A ML-based Query Expansion in Vertical Aggregated Web Search Using Pseudo-Relevance Feedback

Kashia Riaz, Umer Rashid*
Department of Computer Science (Quaid-i-Azam University, Islamabad, Pakistan).
*Correspondence: umerrashid@qau.edu.pk

Citation | Riaz. K, Rashid. U, "A ML-based Query Expansion in Vertical Aggregated Web Search Using Pseudo-Relevance Feedback", IJIST, Special Issue pp 307-323, May 2026.
Received | March 10, 2026 **Revised |** April 27, 2026 **Accepted |** May 02, 2026 **Published |** May 08, 2026.

The massive increase in multimedia content over the internet has posed a significant challenge to vertical search engines in instances where users submit short, ambiguous, or vocabulary-mismatched queries. Current query expansion techniques, including Pseudo-Relevance Feedback (PRF), show limited effectiveness in addressing these challenges, particularly in multimedia and cross-vertical search settings. The developed hybrid query expansion method described in this paper combines PRF, Machine Learning (ML), and Graph Theory to enhance the relevance of search results and enhance the semantic relevance of search results. The methodology consists of applying PRF to extend the first query with the help of relevant terms and then using a machine learning model to narrow the terms according to past search behavior and semantic trends. Also, graph-based techniques are utilized to determine semantic relationships among query terms, thereby improving contextual relevance of the query. The framework identifies semantically related terms and refines the query to improve retrieval relevance and adjusts the search, so the results are more accurate and relevant. This experiment used the TREC Web Track collection, which consists of 50 queries, 25,000 documents, and some artificially generated multimedia sources. The empirical findings show that there are great improvements in essential performance measures. Specifically, the given approach yields a 36.2% increase in MAP, 28.5% in nDCG@10, and 32.6% in P@10, as compared to BM25. In addition, a 30-user study has shown a 40% reduction in the query formulation time and a 25-30 percent increase in user effectiveness. The results indicate better accuracy of retrieval, lower query effort, and increased user satisfaction, especially for novice users involved in exploratory search activities. Additionally, issues of scalability, data privacy, and the computation cost are presented in the real-world application.

Keywords: Graph Theory, Query Expansion, Pseudo-Relevance Feedback, Machine Learning, Multimedia Web Search, Vertical Web Search

Introduction:

Multimedia content that is exponentially increasing on the web poses significant challenges to search engines, especially where user queries are poorly formulated. Current web search systems tend to have issues finding useful information when the search terms are brief, unclear, or when they have a vocabulary mismatch [1][2]. These problems are exceptionally high in vertical aggregated search engines, which retrieve heterogeneous kinds of content, including text, images, videos, and news [3][4]. Even though vertical search engines are the preferred channel when it comes to retrieving a particular type of content [4], they tend to produce irrelevant results because of the complexity of methods of handling the various types of content and the purpose of the query by the user [5][6].

Aggregated search (AS) is among the better techniques to manage the growing volume and complexity of search information by aggregating the search results of various content types of text, images, videos, and news [4]. The main categories of aggregated search include vertical search and relational search, which are used differently in search engines on the web [7]. AS engine retrieves heterogeneous content, and sorting (by relevance score), grouping (clustering), and merging (summary) may therefore be required before presenting search results to the end user. Nevertheless, it is not easy to accumulate such wide-ranging information, especially when the users do not know the terms or the area of the content. Vertical search engines play an essential role in filtering information on domains, yet there are significant problems in returning effective responses as the query presented by the user is rather complex in nature. Moreover, search results are not presented most efficiently, especially when using several tabs. When integrating results into a unified interface, the search experience is not as intuitive as possible [8].

Multimedia information is diverse, including images, audio, videos, and so on, which are usually accessed by search engines. For each type of media in multimedia information, separate search verticals exist, resulting in a scattered form of this information. According to the query of the user, the multimedia vertical snippets are assembled through relational aggregated search, and then they are re – ranked based on their semantic similarity. So, multimedia aggregated search has a proposed framework in which there is a fully blended relational aggregated search to improve diversity and relevance in aggregated search [9].

