words by utilizing BERT and T5 to improve user queries.
- **Hybrid Retrieval**:enhances ranking accuracy by combining vector-based semantic search (FAISS) with keyword-based retrieval (BM25).

D. Machine learning models used:

Query processing, embedding generation, and retrieval efficiency are handled by machine learning models.

- Query Reformulation Models: BERT, T5—improve user queries by producing more semantically rich query expansions.
- Sentence Transformers is a text embedding model that transforms text into vector embeddings for FAISS semantic search.

E. Evaluation Metrics

- Mean Average Precision (MAP) : It evaluates precision at several recall levels to determine retrieving accuracy.
- Normalized Discounted Cumulative Gain: this is used to evaluate highly relevant documents in search results.
- Precision Recall: Assures high relevance by calculating the number of pertinent documents that show up in the top 5 search results.

IV. Expected Outcomes

- Increased Retrieval Accuracy: Dynamic query reformulation results in increased recall and precision. Expanding the query guarantees higher-ranked search results.
- Improved Search Latency: FAISS-based nearest-neighbor search and effective indexing enable quicker document retrieval.
- Better Query Expansion and Ranking Efficiency: By dynamically refining searches and efficiently ranking results, CORAG improves search relevancy.
- Improved User Experience: By enhancing intent identification, query expansion eliminates the requirement for precise keyword matching

V. Conclusion

In order to improve the query expansion, ranking effectiveness and document retrieval in scholarly research this project presents CORAG which is an enhanced retrieval system. This incorporates BM25, FAISS, and large language model driven query reformulation to improve search relevance and retrieval accuracy as demonstrated by its evaluation on the ArXiv Research Papers dataset. CORAG offers an effective and scalable solution for academic IR by bridging gap between keyword based and vector based retrieval

References

[1] S. Bruch et al., "Efficient inverted indexes for approximate retrieval," in *Proc. ACM SIGIR*, 2024.

[2] E. Vendrow et al., "INQUIRE: A Natural World Text-to-Image Retrieval Benchmark," *arXiv preprint arXiv:2411.02537*, 2024.

[3] Z. Li et al., "Attribute-driven Disentangled Representation Learning for Multimodal Recommendation," in *Proc. ACM MM*, 2024.

[4] Y. Tang et al., "Caselink: Inductive graph learning for legal case retrieval," in *Proc. ACM SIGIR*, 2024.

[5] K. Dhole et al., "GenQREnsemble: Zero-shot LLM ensemble prompting for generative query reformulation," in *Proc. ECIR*, 2024.