RAG Fusion

RAG Fusion is an advanced retrieval technique that combines multiple retrieval methods and query variations to improve the overall quality and diversity of retrieved documents.

Core Concept

The fundamental idea behind RAG Fusion is to leverage multiple retrieval approaches and query formulations, then combine the results using a ranking fusion technique. This approach aims to overcome the limitations of individual retrieval methods and capture different aspects of relevance.

Process

1. **Query Expansion**: Generate multiple variations of the original query. This can involve:
    - Using the original query as-is
    - Rephrasing the query
    - Expanding the query with additional context or related terms
2. **Multi-method Retrieval**: For each query variation, perform retrieval using multiple methods, such as:
    - Vector similarity search (e.g., using embeddings)
    - Keyword-based search (e.g., BM25)
    - Other specialized retrieval methods
3. **Result Aggregation**: Collect all the results from the various retrievals into a single pool.
4. **Deduplication**: Remove duplicate documents from the aggregated results.
5. **Rank Fusion**: Apply a rank fusion algorithm to combine and rerank the results. A common choice is Reciprocal Rank Fusion (RRF):
    - For each document, calculate its RRF score across all result lists
    - The RRF score for a document is the sum of the reciprocal of its rank in each list where it appears
    - Documents that appear in multiple result lists and at higher ranks receive higher overall scores
6. **Final Ranking**: Sort the documents based on their fused ranks and return the top-k results.

Advantages

1. **Diversity**: By using multiple query formulations and retrieval methods, RAG Fusion can capture a more diverse set of relevant documents.
2. **Robustness**: The approach is less sensitive to the weaknesses of any single retrieval method or query formulation.
3. **Improved Recall**: The use of multiple approaches increases the chances of retrieving relevant documents that might be missed by a single method.
4. **Balance of Relevance Signals**: The rank fusion step allows for a balanced consideration of different relevance signals (e.g., semantic similarity, keyword matching).

Challenges and Considerations

1. **Computational Cost**: Running multiple retrievals and the fusion process can be more computationally expensive than simpler approaches.
2. **Complexity**: Implementing and tuning a RAG Fusion system can be more complex than single-method retrieval systems.
3. **Parameter Tuning**: The effectiveness of the approach can depend on careful tuning of various parameters (e.g., number of query variations, weights in the fusion process).
4. **Result Coherence**: With diverse retrieval methods, ensuring that the final set of results is coherent and not too disparate can be challenging.

Implementation Considerations

- **Query Variation Generation**: This can be done using rules, templates, or even language models for more sophisticated query expansion.
- **Choice of Retrieval Methods**: The selection of retrieval methods should be based on the characteristics of the document collection and typical queries.
- **Fusion Algorithm**: While RRF is common, other fusion methods (e.g., CombSUM, Borda count) can also be considered.
- **Efficient Implementation**: For large-scale systems, efficient implementation of the fusion and reranking steps is crucial for performance.

Use Cases

RAG Fusion can be particularly effective in scenarios such as: - Web search engines - Enterprise search systems - Academic literature retrieval - Legal document search - E-commerce product search

By combining multiple retrieval strategies and query formulations, RAG Fusion offers a powerful approach to improving retrieval performance, especially in

complex information retrieval scenarios where different aspects of relevance need to be considered.