Grenze International Journal of Engineering and Technology, June Issue

Hybrid Retrieval Systems: Combining Graph-Structured and Embedding-based Models for Efficient Document Processing

Jeevanantham G[1], Sanmuga Priya M[2] and Barakkath Nisha U[3]
[1-3]Sri Krishna College of Engineering and Technology/M. Tech CSE, Coimbatore, India
Email: jeevananthamg73@gmail.com, sanmugapriyam, barakkathnisha}@skcet.ac.in

***Abstract*— In the current realm of information retrieval, the adoption of hybrid machine learning models offers a significant opportunity to enhance document processing and knowledge-driven question answering. This paper presents the development of a Hybrid Retrieval System that effectively merges graph-based and vector-based retrieval methodologies. The system leverages HuggingFace embeddings alongside Neo4j graph databases to improve the precision and relevance of information extracted from large document collections. A key element of our strategy involves constructing both a knowledge graph and a vector store from unstructured documents, employing semantic chunking to break down text into significant segments. We propose a robust retrieval framework that combines vector search with knowledge graph navigation, thereby optimizing the retrieval of contextually pertinent answers. Additionally, the integration of Contextual Knowledge Distillation (CKD) facilitates the extraction and compression of relevant subgraphs in response to user queries, thereby enhancing overall retrieval accuracy. The system's capabilities are demonstrated through various use cases, highlighting improvements in response accuracy, document indexing efficiency, and the speed of knowledge retrieval. This hybrid methodology not only streamlines the retrieval process but also provides scalability, flexibility, and adaptability across diverse sectors, including legal research and healthcare. The results emphasize the potential of the Hybrid Retrieval System to transform the construction, querying, and distillation of large-scale knowledge bases tailored to specific domain requirements.**

***Index Terms*— Hybrid Retrieval, Knowledge Graph, Semantic Chunking, Machine Learning (ML), Retrieval Augmented Generation (RAG).**

I. INTRODUCTION

Efficient retrieval of information across diverse data sources is a critical challenge in today's data-driven landscape. Traditional retrieval models that address either structured or unstructured data in isolation often struggle with the scale and complexity of modern queries, leading to inefficiencies in processing and suboptimal outcomes. In contrast, hybrid retrieval systems offer a transformative solution by fusing graph-based and vector-based approaches to extract highly relevant insights.

Graph-based retrieval methods excel at uncovering the intricate relationships among entities, enabling a deep contextual analysis through knowledge graphs. Conversely, vector-based retrieval employs machine learning embeddings to process large volumes of unstructured text, facilitating rapid and salient query responses.

Grenze ID: 01.GIJET.11.2.1083_22

By integrating these complementary techniques, the proposed hybrid framework advances document processing capabilities and meets the diverse needs of complex information queries.

This paper introduces a comprehensive Hybrid Retrieval System that leverages both Graph Retrieval Augmented Generation (Graph RAG) and Vector Retrieval Augmented Generation (Vector RAG). The system is further enhanced through semantic chunking—a method that segments lengthy documents into contextually meaningful pieces—and query-specific knowledge distillation, which focuses on extracting the most relevant parts of the knowledge graph via subgraph extraction. A custom retriever dynamically selects the optimal retrieval method based on the unique features of each query, thereby maximizing both accuracy and efficiency.

By combining these advanced techniques, the proposed solution not only improves retrieval precision and contextual understanding but also enhances scalability and response times. This unified approach is especially valuable in technical domains such as healthcare, finance, and technical writing, where complex and multidimensional queries demand both structured and unstructured data analysis.

II. Literature Review

Sarmah et al. introduce HybridRAG, an innovative methodology that seamlessly fuses Vector Retrieval Augmented Generation (VectorRAG) with Graph-based Retrieval Augmented Generation (GraphRAG) for enhanced information extraction from unstructured financial documents. Their framework capitalizes on the strengths of semantic similarity retrieval via vector embeddings and the deep relational insights provided by structured knowledge graphs. By demonstrating improved retrieval accuracy and reduced hallucinations—especially in the context of financial earnings call transcripts—the study positions HybridRAG as a pivotal tool for not only the financial sector but also other domains requiring precise and contextually enriched information extraction.