The incompatibility of old query expansion techniques to capture the contextual relationships of words entirely in a query only worsens this problem. Query modification is applied to solve problems of ambiguity, vagueness in the query, and vocabulary problems in most information retrieval systems. It improves the accuracy and retrieval rate of previous search results. It removes ambiguity in the original query. Several techniques are built based on query modification, user feedback, and the result list of previous queries. Query suggestion, query refinement, and query expansion are considered as the most well – known techniques to modify the entered query by the user [10].

A well-known method to enhance search results is query expansion, which extends the initial query by adding new words. The current techniques of query expansion may generally be divided into manual, automated, and interactive query expansion methods. Nevertheless, many traditional methods, such as Pseudo-Relevance Feedback (PRF), are based on simple term co-occurrence and relevance feedback schemes, and thus, cannot be used to support more complex search tasks, e.g., in the multimedia and cross-vertical search environment [11][12]. Such approaches do not always incorporate semantic interrelations among words and cannot enhance the relevance of queries in the presence of heterogeneous types of content. The most important challenge is to provide relevant information to users so that they can find the relevant information by observing the interests and requirements of users. Web search is also going to be social web search nowadays. The general process of Information Retrieval

does not consider the social dimension, and results are returned in the form of web pages using any search engine [13].

Query expansion (QE) is a well-known method employed to enhance the search results by incorporating new words into the initial query, to reduce the amount of time required in processing and retrieving queries. Historical query expansion techniques, including pseudo-relevance feedback (PRF), tend to use top-ranking documents as the basis of determining terms of relevance to expand. Although PRF has proven successful in query relevance-enhancement, it is susceptible to several drawbacks, especially where irrelevant documents are present in the feedback, and this disrupts query drift [14]. The second limitation of the conventional QE techniques is that they cannot consider contextual links between query words. These limitations are more critical in multimedia queries and complex queries, as the relationship between words is more complex.

Semantic query expansion (QE) has received much attention in enhancing information retrieval by solving vocabulary mismatch by expanding terms by meaning. Recent surveys note the development of semantic QE methods, such as linguistic, ontology-based, and hybrid methods, and the incorporation of current AI methods, such as word embeddings and transformer-based models. These methods are more effective at retrieving user intent as they more precisely capture user intent, but there are still open research problems of query drift, computational complexity, limited resource availability, and multilingual and multimedia data [15].

Web knowledge-based QE is a proposed model to deal with word mismatch. In response to the seed query, pseudo-relevant web knowledge is being employed so that the initial set of terms for expansion can be extracted. Each of the individual terms from the content of the web is calculated by the tf-idf approach. The relation among terms is obtained by computing the score of expansion terms through an approach of KNN-based cosine similarity. The connection of selected expansion terms with the original query is weighted by correlation score [16]. Query expansion improves retrieval in federated search, particularly for single databases, but shows inconsistent performance across multiple sources due to data heterogeneity; selecting robust expansion terms (e.g., from URLs) remains a key challenge [17].

To overcome the limitation of blindly selecting the top-n documents, a knowledge-based query expansion framework is proposed that integrates corpus-based and knowledge-driven techniques with relevance feedback. The relevance feedback process is automated, and high-quality expansion terms are selected using semantic relatedness measures. The proposed framework enhances the information retrieval process by improving both the diversity and effectiveness of search results. Experimental evaluation is performed on standard IR datasets, namely ADI, CISI, and CACM, using unigram and Okapi BM25 as baseline models, with precision and mean average precision (MAP) as performance metrics. This framework depends on initial retrieval quality and semantic relatedness, which may introduce noise and reduce effectiveness in large-scale, real-world scenarios [18]. A novel relevance feedback (RF) framework is proposed that jointly exploits visual and textual features from retrieved results, augmented with spelling correction and cross-lingual query translation. RF methods are evaluated based on the modality of feedback signals. Explicit, short-term RF strategies leveraging visual descriptors and textual representations are investigated for medical image retrieval. This framework is constrained by challenges in balancing visual–textual features and potential errors from spelling correction and cross-lingual translation, affecting retrieval precision and adaptability [19].

