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

CSCE 5200 Information Retrieval and Web Search

Abjal Hussain Shaik, Hari Krishna sai Rachuri, Tharun Ramula

Yuan Li, Ph.D.

Computer Science and Engineering Department

Outline

- Abstract (optional)
- Introduction
- Literature Review Overview
- Gaps in Current Research (Motivation)
- Proposed Methodology
- Expected Outcomes
- Timeline/Milestone
- Conclusion
- References

Abstract

- Research Focus: Introduces CORAG, a novel approach to improve academic information retrieval.
- Problem: Traditional IR models struggle with query expansion, ranking, and retrieval accuracy.
- Solution: Uses hybrid retrieval (BM25 + FAISS) and BERT/T5 for query expansion.
- Impact: Enhances search relevance, speed, and ranking efficiency in large academic datasets.

Introduction

- Feature Engineering in Information Retrieval:

Feature engineering involves selecting and transforming data attributes to improve retrieval accuracy. It plays a crucial role in refining search results and optimizing ranking models.

- Motivation:

Traditional IR models struggle with context understanding, query expansion, and ranking optimization. Enhancing these aspects is vital for efficient academic research and large-scale information access.

Introduction (cont.)

- Relevance:

With the exponential growth of digital scholarly content, improving retrieval techniques ensures faster, more relevant, and scalable search systems.

- Research Question:

How can hybrid retrieval models (BM25 + FAISS) and dynamic query expansion (BERT/T5) enhance precision, ranking efficiency, and scalability in academic information retrieval?

Literature Review Overview

Paper	Contribution	Relevance
Seismic (SIGIR 2024)	Optimized inverted index for fast retrieval.	Improves query efficiency but lacks reformulation.
Inquire (NeurIPS 2024)	Text-to-image retrieval benchmark.	Highlights semantic search challenges.
AD-DRL (ACM MM 2024)	Multimodal recommendation using disentangled representations.	Improves retrieval robustness but not general IR.
CaseLink (SIGIR 2024)	GNN-based legal case retrieval.	Enhances retrieval but high computational cost.
GenQREnsemble (JIR 2024)	Zero-shot LLM-based query expansion.	Advances IR but lacks domain-specific optimization.

Literature Review Overview (cont.)

- Trends in Information Retrieval

Hybrid Retrieval: Combining keyword-based (BM25) and vector-based (FAISS) for better ranking.

Query Expansion: LLMs (BERT/T5) improve search relevance. Semantic

Search: Embedding-based models enhance contextual understanding.

Scalability Challenges: Large-scale datasets require efficient indexing methods.

Gaps in Current Research

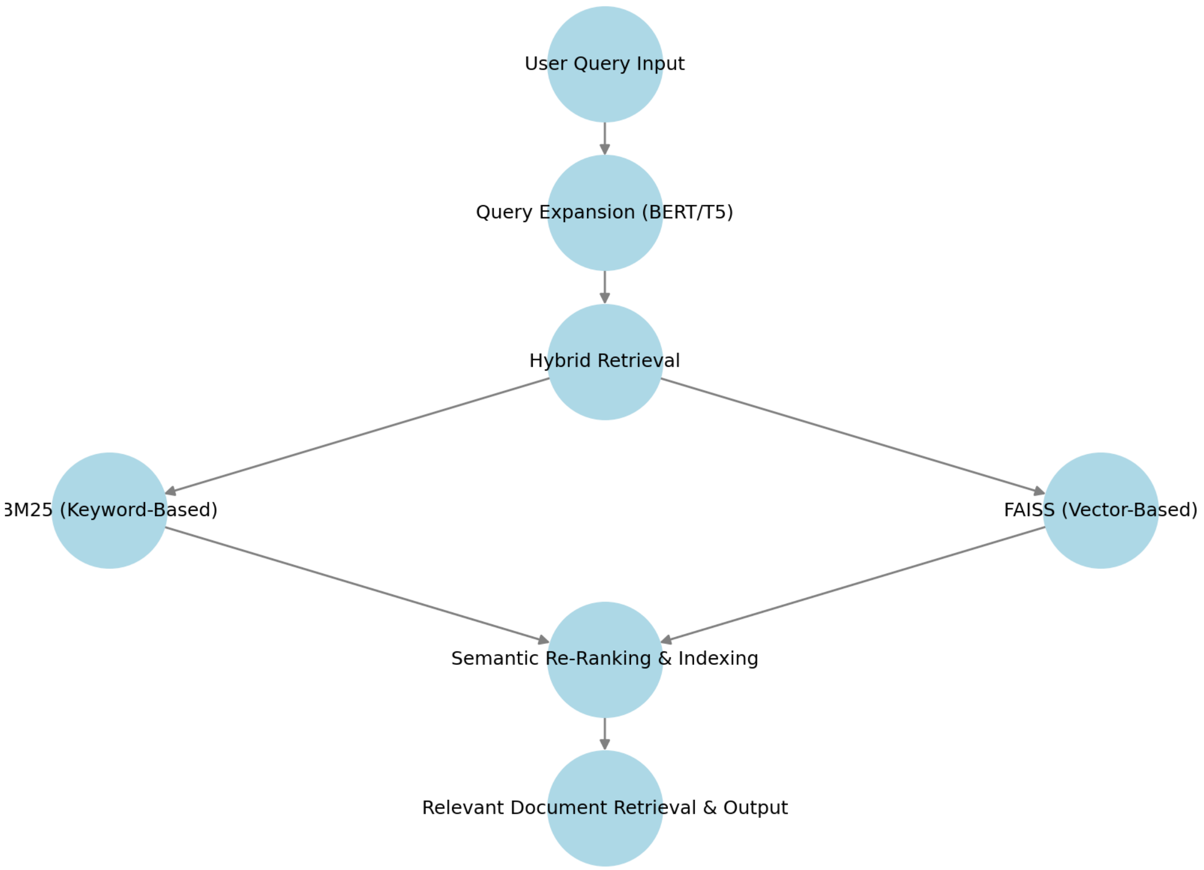
- Gaps: No dynamic query expansion, high retrieval latency, weak semantic understanding, poor domain adaptability.
- Query Expansion: BERT/T5 enhances search relevance.
- Hybrid Retrieval: BM25 + FAISS improves accuracy and efficiency.
- Optimization: Semantic reranking + hierarchical indexing boost speed.