Yuan et al. present an advanced Hybrid RAG system designed to augment the reasoning capabilities of large language models (LLMs). Their work integrates classic retrieval-augmented generation with enhanced reasoning mechanisms that draw on both external knowledge bases and knowledge graphs. Extensive evaluations on the CRAG benchmark illustrate the system's prowess in handling multistep reasoning and complex query aggregation. The modular design and significant improvements in retrieval latency underscore its potential applicability across diverse fields such as medical diagnosis, legal documentation, and financial analysis.

Zheng et al. introduce KGBERT, a novel adaptation of the BERT model tailored for knowledge graph completion tasks. By recasting KG completion as a sentence classification challenge, KGBERT leverages BERT's deep contextual understanding to predict missing entities and relationships effectively. Validated on benchmark datasets like FB15k237 and WN18RR, the model demonstrates how integrating pretrained language representations with structured graph paradigms can substantially enhance prediction accuracy. Its adaptable nature across multiple domains further reinforces the model's value in advancing knowledge graph methodologies. Lewis et al. propose the Retrieval Augmented Generation (RAG) framework, which combines dense retrieval techniques with generative models to address knowledge-intensive NLP tasks. Their dual-encoder architecture—comprising one encoder for retrieving relevant passages from extensive corpora and another for generating informed responses—ensures that external knowledge is incorporated in a scalable, efficient manner. Evaluations on datasets such as Natural Questions and TriviaQA reveal improvements in factual consistency and response relevance, while the approach also successfully mitigates issues like hallucinations common in conventional language models.

Collectively, these studies underscore a significant evolution in retrieval-augmented systems. By blending vector-based retrieval with graph-centric methods and infusing advanced reasoning capabilities, these approaches demonstrate a robust solution for extracting pertinent information from complex documents. This integrated perspective not only enhances accuracy and efficiency in knowledge-intensive tasks but also lays the foundation for future cross-domain applications and further improvements in retrieval methodologies.

III. Proposed Methodology

A. Data Preprocessing and Collection Module

Description: This module is responsible for gathering documents from diverse sources—including research articles (PDFs), internal knowledge bases, web pages, technical reports, and other structured/unstructured data—and preparing the data for subsequent processing. A key component is semantic chunking, which segments documents into contextually coherent fragments rather than fixed-length pieces, ensuring that entire ideas or concepts are preserved.

Significance: By breaking down documents into semantically meaningful units, the module ensures that the following retrieval stages have access to cohesive and context-rich content, thereby enhancing the overall accuracy and relevance of the system's responses.

B. Feature Extraction and Embedding Module

Description: This module extracts key features from each semantically chunked segment. It performs triplet extraction (subject-predicate-object mapping) to convert unstructured text into structured representations and leverages pretrained language models (e.g., from Hugging Face) to generate high-dimensional vector embeddings.
Significance: Transforming unstructured text into both structured triplets and vector embeddings facilitates efficient similarity-based searches and supports the development of robust, semantically aware retrieval systems.

C. Knowledge Graph Construction and Contextual Distillation Module

Description: Using the extracted triplets, this module constructs a dynamic knowledge graph where entities are represented as nodes and their relationships as edges. It continuously augments the graph as new documents are processed.
Additionally, the module applies contextual knowledge distillation techniques to extract query-specific subgraphs, focusing on the most relevant nodes and relationships for each query.
Significance: This approach enriches the system's ability to understand intricate relationships and to narrow down retrieval efforts to pertinent segments of the knowledge graph, thereby improving both efficiency and the accuracy of complex, multi-step reasoning tasks.

D. Hybrid Retrieval Mechanisms Module

Description: This module develops both a vector store and a graph store (using a graph database such as Neo4j). The vector store maintains document segment embeddings, while the graph store organizes the knowledge graph.
The integration of these two stores supports the retrieval of information through both vector similarity (capturing semantic meaning) and graph-based relational queries.
Significance: By fusing vector-based and graph-based retrieval methods, this module mitigates common challenges seen in standalone systems—such as retrieving vast amounts of irrelevant information or dealing with incomplete graphs—while ensuring that retrieval is both precise and contextually aware.

E. Hybrid Retriever Architecture Module

Description: At the core of the system lies a two-stage retrieval architecture that features handcrafted retrievers. The first stage employs a vector retriever to identify semantically similar document segments, while the second stage uses a graph retriever to extract relevant entities, concepts, and relationships. This dual approach leverages the strengths of both retrieval paradigms.
Significance: The combined retriever architecture facilitates an end-to-end contextual understanding by efficiently synthesizing information from the vector space and the knowledge graph, leading to improved factual accuracy and explainability of the system's responses.

