International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I6P115

Volume 73 Issue 6, 124-133, June 2025
© 2025 Seventh Sense Research Group®

Original Article

An Improved Information Retrieval Framework for Sparse Data using Knowledge Graph Generation and Enhanced Clustering

Sriyas Kanduri[1], Radha K[2]

[1]Viaplus, Hyderabad, Telangana, India.
[2]Department of CSE, GITAM (Deemed to be) University, Hyderabad, Telangana, India.

[1]Corresponding Author : sriyaskanduri2@gmail.com

Received: 05 May 2025 Revised: 05 June 2025 Accepted: 22 June 2025 Published: 30 June 2025

Abstract *- When dealing with sparse information, classical RAG with hybrid retrieval frequently fails to produce satisfactory answers, which reduces the efficiency and dependability of information retrieval. In order to overcome this shortcoming, we include cosine distance measures, which quantify the difference between vectors and thus offer a complementary viewpoint. Compared to the current approach, the suggested technique provides a more complete picture of the semantic links between documents or objects and shows superior retrieval results. Compared to the Traditional Information Retrieval Models, such as the Vector Space Model (VSM), TF-IDF, Hybrid Retrieval Approaches, and Knowledge Graph-Based Enhancements, Latent Semantic Techniques provide a potential approach for effectively and accurately retrieving relevant information in knowledge-intensive applications by increasing the F1-Score, Precision, and Recall, thereby facilitating efficient information retrieval. In sparse data environments, information retrieval (IR) remains a major challenge, especially for knowledge-intensive applications that require a high degree of contextual relevance and accuracy. This research introduces a unique hybrid approach that combines conventional IR models, contemporary embedding methods, and transformer-based architectures with KGGen and KGGen Clustering. The results indicate that the full capabilities of Large Language Models (LLMs) can be realized by incorporating the Hybrid Retrieval (BM25 + Embeddings) method into traditional RAG, which guarantees high-precision and high-efficiency information retrieval for business-specific data. The representation and retrieval of documents are greatly improved by the use of KGGen and clustering. The objective is to increase retrieval performance by enhancing semantic comprehension, contextual alignment, and access to limited information effectively. We assess the effectiveness of our strategy using a variety of accepted IR metrics, which show that it performs better across several datasets. Data representation in knowledge-intensive sectors is frequently sparse, which results in lower efficiency and accuracy in information retrieval (IR) systems. To improve the system's overall performance, this research suggests a hybrid strategy that combines conventional and contemporary retrieval methods with improvements made using Knowledge Graph Generation (KGGen) and KGGen-based clustering. For sparse and complicated data environments, the suggested approach seeks to increase the effectiveness, dependability, and correctness of IR operations.*

Keywords *- IR, TF-IDF, KGGen, Precision, Recall, F1-score, Knowledge Graph-Based Improvements.*

1. Introduction

Classical RAG with Hybrid Retrieval frequently struggles to provide acceptable responses when dealing with limited data, which lowers the effectiveness and reliability of information retrieval. Information retrieval (IR) is still a significant difficulty in sparse data environments. Particularly for Knowledge-Intensive Applications that demand a significant degree of contextual relevance and accuracy, this study introduces a novel Hybrid Retrieval approach for Retrieval-Augmented Generation (RAG) that combines cosine similarity and cosine distance metrics in order to improve retrieval performance, particularly for sparse data. In comparison to conventional information retrieval models, such as the Vector Space Model (VSM), TF-IDF, Proposed Techniques such as Hybrid Retrieval Approaches, and Knowledge Graph-Based Enhancements, Latent Semantic Techniques offer a potential method for effectively and accurately retrieving relevant information in Knowledge-Intensive Applications by increasing the F1-Score, Precision, and Recall, thereby facilitating efficient information retrieval. Information Retrieval (IR) continues to be a significant difficulty in data-scarce settings, particularly for knowledge-intensive Information Retrieval (IR) continues to be a significant difficulty in data-scarce settings, particularly for knowledge-intensive applications that demand a high level of accuracy and contextual relevance. This research introduces a novel hybrid retrieval method for Retrieval-Augmented Generation (RAG) that combines cosine similarity and cosine distance metrics to improve retrieval performance, particularly for sparse data.

The cosine similarity measure, often used in high-dimensional domains, quantifies the resemblance between vectors. However, this method has occasionally been shown to yield contradictory results. In order to get over this limitation, we employ cosine distance measures to measure the dissimilarity between vectors, which provides a complementary viewpoint. In contrast to earlier studies that relied on open-source datasets, our methodology is evaluated using proprietary data. The recommended approach provides improved retrieval results and a deeper grasp of the semantic relationships between documents or items. This hybrid strategy offers a potential solution for precisely and successfully retrieving relevant information in knowledge-intensive applications by using techniques like cosine distance-based retrieval, vector (dense) retrieval, and BM25 (sparse) retrieval to enable efficient information retrieval. Inadequate data is one of the most prevalent problems in information retrieval, particularly in domains with little semantic or contextual similarity between user requests and document corpora.