To overcome these limitations of conventional techniques of QE, various techniques of semantic-based QE have been proposed. These techniques use external semantic tools like WordNet or Wikipedia to identify semantically related words that are not included in the original query [15]. Another technique, semantic filtering like BERTScore and WordNet

filtering, minimizes uncertainty and maximizes the accuracy of QE [20]. Although these techniques are beneficial for capturing semantic relationships, they are still limited, as they do not address dynamic user queries or content. In addition, Knowledge Graph QE was also made Knowledge Graph-based QE has attracted considerable attention due to its ability to establish more semantic relationships. Knowledge graphs (KG) are more structured representations of queries and their enhanced context of QE. However, these techniques may face issues with matching words and semantic uncertainty with dynamic web search targets, especially in multimedia search.

Graph methods, such as term-term co-occurrence graphs and semantic networks, have also been used for query expansion [7]. These are also represented using graph theory. These methods have the limitation of not being able to adapt to changes promptly, and the limitation of using a fixed set of data. The latest advancement in the graph-based query expansion method is the Wikipedia WordNet-based query expansion (WWQE), which produces better results since the semantic association of the terms is found [12]. Nevertheless, there is still a gap in having a dynamic and user-specific adaptation on query expansion. Besides, the multimodal associations between various content types, reflected in images, text, and videos that are central to vertical aggregated search, have remain insufficiently addressed by many graph-based approaches.

A framework as a novel approach for expanding queries is proposed so that the ranking of web search can be refined in a better way. There is a need for an approach that can help in choosing the most relevant terms that can be further added to the query. The proposed approach is an extended version of the WWQE technique. In this approach, Google Wikipedia Search is done after the initial query. Then, the top Wiki documents are selected, out of which the two best articles are selected. The selected articles are taken for content extraction of articles by using the scores of articles in terms of in-link and out-link approaches. Then the links that are related to each other in terms of semantic measures are combined. Then the CETs are weighted, ranked, and selected using in-links and cosine similarity to select the final expansion terms. Wikipedia is the main source on which the development of the QE prototype in real time is dependent. The main concern of this approach is to weight the documents in an effective way, along with an enhancement in candidate expansion terms. Evaluation is done by using the TREC dataset [21].

Query2doc is a simple yet effective query expansion method that leverages large language models (LLMs) to generate pseudo-documents using few-shot prompting. These generated documents are appended to the original query and integrated with both sparse and dense retrieval models to improve retrieval performance. The approach is motivated by distilling semantic knowledge from LLMs through prompting, and empirical results show consistent improvements across multiple datasets and retrieval systems. However, the method introduces efficiency challenges, as LLM inference increases latency and expanded queries slow down index search, making real-world deployment computationally demanding [22].

In case of retrieving the results from vertical web search, it is challenging for the users to formulate the query by themselves according to their data requirements. Irrelevant results are retrieved due to short length vague queries, a deficiency of field data, and a lack of skills to modify the query. The retrieval of documents in a web search engine in the form of vertical AS is causing problems due to fewer keywords in a query, uncertain, vague, and poorly defined expressions of the query. Query enhancement approaches are presented for ill – defined queries. Relevance feedback is also considered important to retrieve effective search results. Query recommendation approaches are also presented to recommend well-structured queries. The proposed approach extracts field data using pseudo-relevance feedback, alters it into an integrated text-to-text summary, and supports users in generating precise and well-structured query recommendations. This approach is evaluated by empirical Analysis in comparison to

the Google search engine as a baseline. The accuracy measure for this approach is 89%. In user behavior analysis, like clicks, keystrokes, time, log, and so on, users rated this approach as 80/100 [23].