F. User Query Processing and Contextual Response Module

Description: This module handles user queries by first processing them through natural language understanding (NLU) to extract intent, key entities, and relational context. Based on the query's complexity, it dynamically invokes the vector-based, graph-based, or integrated retrieval mechanisms. The retrieved segments and subgraph are then merged to form a comprehensive context.
Significance: Tailoring the retrieval process to the specific nature of each query enables the system to handle complex and multistep reasoning tasks effectively, ensuring that the final responses are both accurate and richly informed.

G. Response Generation Module

Description: Once the contextual information is compiled, a well-structured prompt is formulated for a Large Language Model (LLM). The LLM then generates a coherent, contextually relevant response that synthesizes both the factual and relational data retrieved from the system.
Significance: This module is vital for producing high-quality responses that are precise, well-supported by retrieved evidence, and adaptable to various types of queries, thereby overcoming limitations like hallucination and data staleness found in traditional systems.

H. Continuous Learning and Improvement Module

Description: This module implements periodic updates based on user feedback, query performance metrics, and evolving document datasets. It fine-tunes the underlying embedding models, adjusts retrieval strategies, and continuously enriches the knowledge graph with new documents and relationships.

Significance: By embracing a modular and iterative learning approach, the system remains adaptive and maintains high retrieval accuracy over time, ensuring its effectiveness in dynamic and knowledge-intensive environments.

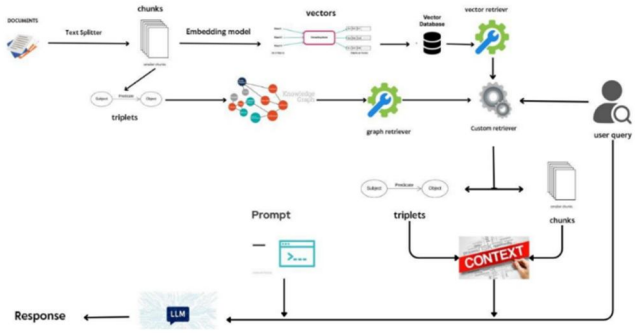

Figure 1. Pipeline Diagram for Hybrid Retrieval Process

This diagram illustrates the end-to-end data flow from document collection through feature extraction, hybrid retrieval, user query processing, and final response generation—demonstrating how each module interacts to produce a contextually enriched and accurate response.

IV. Result

This section details and interprets the results obtained from the project.

We employ three main metrics to evaluate the Hybrid Retrieval System's performance: faithfulness, answer relevancy, and answer similarity. These metrics are vital in determining whether the system's responses are correct, relevant, and contextually appropriate. Below, we outline each metric thoroughly, including the formula for calculation and sample results based on the system's performance.

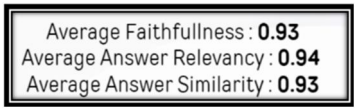

Figure 2. Evaluation metrics and their results

The findings highlight the performance metrics of an integrated system that utilizes both vector RAG (Retrieval-Augmented Generation) and graph RAG. This system attained impressive scores in three critical dimensions: faithfulness (0.93), which ensures that responses are based on retrieved information; answer relevancy (0.94), which reflects a strong contextual alignment; and answer similarity (0.93), demonstrating consistency in the generated responses. These metrics indicate that the system effectively balances factual accuracy with relevance.

V. Future Scope

The future scope of this system encompasses incorporating a broader array of sophisticated models, including deep neural networks, reinforcement learning, and graph neural networks (GNNs), to further enhance its ability to discern complex patterns and semantic relationships within documents. By integrating these advanced models, the system will be better positioned to refine both vector-based and graph-based retrieval methodologies, ultimately leading to improved reasoning and synthesis. This direction mirrors the need for expanding analytical capabilities seen in other state-of-the-art systems, where enhanced model integration drives better performance and accuracy in handling intricate and large-scale data.