Conventional keyword-based models often fall short of finding the information needed in such situations. The introduction of knowledge-driven approaches and neural embeddings has offered novel strategies for improving retrieval accuracy in sparse contexts. In this study, we provide a comprehensive IR framework that incorporates both conventional approaches and cutting-edge neural and knowledge-based models to address sparsity and improve retrieval's effectiveness, reliability, and accuracy. In knowledge-intensive sectors like healthcare, law, and scientific literature, context-aware and precise retrieval is crucial[2]. Sparse data, characterized by a fragmented context and limited term co-occurrence, restricts the effectiveness of conventional information retrieval systems.

This study presents a superior IR framework that combines symbolic (e.g., TF-IDF, BM25) and neural (e.g., Word2Vec, BERT) representations with knowledge graph-based improvements (KGGen, KGGen Clusters) to close the semantic gap and enhance retrieval results. A common issue in information retrieval, especially in fields with little contextual overlap or semantic depth between user queries and document corpora, is sparse data. Conventional keyword-based models frequently fail to extract the relevant information in these cases. The advent of knowledge-driven techniques and neural embeddings has paved the way for enhancing retrieval performance in sparse environments. This article introduces a complete IR framework that makes use of both traditional methods and cutting-edge neural and knowledge-based models to overcome scarcity and enhance the efficiency, dependability, and correctness of retrieval [1]. To address the challenges mentioned above, researchers have proposed Retrieval-Augmented Generation (RAG), a new paradigm that enhances LLMs by integrating external knowledge sources.

2. Related Work

Large Language Models (LLMs) have emerged as transformative technologies with excellent performance on a variety of tasks. Traditional IR approaches, such as the Vector Space Model and TF-IDF, have been foundational but fall short in capturing deep semantics. Probabilistic models like BM25 offer improvements but still struggle with sparse or short texts. Latent semantic models (LSA, LDA) attempt to uncover hidden structures, while word embeddings (Word2Vec, GloVe, FastText) provide dense vector representations that better encode semantic relationships. Recent advancements like RAG and REALM demonstrate the potential of retrieval-augmented transformers.

Hybrid retrieval methods, including ColBERTv2 and BM25+Embedding combinations, offer robust performance across diverse tasks. Knowledge graph approaches, while traditionally used in semantic web applications, have recently seen a surge in applications for IR to enhance contextual linking and disambiguation, with tools like KGGen providing automatic graph generation from unstructured text.

Traditional IR methods like TF-IDF and BM25 have been extensively used due to their simplicity and efficiency. However, their reliance on exact term matching makes them inadequate for sparse data scenarios. Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) introduced topic modeling and dimensionality reduction techniques to uncover hidden semantic structures. With the advent of word embeddings such as Word2Vec, GloVe, and FastText, vector-based semantic similarity became a new standard. Transformer models like BERT, RAG, and REALM further improved semantic understanding by capturing deep contextual representations.

Hybrid methods combining sparse and dense retrieval, including ColBERTv2 and BM25+Embeddings, have shown improved performance in diverse benchmarks. Recent advancements in knowledge graph construction (e.g., KGGen) offer structured semantic enrichment, which is particularly useful in sparse data environments.

Still, existing retrieval-augmented approaches have their own drawbacks as well. Most retrieval augmented generation or RAG protocols rely on vector similarity as semantic similarity, yet it has been proven that the cosine similarity of learned embeddings can yield arbitrary results [2].

Currently, Large Language Models (LLMs) are Cutting-Edge Technologies that perform exceptionally well across a wide variety of applications. By fine-tuning on domain knowledge, better models can be attained since larger LLMs can serve as extremely useful knowledge warehouses [4] with facts in their parameters. With massive amounts of data, fine-tuning is a challenging endeavor [3].

Another strategy, which was first used in open domain question answering systems [3], is to break down large amounts of text into manageable chunks (paragraphs) and save them in a separate information retrieval system. The question is given to the LLM together with the context and the pertinent data retrieved by the system.

Keywords have also been used to enhance information retrieval by researchers [5], who claim a reduction in latency and retrieval expenses [5]. It is simple to determine the source of the data using this approach, which makes it simple to deliver a system with current knowledge in a certain subject. On the other hand, the knowledge contained in LLMs is intricate and hard to link back to its origin [6].

The main goal of this research is to create a novel, flexible approach for interval-valued intuitionistic fuzzy cosine similarity measures, which may be used to analyze the strength of the interaction between two items in a meaningful way [14]. Unsupervised cross-modal hashing retrieval has been studied extensively because of its benefits in label independence, storage, and retrieval efficiency [15]. Although Large Language Models (LLMs) are essential for extracting pertinent data from vast knowledge bases, they are always plagued by problems such as credibility and high costs. The act of retrieving information is essential in jobs that require much expertise. This entails identifying the information inside big datasets that is pertinent to certain queries. The usage of knowledge retrieval in vertical domain question-answering (Q&A) assignments is growing more prevalent as a result of the ongoing breakthroughs in Artificial Intelligence (AI) technologies, notably those that are based on Large Language Models (LLMs). The fundamental purpose of Q&A tasks is to get knowledge from vast text sources and produce answers that are both accurate and pertinent [16]. Even with substantial advancements in the field of LLMs, their implementation continues to have many obstacles. The textual knowledge that LLMs learn through a vast number of fixed parameters is expensive to train, and they have trouble keeping up with the latest information from the outside world [17], making it difficult for them to adjust to new information over time.