Effective ways are needed in processing and extracting data from vast, noisy data available on the internet. A framework to represent the knowledge is known as a Knowledge Graph (KG). The state-of-the-art techniques in KG are knowledge base and RDF (Resource Description Framework) in KG, KGs Construction tasks, reasoning in KG, general purpose KGs (DBpedia, Wikipedia, etc.), expert KGs (FIBO (Financial Industry Business Ontology), etc.), and industrial KGs (eBay, IBM, etc.). Some challenges of KG are entity disambiguation, semantic embedding, dynamic knowledge, knowledge extraction, privacy, security, and so on. Information deficiency, inadequate and inappropriate information, variations, integration of knowledge, etc., are major problems of KG [24]. Knowledge Graph Embedding (KGE) is an approach used to preserve the structural information of KG and represent it in some vector space. KGE generates a vector representation of entities and relationships and to preserve relational semantics and improve scoring functions so that embedding can be learnt. Knowledge graph embedding (KGE) is employed in entity-based learning, textual learning, translation model, semantic model, graph attention network, multi modal KG (Multiple info like image and text, etc.). The applications of KGE are KGE with textual data and KGE based on deep learning [25].

Knowledge graphs have emerged as a powerful means for organizing and representing large-scale, complex information, playing a crucial role in the development of intelligent systems and data-driven applications. They enable effective knowledge representation and support a wide range of AI applications across various domains. Recent research gives a thorough overview of the possibilities of knowledge graphs, such as how to implement them into AI systems and in various application fields, but also identifies several central challenges, such as knowledge acquisition, embedding, completion, fusion, and reasoning. Although these technical challenges are becoming increasingly significant, they still restrict their full potential and need new studies to be able to be scaled and implemented efficiently [26].

ConvGQR is a conversational query reformulation method that incorporates query rewriting and query expansion with generative pre-trained language models (PLMs) to capture user intent in multi-turn conversations. It uses a rewrite model to rewrite queries dependent on context, a model to generate possible answers, which are the expansion signals, and a knowledge infusion mechanism to match reformulation with retrieval tasks. Experimental findings using numerous conversational search data illustrate a large enhancement of performance. Nonetheless, the framework presents an extra computational burden and involves the presence of many PLMs, relying extensively on generated responses, which can be inefficient and weak in practice [27]. The search Retrieval-Augmented Generation (RAG) is not a query expansion method but is a complementary method that can improve retrieval through the combination of generative models with search processes. Query reformulation and query expansion techniques can be used to further enhance retrieval performance in RAG-based systems, so such methods can be useful in maximizing retrieved context quality as input to downstream generation processes [28].

Recent surveys emphasize the shift of query expansion (QE) towards a new set of methods that make use of pre-trained and large language models (PLMs/LLMs) rather than the traditional corpus-based and lexicon-based approaches to query expansion. These techniques are classified according to integration in retrieval pipelines, interaction mechanism, and incorporation of knowledge graphs, and have evidence of greater performance in dealing with ambiguity and retrieval performance. Nevertheless, problems with reliability, efficiency, scalability, and adaptability to dynamic environments are also major concerns of real-world deployment [29].

One of the most diverse problems in web search is the fact that users cannot always adequately express their needs in terms of the information they need, and they end up with poorly formulated queries. This causes poor search outcomes, particularly in a query of multimedia information in a vertically aggregated search environment. The classical method that depends on the query made by the user directly has not been adequate, especially with users who might not entirely understand how to construct an effective query. We recommend a hybrid query expansion technique to overcome this by using a combination of Pseudo-Relevance Feedback (PRF), Machine Learning (ML), and Graph Theory to solve query ambiguities, vocabulary differences, or poor query formulations.