In addition to advanced modelling, further development will focus on creating more context-sensitive retrieval mechanisms that adaptively adjust search results based on user history, query intent, and document structure. By leveraging user behaviour data alongside contextual cues, the system can be tuned to deliver more personalized and precise responses. This approach not only enhances retrieval relevance but also ensures that the system can

dynamically respond to the evolving nature of queries, much like how modern platforms integrate real-time data to improve overall efficiency, transparency, and user satisfaction.
These promising avenues aim to reinforce the system's robustness, scalability, and precision, making it a compelling solution for handling complex, multistep reasoning tasks in knowledge-intensive domains.

VI. Conclusion

In conclusion, the proposed Hybrid Retrieval System offers a transformative advancement in document processing by uniting vector embeddings with graph-based frameworks to deliver retrieval outcomes that are both highly precise and contextually enriched. By integrating semantic chunking, dynamic knowledge graph development, and tailored retrieval methods, the system not only addresses the limitations of traditional techniques but also provides deep insights through contextual knowledge distillation. This innovative approach extends beyond standard retrieval practices by extracting meaningful inter-entity relationships and synthesizing comprehensive query responses, which significantly enhances both operational efficiency and user experience. Moreover, the system lays a robust foundation for future enhancements—such as advanced machine learning integrations and adaptive retrieval strategies—that promise to further revolutionize document processing across diverse fields including legal, academic, and technical domains.

References

[1] Baldini, L., Baroni, E., Ghidini, C., & Rospocher, M. (2020). A hybrid retrieval system based on knowledge graphs and embeddings. Journal of Web Semantics, 62, 100589.
[2] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning-based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1).
[3] Rizzo, G., Falquet, G., & Hyvönen, E. (2021). Knowledge graph construction for improving document retrieval performance. Semantic Web Journal, 12(4), 559–577.
[4] Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.
[5] Singh, A., Gupta, P., & Jain, R. (2021). A hybrid RAG framework for efficient document retrieval in large-scale databases. Knowledge-Based Systems, 215, 106774.
[6] Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 855–864).
[7] Zhao, L., Zou, D., Zhang, X., & Wang, J. (2020). Enhancing document retrieval through knowledge graph-based contextual reasoning. Information Processing & Management, 57(6), 102337.
[8] Bao, Y., Zhang, Q., Cao, Y., & Fang, Z. (2022). Hybrid retrieval systems combining semantic chunking and graph traversal for question answering. IEEE Transactions on Knowledge and Data Engineering, 34(5), 1964–1979.
[9] Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724–2743.
[10] Lee, K., Heo, M., Lee, J., & Lim, H. (2023). Contextual knowledge distillation in graph-based document retrieval systems. Expert Systems with Applications, 211, 11854.
[11] Wang, Y., Xu, X., Li, Z., & Wang, S. (2020). A hybrid neural network model for knowledge graph-based document retrieval. Information Processing & Management, 57(3), 102256.
[12] Li, H., Zhang, Z., Jiang, S., & Wang, H. (2021). Enhancing question answering systems with dynamic knowledge graphs. IEEE Transactions on Knowledge and Data Engineering, 33(12), 3285–3298.
[13] Sun, C., Qin, T., & Liu, T. (2019). Multi-hop knowledge graph reasoning with rewards and penalties. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 1, pp. 464–471).
[14] Gao, L., Xie, Y., & Tang, J. (2020). Knowledge graph-based hybrid retrieval for legal information systems. Journal of Information Science, 46(4), 529–545.
[15] Huang, Z., Wang, X., & Yang, Q. (2019). Deep retrieval models for knowledge graph-enhanced question answering. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 1457–1465).
[16] Liu, J., Chen, T., & Shen, H. (2021). A hybrid approach for document retrieval based on semantic understanding and knowledge graphs. Expert Systems with Applications, 169, 114357.
[17] Feng, W., Shi, X., & Yang, P. (2018). Neural semantic parsing and hybrid retrieval for domain-specific question answering. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 146–160).
[18] Zhang, Y., Zhou, L., & Chen, F. (2017). Hybrid document retrieval based on knowledge graphs and semantic embeddings. Journal of Artificial Intelligence Research, 58, 385–409.
[19] Yu, H., Sun, Z., & Guo, J. (2020). Leveraging hybrid retrieval models for improving question answering over knowledge bases. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2301–2310).
[20] Cheng, W., Qiu, X., & Wang, X. (2019). Knowledge graph-augmented neural networks for document retrieval and reasoning. IEEE Access, 7, 65583–65592.