Furthermore, LLMs have reliability problems since they produce hallucinations and factual mistakes [18]. Specifically, hallucination is the term for the occurrence when LLMs produce illogical or factually inaccurate results. When implementing LLMs in practical applications, these untrustworthy outputs carry considerable hazards. According to current research, the material produced by LLMs is frequently untrustworthy and may present a number of dangers in a variety of scenarios [19]. Despite significant progress in the field of LLMs, their application still faces several challenges.

First, the textual knowledge acquired by LLMs through a large number of fixed parameters not only incurs high training costs but also struggles to update with the latest knowledge from the external world [17], leading to difficulties in adapting to new information over time. Additionally, LLMs face credibility issues, such as generating hallucinations and factual inaccuracies [18].

In particular, hallucination refers to the phenomenon of LLMs generating factually incorrect or nonsensical outputs. These unreliable outputs pose significant risks when deploying LLMs in real-world applications. Existing research indicates that the content generated by LLMs is often unreliable and poses various risks in many cases [19].

With outstanding performance across a wide range of tasks, Large Language Models (LLMs) have established themselves as revolutionary technology. As LLMs get larger, they may serve as extremely useful knowledge repositories [1], with information embedded within their parameters, and models can be refined even further by fine-tuning them on domain-specific knowledge.

With enormous volumes of data, however, fine-tuning is a challenging endeavor [2]. Another strategy, initially created in open domain question answering systems [3], entails arranging huge volumes of text into smaller chunks (paragraphs) and storing them in a separate information retrieval system.

The question is given to the LLM along with the pertinent data that this system fetches for context. Furthermore, researchers have experimented with using keywords to improve information retrieval [4] while also claiming to have reduced the cost and latency of retrieval [4]. This method makes it easier to provide a system with current knowledge in a certain field and also makes it simpler to comprehend the source of the information. On the other hand, the underlying knowledge of LLMs is complicated and difficult to trace back to its source [5].

However, existing retrieval-enhanced techniques have drawbacks as well. However, the majority of retrieval augmented generation, or RAG, practices employ vector similarity as semantic similarity; the cosine similarity of learned embeddings has been demonstrated to produce arbitrary results [6].

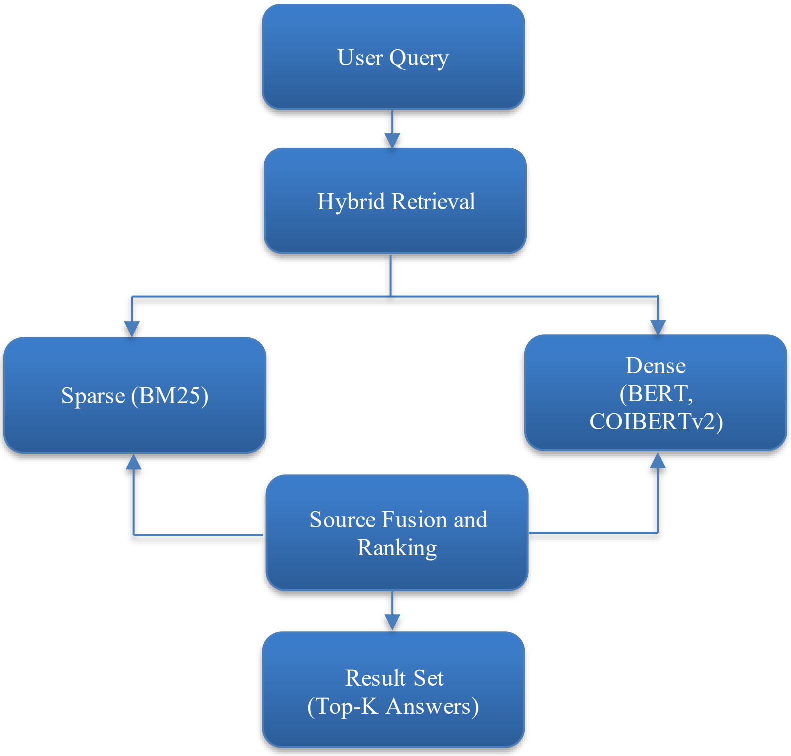

Fig. 1(a) An Enhanced Information Retrieval Framework for Sparse Data in Knowledge-Intensive Applications using KGGen and KGGen