The proposed method is a combination of the three techniques, and the process of query formulation, on the one hand, can help to improve the retrieval accuracy in the context of vertical aggregated search. Our architecture is built in a series of steps, with the query entered by the user being filtered out in a series of transformations. The proposed system architecture involves the following phases:

User Search Input: The user types in a query (e.g., short keywords, natural language query), and it is handled by the system.

Vertical Aggregated Search: The Query is used to find different content of various types (text, images, videos, news) in multiple verticals.

Pseudo-Relevance Feedback (PRF): The documents most frequently accessed are relevant, and words on these documents are used to add to the initial query.

Machine Learning Model: The query is narrowed to build upon the previous user behavior and semantic patterns, which are captured in the history of search information.

Graph Theory Model: This model is used to develop semantic networks among the words of the query and increase the context relevance.

Query Generator: This system is a combination of the refined query output of the ML and graph-based modules, which produces a final enhanced query.

Better Results Retrieval: The improved query is then applied to extract the most significant results, which are more in line with the information requirements of the user.

The following are the aims of this research:

To construct a hybrid query expansion scheme (PRF-ML-Graph) that combines pseudo-relevance feedback, machine learning, and the use of graph theory in enhancing query formulation in vertical aggregated web search.

To overcome the problems of formulating poor queries, vocabulary differences, and semantic ambiguity during multimedia search.

To improve query representation by including patterns of user behavior (ML) and semantic relationships among query words with the help of graph-based modeling.

To enhance the retrieval performance of heterogeneous content (text, images, videos, news) in vertical aggregated search.

To measure the performance of the proposed framework on the standard information retrieval measures, such as Mean Average Precision (MAP), nDCG@10, and Precision at K (P@10).

To compare the suggested hybrid method to baseline methods, such as BM25, BM25 with PRF, and PRF with graph-based expansion, in terms of retrieval accuracy and relevance.

To evaluate whether the suggested approach is effective in helping users with limited domain knowledge to create more useful queries and find the necessary results.

To quantitatively compare the gain of the proposed framework on TREC Web Track and multimedia datasets (e.g., in MAP, nDCG@10, and P@10) over baseline approaches, as well as query formulation times (up to 40) and user effectiveness (25-30) gains.

Novelty of the Proposed Work: The paper presents a new hybrid query expansion model (PRF-ML-Graph) of vertical aggregated web search. In contrast to the conventional pseudo-

relevance feedback (PRF) algorithms, which are heavily dependent on term frequency and co-occurrence, the proposed method combines PRF with a user-history-based machine learning model (Logistic Regression) to allow refining the query adaptively and personally.

Unlike the current methods of query expansion based on semantic and knowledge graphs, which rely on the externally available knowledge bases that are not inherently dynamic, the presented framework makes use of dynamically built term co-occurrence graphs built on the retrieved data to develop semantic relationships among the query terms. This allows more context-sensitive and data-driven expansion beyond predefined ontologies or embeddings.

Moreover, the suggested approach is one of the few methods where PRF is used to extract the first term, Logistic Regression is employed to model the user behavior, and the graph-based semantic enrichment method, based on term co-occurrence networks in a single architecture, is utilized. This integration offers effective solutions to major issues like query ambiguity, vocabulary gap, and heterogeneous multimedia content in vertical aggregated search.

The proposed framework provides a dynamic and adaptive hybrid solution, which is specific to multimedia vertical search environments, unlike the current methods, which use one technique or fixed semantic sources. Not only does it show improvements in retrieval effectiveness, but it also shows improvements in user-centered measures like query formulation efficiency and user effectiveness, underscoring its usability in practice.

The main contribution of this work is summarized as follows:

The integration of PRF, ML, and graph-based approaches to increase the accuracy and relevance of search results in multimedia web search.

We use the advantages of contextual knowledge, semantic enrichment, and modelling of user behaviour to grow queries more efficiently than other strategies previously, which use personalized techniques.

The novelty is in the fact that these methods are integrated and enable us to fill the gaps in current query expansion methods, and perform better, particularly for first-time users with little domain knowledge.