Proposed Methodology

- Dataset: ArXiv Research Papers (titles, abstracts, full-text PDFs) for academic search evaluation.
- Retrieval Techniques: Hybrid approach using BM25 (keyword-based) + FAISS (vector-based) with BERT/T5 query expansion.
- Evaluation: MAP, NDCG, and Precision-Recall to measure accuracy, ranking efficiency, and retrieval speed.

Proposed Methodology (cont.)

CORAG Workflow Flowchart



Expected Outcomes

- Improved Model Performance

Hybrid retrieval (BM25 + FAISS) + BERT/T5 enhances ranking and relevance.

Better semantic understanding improves query expansion and retrieval accuracy.

- Higher Retrieval Accuracy

Increased precision, recall, and NDCG scores validate model effectiveness.

Adaptive query expansion ensures more context-aware search results.

Expected Outcomes (cont.)

- Faster & Scalable Search Efficiency

Hierarchical indexing + FAISS-based retrieval reduces search latency.

Enables real-time document retrieval, even for large academic datasets.

- Transforming Information Retrieval

Bridges keyword-based and vector-based retrieval, improving search relevance.

Helps researchers access relevant papers faster, enhancing knowledge discovery.

Timeline

Task	Deadline	Description
Literature Review	Week 1-2	Analyze recent IR techniques, identify gaps.
Dataset Collection & Preprocessing	Week 3	Acquire ArXiv dataset, clean and format data.
Query Expansion Implementation	Week 4	Implement BERT/T5 for query refinement.
Hybrid Retrieval Setup	Week 5	Integrate BM25 (Whoosh) & FAISS for search.
Semantic Re-Ranking & Indexing	Week 6	Optimize ranking with hierarchical indexing.
Model Evaluation	Week 7-8	Measure MAP, NDCG, Precision-Recall.
Optimization & Fine-Tuning	Week 9	Improve latency, retrieval accuracy.
Results Analysis & Validation	Week 10	Compare performance with baseline models.
Final Report & Presentation	Week 11	Document findings, prepare presentation.

Conclusion

- Key Points Recap: CORAG enhances query expansion, retrieval accuracy, and ranking.
- Significance: Improves academic search efficiency using hybrid retrieval + AI models.
- Expected Contribution: Bridges gaps in semantic understanding and search scalability.
- Call to Action: How can this approach be further optimized for real-world academic IR systems?

References

- Bruch, S., et al. (2024). Efficient inverted indexes for approximate retrieval. Proceedings of ACM SIGIR 2024.
- Vendrow, E., et al. (2024). INQUIRE: A Natural World Text-to-Image Retrieval Benchmark. arXiv preprint arXiv:2411.02537.
- Li, Z., et al. (2024). Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. Proceedings of ACM MM 2024.
- Tang, Y., et al. (2024). CaseLink: Inductive Graph Learning for Legal Case Retrieval. Proceedings of ACM SIGIR 2024.
- Dhole, K., et al. (2024). GenQREnsemble: Zero-shot LLM Ensemble Prompting for Generative Query Reformulation. Proceedings of ECIR 2024.