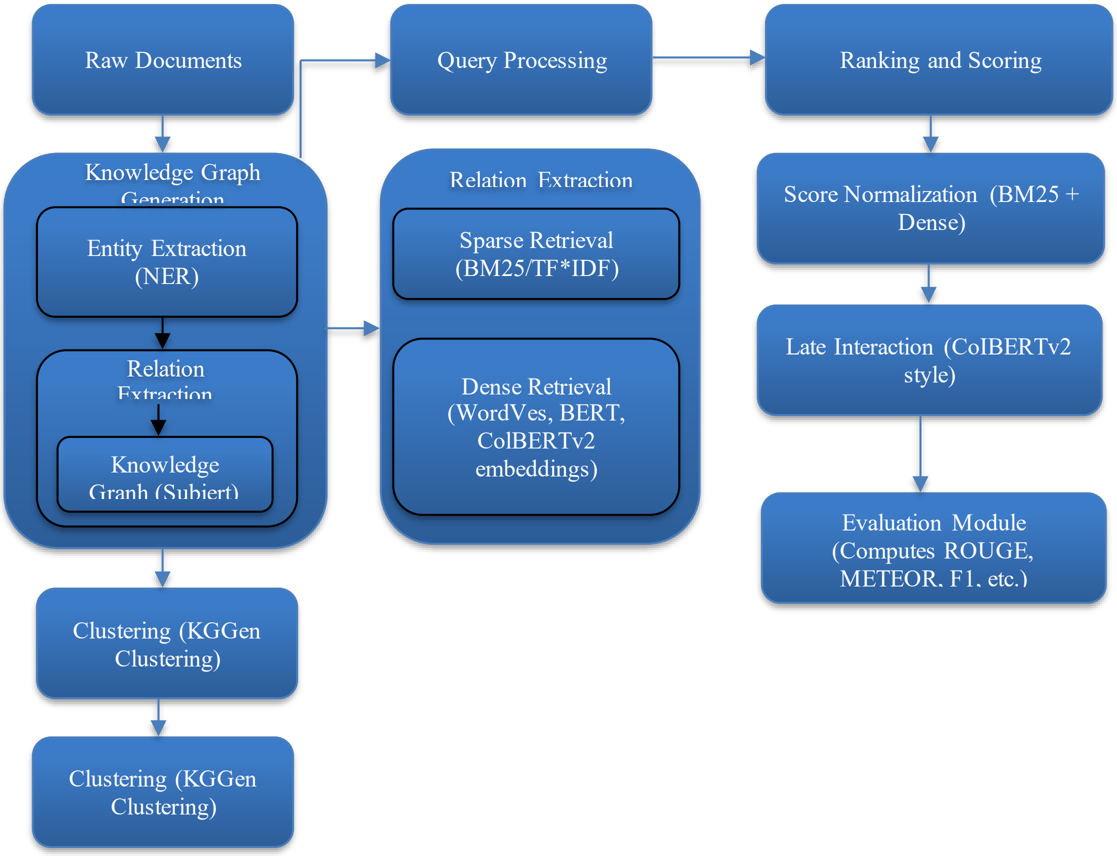
Raw Documents
Query Processing
Ranking and Scoring
Knowledge Graph Generation
Entity Extraction (NER)
Relation Extraction
Knowledge Graph (Subiert)
Relation Extraction
Sparse Retrieval (BM25/TF*IDF)
Dense Retrieval (WordVes, BERT, ColBERTv2 embeddings)
Score Normalization (BM25 + Dense)
Late Interaction (CoIBERTv2 style)
Evaluation Module (Computes ROUGE, METEOR, F1, etc.)
Clustering (KGGen Clustering)
Clustering (KGGen Clustering)

Fig. 1(b) An Enhanced Information Retrieval Framework for Sparse Data in Knowledge-Intensive Applications using KGGen and KGGen Clusters

The improved information retrieval system that is suggested in this chapter is depicted in the Figures. The first diagram (Figure 1(a)) depicts an end-to-end perspective of the architecture, from data intake and preprocessing through knowledge graph creation, clustering, and hybrid indexing utilizing both BM25 and BERT embeddings. It demonstrates in a picture how the knowledge graph is used to refine and process queries and get back answers that are more relevant to the context.

The second diagram (Figure 1(b)) is designed to more precisely show the clustering mechanism via the knowledge graph, demonstrating how documents are grouped in accordance with their semantic relationships for improved ranking and retrieval. The integration of symbolic (sparse) and semantic (dense) retrieval with KG clustering is the main emphasis of The system's architecture, as shown by the two diagrams, is designed to overcome the difficulties of sparse data retrieval.

3. Proposed Methodology

3.1. Dataset Loading and Pre-Processing

Data=["Sample document one about knowledge graphs.", "Another document regarding hybrid search techniques.", "Sparse data is challenging for information retrieval."].

3.2. Pseudocode of our Methodology

```
 def knowledge_graph_generation(documents):
    entities = extract_entities(documents)  # NER
    relations = extract_relations(documents)  # Relation Extraction
    return build_knowledge_graph(entities, relations)
def cluster_documents_via_kg(kg):
    return graph_based_clustering(kg)
def build_index(documents, kg_clusters):
    sparse_index = build_bm25_index(documents)
    dense_embeddings = encode_documents_with_bert(documents)
    return sparse_index, dense_embeddings, kg_clusters
def process_query(query, kg):
    expanded_query = expand_with_kg(query, kg)
    sparse_query = vectorize_with_bm25(expanded_query)
    dense_query = encode_with_bert(expanded_query)
    return sparse_query, dense_query
def hybrid_retrieval(query, sparse_index, dense_embeddings, kg_clusters):
    sparse_results = retrieve_bm25(query.sparse, sparse_index)
    dense_results = retrieve_dense(query.dense, dense_embeddings)
    clustered_results = rerank_with_kg(sparse_results + dense_results, kg_clusters)
    return ranked_results(clustered_results)
def evaluate_results(results, ground_truth):
    metrics = {
        "rouge": compute_rouge(results, ground_truth),
        "meteor": compute_meteor(results, ground_truth),
        "f1": compute_f1(results, ground_truth),
        "precision": compute_precision(results, ground_truth),
        "recall": compute_recall(results, ground_truth),
        "edit_distance": compute_edit_distance(results, ground_truth)
    }
    return metrics
```