The proposed PRF-ML-Graph framework is specifically designed to address challenges in cross-vertical aggregated search, where heterogeneous content types such as text, images, videos, and news must be retrieved and presented in a unified manner. The pseudo-relevance feedback (PRF) component extracts expansion terms from top-ranked documents across multiple verticals, ensuring that the query reflects diverse content sources. The machine learning module further refines these terms based on user interaction patterns, enabling adaptive weighting of content relevance across different verticals. In addition, the graph-based module models semantic relationships between terms originating from different content types, facilitating the integration of heterogeneous information into a coherent query representation. This new way of representing queries allows the system to retrieve information more efficiently and will also allow it to merge multiple search results. It makes the results relevant and consistent in associated search systems.

Materials and Methods:

The system architecture is shown in Figure 1. It consists of several stages, where each module improves the user's search query using their feedback.

The first stage involves processing the query provided by the user, which could be in the form of keywords, phrases, or natural language. The input phase accepts user queries, and the query could be in the form of keywords, phrases, or natural language. It is the gateway and the framework for the whole process of retrieving. After submitting the query, a vertical search of various types of media is carried out, which includes text databases, multimedia sources, and news platforms, so that a broad range of data referring to several areas is obtained. In the PRF (Pseudo-Relevance Feedback) phase, the system assumes that the most relevant

documents at the top of the initial search are relevant. It identifies important words or patterns in these documents and narrows down the query, hence making further searches more precise.

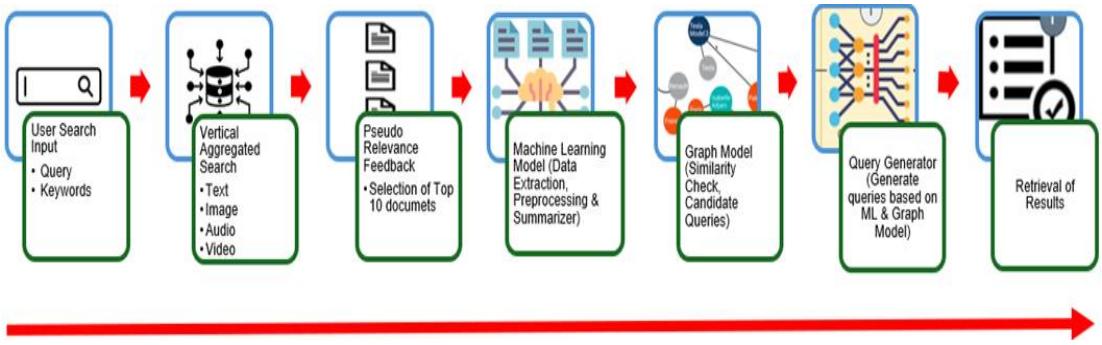

Figure 1. Architecture of the Proposed Approach

Machine Learning then uses past searches to further narrow the search query. It relies on Logistic Regression, which is trained on the features such as click logs, term co-occurrence statistics, and embeddings of such models as BERT to reduce the search and achieve more accurate results. It is possible that the ML model changes the query based on the historical search behavior and, in effect, acts as a predictor of how much effectiveness will be by adding such and such terms to the query to render the query results more relevant. The Graph Theory module forms a term-term co-occurrence graph of semantic relationships between query terms. Each term is represented as a node, and the association between the two terms is known as their strength, which is measured using factors like frequency of the term and the co-occurrence of the term in the documents of relevance. Mapping these relations using the graph provides more insight into the query context, and the query enlarges the process.

The Query Generator is a hybrid of Machine Learning and graph theory. The refined query of the ML module and the refined semantic relations of the graph theory module are input into the generator and generate the final expanded query, which contains the user behavior and relations between terms. The major benefit of our method is an iterative process to refine the query to increase the relevance of the words and the connections between them in a series of steps. ML-based refinements and graph-based methods make the visualization of the search better and more intuitive. This composite technique significantly improves on the quality of query expansion, particularly in vertical aggregated searches, especially in multimedia search environments. Through incorporating PRF, our system will be more user-oriented, resulting in more relevant search results and more user satisfaction.