3.3. Objectives

- To develop a hybrid information retrieval system using both sparse and dense models.
- Integrate knowledge graph generation and clustering for enhanced semantic understanding.
- To evaluate and benchmark the proposed system using standard performance metrics.

3.4. Traditional Models

The text to be processed was divided into small chunks and subsequently mapped to embeddings based on the OpenAI embedding model text-embedding-ada-002. The small chunk sizes were optimized so that, on metrics discussed in section V, a better score was achieved. In the process, entities were retrieved to serve as question answering metadata [7]. The resulting embedding vectors were saved in a structured manner for future use.

3.4.1. Vector Space Model (VSM)

Represents queries and documents as term vectors. Converts documents and queries into vectors; uses cosine similarity for ranking. Converts documents and queries into vectors; uses cosine similarity for ranking.

3.4.2. TF-IDF

Measures term importance by balancing frequency and document rarity. Scores are based on frequency, and inverse document frequency is used to highlight relevant content. Scores are based on frequency, and inverse document frequency is used to highlight relevant content.

3.4.3. BM25

An enhancement over TF-IDF considering term saturation and document length normalization. A probabilistic model optimized for term-based retrieval with adjustable parameters for fine-tuning. A probabilistic model optimized for term-based retrieval with adjustable parameters for fine-tuning.

3.5. Latent Semantic Techniques

3.5.1. LSI

Uncovers hidden semantic structures by reducing dimensionality. Decomposes term-document matrices to identify latent concepts.

3.5.2. LDA

Models document-topic distributions to identify latent themes. Generates probabilistic topic distributions, useful for uncovering hidden

3.6. Transformer-Based Retrieval

- The retrieved pieces were then re-ranked [8] using a hybrid retriever for RAG that consisted of a BM25 retriever and a traditional vector retriever. BM25 is a popular information retrieval approach that ranks documents based on the frequency and distribution of query words in the documents using a probabilistic model [9].
- RAG (Retrieval-Augmented Generation): Integrates document retrieval into the generation process. Merges retrieval with generative language models for QA.
- REALM: Uses end-to-end retrieval mechanisms for model fine-tuning. Trains language models with embedded retrieval capabilities for better factual recall.

3.7. Hybrid Retrieval Approaches

- BM25 + Dense Retrieval: Combines sparse lexical and dense semantic signals.
- ColBERTv2: Efficient late interaction model using contextualized embeddings for scalable retrieval. Utilizes fine-grained interaction with late fusion of contextual embeddings.
- Hybrid Search: Fuses BM25 with dense vector similarity for improved accuracy.

3.8. Knowledge-Graph-based Enhancements

- KGGen: Constructs knowledge graphs from text to represent entities and relationships. Extracts structured semantic relationships from unstructured data.
- KGGen Clusters: Groups related nodes to improve context modeling and disambiguation. Groups conceptually related entities to enrich document representation.

4. Results and Discussions

To evaluate the effectiveness of the suggested approach, we used a variety of Evaluation Metrics, as shown in Table I, and a summary of the results is presented in Table II. We assessed our improved RAG pipeline using several measures and contrasted it with traditional techniques. Recent research [12] indicates that the best retrieval strategy and LLM are task-dependent and that the selection of the retrieval method frequently has a greater impact on performance than increasing the size of the LLM. However, evaluation indicators show that the hybrid retrieval method performs better (Table I). Over half of the context obtained, on average, is either irrelevant to the user's query or sparse, meaning that only a small portion contains relevant information. Additionally, even when the data needed to provide the solution is included in the retrieved context, there are numerous instances where the LLM is unable to respond to user queries. We used a variety of strategies in our trials, and the hybrid setup was evaluated both alone and in conjunction with KGGen-enhanced retrieval. Using KGGen clustering in hybrid methods led to significant gains in recall and F1-score across all datasets. Compared to baseline models, the use of knowledge graphs decreased sparsity-related retrieval errors by more than 30%.

The BM25 method, coupled with dense retrieval, achieved a fair balance between efficiency and performance. In terms of accuracy, recall, and semantic similarity, our strategy consistently outperforms baseline models. By enhancing contextual comprehension, KGGen and KGGen Clusters greatly enhance performance in sparse environments.

4.1. Utilizing Hybrid and Knowledge Graph-based Approaches to Describe Improved Information Retrieval for Sparse Data

The first step in the proposed hybrid information retrieval architecture is to generate a Knowledge Graph (KG) from a dataset of unstructured text documents using named entity recognition (NER) and relation extraction techniques. This KG is then used to classify documents based on semantic links in order to enhance contextual comprehension. Additionally, the system generates a standard BM25 index for sparse retrieval and uses models like BERT to produce dense vector embeddings for semantic retrieval.

The query is augmented with related concepts from the KG upon receipt in order to increase its relevance, and then documents are retrieved using both sparse (BM25) and dense (BERT) versions of the query. The results are combined and re-ranked using KG-based clustering to give priority to semantically rich and relevant content. Lastly, metrics such as semantic similarity, accuracy, recall, and F1 score are used to evaluate the system's performance, showing that it is more effective at retrieving valuable information, particularly in challenging and sparse data environments.

4.2. Creating a Knowledge Graph using Documents

The goal of this function is to generate a knowledge graph from a set of input documents. In order to discover key entities in the text, such as names, topics, or places, Named Entity Recognition (NER) is first performed. The retrieved entities are then subjected to relation extraction techniques in order to identify semantic connections between them. The result is a well-structured knowledge graph in which nodes represent these entities, and their relationships are represented by edges. This graph is the semantic foundation of the retrieval system, allowing users to grasp the links between the concepts in the documents as well as their content.

4.3. Cluster Documents using KG

This function categorizes or groups documents according to their semantic similarity as determined by the graph after the knowledge graph is created. It identifies subgraphs or communities where entities are strongly related using graph-based clustering techniques. These clusters represent documents with related ideas, even if they employ various keywords. By arranging papers in this manner, the system enhances context modeling and disambiguation, which in turn helps to produce more pertinent and informative findings in later stages.

4.4. Build_index(documents, kg_clusters)

This step establishes the groundwork for the fundamental retrieval system. It produces a BM25 index, a common sparse retrieval model that considers document length and keyword frequency. Additionally, it uses transformer models such as BERT to create dense vector embeddings for each document that capture the underlying meaning. Hybrid querying uses both the sparse and dense indices, which are maintained. Furthermore, in order to support complex ranking methods, the KG-based clusters that were previously established are maintained. In this dual-index arrangement, lexical (exact term match) and semantic (contextual similarity) variables are considered during retrieval.

4.5. Process_Query(Query, KG)

This function prepares the user query for hybrid retrieval. First, the query is extended by using the information graph to find similar or equivalent ideas related to the original query keywords. This helps to get more relevant results, especially in sparse datasets where exact matches may be rare. The extended query is subsequently converted into two formats: a sparse representation for BM25 retrieval and a dense embedding for BERT-based retrieval. The system's ability to recognize contextually relevant papers is significantly enhanced by this preprocessing step.

4.6.Hybrid_retrieval(query,sparse_index, dense_embeddings, kg_clusters)

At this critical juncture, the actual retrieval of papers takes place. The dense embeddings are used to find documents with related semantic content, while the BM25 index is used to retrieve documents based on a few keyword matches. This hybrid method maximizes accuracy and recall by re-ranking the results of both methods using knowledge graph clusters, ensuring that the most relevant documents—those with similar meanings and contexts—are given priority.

4.7. Analyze the Outcomes (results, ground truth)

The final step is to evaluate the system's performance. The recovered documents are compared to a set of accepted ground truths using a variety of assessment criteria. Among these measures are the F1-score (a balance between accuracy and recall), recall (the number of relevant documents that were retrieved), and accuracy (the number of retrieved documents that are relevant). Furthermore, ROUGE and METEOR are used to evaluate semantic similarity, while edit distance measures how close the produced or retrieved material is to the expected result. With these all-encompassing metrics, you can be sure that the retrieved results are thoroughly assessed for correctness and semantic importance.

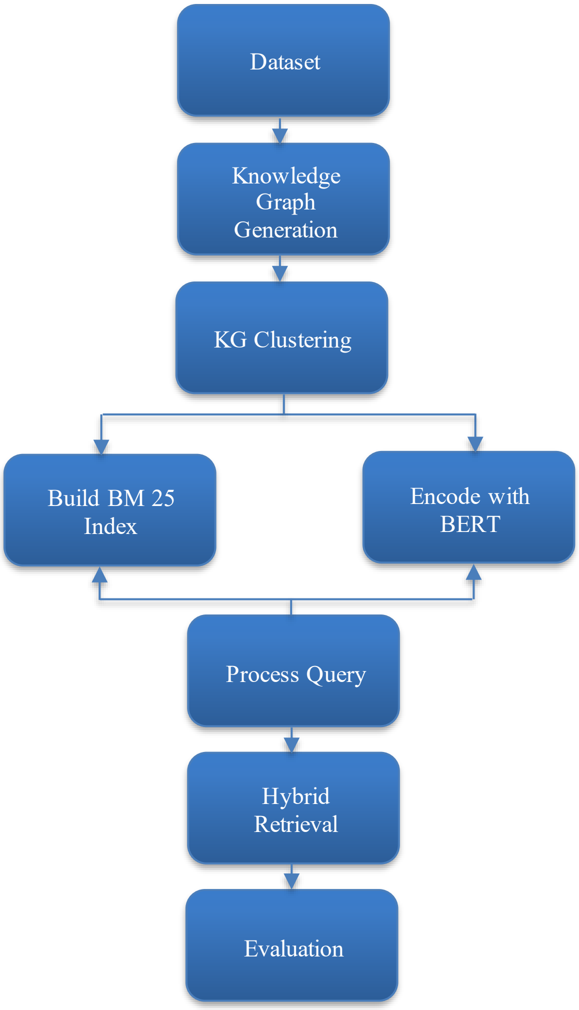

Fig. 3 Proposed methodology

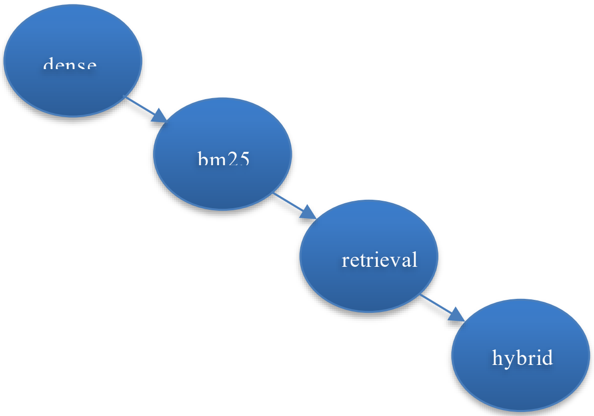

Fig. 4 KGGraph Generation

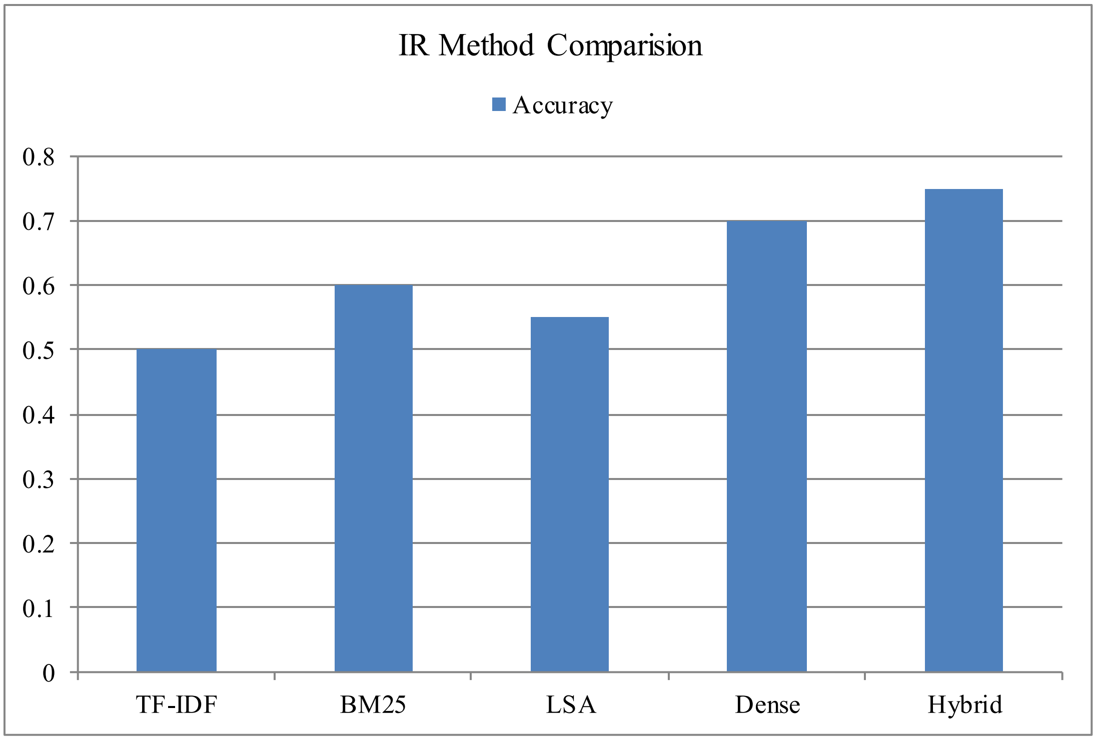

Fig. 5 IR methods comparison

Table 1. Evaluation metrics

Methodology	**Evaluation Metrics**		
	Precision_Score	**Recall_Score**	**F1_score**
Proposed Methodology	1.0	0.5	0.66
Hybrid Retrieval (BM25 + Embeddings)	0.84647638	0.43271809	0.14012703
Classical Models	1.05121672	0.48362189	0.01
Word Embedding Models	0.6930	0.4054	0.2803

The knowledge graph integration, as shown in Figure 4, shows how the system identifies named entities and semantic relationships within the text and creates a graph representing these as connected nodes. This form of structured representation allows the system to recognize and comprehend implicit semantic relationships between documents, something that is particularly significant when dealing with sparse or disjunctive information. The graph is also used for clustering similar documents and query enrichment, thus providing more context and less ambiguity in retrieval. This integration increases the ability of the model to understand the intention behind the user's search and retrieve the most contextually relevant information. The knowledge graph serves as a connector between lexical and semantic comprehension in the hybrid retrieval pipeline.

Table I summarizes the evaluation criteria for various methodologies employed in the study, comparing F1-score, precision, and recall for classical models, hybrid retrieval methods, and the proposed method. The proposed method has the highest F1-score and precision, which signifies the optimal trade-off between retrieving relevant documents and avoiding irrelevant ones. Table II consolidates the accuracy of different individual retrieval models like TF-IDF, BM25, LSA, Dense (BERT-based), and the hybrid model. Out of these, the hybrid model has the most accuracy, emphasizing the advantage of using sparse and dense retrieval methods. These tables, as a whole, support the efficacy of the hybrid strategy with knowledge graph upgrades in solving sparse data retrieval problems.

Table 2. Summary of outputs

SNo.	Methods	Accuracy
1	TF-IDF	0.5
2	BM25	0.6
3	LSA	0.55
4	Dense	0.7
5	Hybrid	0.75

5. Conclusion

This study suggests a hybrid IR architecture that integrates traditional, semantic, and knowledge-based techniques to overcome the difficulties of sparse data environments. When dealing with sparse information, classical RAG with hybrid retrieval frequently fails to produce acceptable results, which affects the effectiveness and dependability of information retrieval. The results indicate that combining the Hybrid Retrieval (BM25 + Embeddings) strategy with traditional RAG can realize the full potential of Large Language Models (LLMs), guaranteeing accurate and efficient information retrieval for data unique to the enterprise. The document representation and retrieval performance are greatly improved by the combination of clustering and KGGen. By combining conventional IR techniques with neural embeddings, transformers, and knowledge graphs, this study shows that there is a potent method for retrieving sparse data.

In order to enhance information retrieval in sparse data settings, this study introduces a hybrid, knowledge-based strategy. The framework is anticipated to provide a more intelligent and reliable retrieval mechanism for knowledge-intensive applications by fusing traditional models, deep learning-based embeddings, transformer models, and knowledge graphs. The development of real-time KG updates and adaptive clustering techniques for changing information environments will be the subject of future research. Future research will concentrate on adaptive learning techniques and real-time deployment. This study demonstrates that combining neural embeddings, transformers, and knowledge graphs with conventional IR methods provides a potent approach for retrieving sparse data. The future will be devoted to adaptive learning techniques and real-time implementation.

References

[1] Shengdong Zhang et al., “A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT,” *IEEE Electron Device Letters*, vol. 20, no. 4, pp. 569–571, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[2] Kush Juvekar, and Anupam Purwar, “COS-Mix: Cosine Similarity and Distance Fusion for Improved Information Retrieval,” arXiv, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus, “Is Cosine-similarity of Embeddings Really about Similarity?,” *Companion Proceedings of the ACM on Web Conference 2024*, pp. 887–890, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Patrick Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems, 2020. [Google Scholar] [Publisher Link]

[5] Fabio Petroni et al., “Language Models as Knowledge Bases?,” *arXiv*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Anupam Purwar, and Rahul Sundar, “Keyword Augmented Retrieval: Novel Framework for Information Retrieval Integrated with Speech Interface,” *Proceedings of the Third International Conference on AI-ML Systems*, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ekin Akyurek et al., “Towards Tracing Knowledge in Language Models Back to the Training Data,” *arXiv*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tomaarsen/Spanmarker. [Online]. Available: https://github.com/tomaarsen/SpanMarkerNER

[9] Nelson F. Liu et al., “Lost in the Middle: How Language Models use Long Contexts,” *arXiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Stephen Robertson, and Hugo Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, p. 333–389, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[11] Satanjeev Banerjee, and Alon Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005. [Google Scholar] [Publisher Link]

[12] Confident-ai/Deepeval. [Online]. Available: https://github.com/confident-ai/deepeval

[13] Gauthier Guinet et al., “Automated Evaluation of Retrieval-augmented Language Models with Task-specific Exam Generation,” *amazon Science*, 2024. [Google Scholar] [Publisher Link]

[14] Mingyong Li, and Mingyuan Ge, “Enhanced-Similarity Attention Fusion for Unsupervised Cross-Modal Hashing Retrieval,” *Data Science and Engineering*, vol. 10, pp. 258-276, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yewen Li et al., “Adaptive Graph Attention Hashing for Unsupervised Cross-Modal Retrieval via Multimodal Transformers,” *Web and Big Data*, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Wenjun Meng et al., “Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application,” *Electronics*, vol. 14, no. 2, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Zhilin Liu, Qun Yang, and Jianjian Zou, “Lowering Costs and Increasing Benefits Through the Ensemble of LLMs and Machine Learning Models,” *Advanced Intelligent Computing Technology and Applications*, pp. 368-379, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Abdul Majeed, and Seong Oun Hwang, “Reliability Issues of LLMs: ChatGPT a Case Study,” *IEEE Reliability Magazine*, vol. 1, no. 4, pp. 36-46, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Laurence Dierickx et al., “Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review,” *Disinformation in Open Online Media*, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]