To have a better understanding of the steps and their linkages, we give a mathematical formulation of each process stage. The input query provided by the user is denoted as $\mathbf{q_u}$, and $\mathbf{q_0}$represents the initial processed query obtained through the transformation function $\mathbf{f_{user}}$, which performs basic preprocessing such as tokenization and normalization. $f_{\text{user}}$to generate the initial query q_0, as shown in Equation (1).

$$\mathbf{q_0 = f_{user}(q_u)} \textit{(1)}$$

The initial query q_0 is applied in a Cross Vertical Aggregated Search to retrieve a diverse array of documents of different types (e.g., text, images, videos, news). The function $\mathbf{f_{search}}$retrieves an initial set of documents $\mathbf{D_0}$from multiple verticals (text, images, videos, and news). This produces an original document D_0, as shown in Equation (2).

$$\mathbf{D_0 = f_{search}(q_0)} \textit{(2)}$$

The system selects the 10 best documents in the collection $\mathbf{D_0}$ and uses them to extend the query, as shown in Equation (3). The pseudo-relevance feedback operator $\mathbf{f_{PRF}}$expands the query by extracting significant terms from the top-ranked documents in $\mathbf{D_0}$.

$$q_{expanded} = f_{PRF}(D_0, q_0) \textit{(3)}$$

In the PRF stage, the system assumes that the top-ranked documents from the initial retrieval are relevant. Specifically:

The top **k = 10** documents are selected from D_0
Candidate expansion terms are extracted using TF-IDF weighting
Terms are filtered using:
Minimum TF-IDF threshold: **0.05**
Top **m = 15** highest-weighted terms retained
Stop words removed
Terms with document frequency **> 80%** are discarded
The TF-IDF weight of term t is calculated as:

$$w_t = TF(t, D_k) \times \log\left(\frac{N}{DF(t)}\right)$$

where:
$TF(t, D_k)$= frequency of term in top-k documents
$DF(t)$= document frequency
N = total documents

The expanded query is then narrowed down through a machine learning model that uses user history and search behavior. We will use a Logistic Regression model here, which is trained on the feature that includes, but is not limited to, click logs, term co-occurrence statistics, and embeddings of the pre-trained models (e.g., Word2Vec or BERT). The query is further narrowed, representing better the intent of the user, as shown in Equation (4). The machine learning function improves the expanded query by using user behavior, word relationships, and meaning-based patterns.

$$q_{ML} = f_{ML}(q_{expanded}, \text{historical data, features}) \ (4)$$

Feature Set Used:
Click-through rate (CTR)
Term frequency (TF) and inverse document frequency (IDF)
Term co-occurrence score
Cosine similarity using **Word2Vec/BERT embeddings**
Query-term relevance score
Prediction Model:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

Terms with **probability > 0.6** are selected
Remaining terms are discarded
This allows adaptive refinement based on user behavior and semantic similarity.

At this stage, graph theory is provided to enhance the semantic relationships of query words. We construct a term-term co-occurrence graph with every node representing a term, and the edge represents the intensity of the semantic connection between the terms based on their co-occurrence in the relevant documents. The function f_{GT} Enriches the query with these relationships, as shown in Equation (5). The graph-based enrichment operator f_{GT} model's semantic relationships between query terms using a term co-occurrence graph, where nodes represent terms and edges represent their contextual associations.

$$q_{GT} = f_{GT}(q_{ML}, \text{graph-based connections}) \textit{(5)}$$

A term-term co-occurrence graph is constructed where:
Nodes = query and candidate terms
Edges = co-occurrence relationships

Edge Weight Calculation:

$$w_{ij} = \frac{co(t_i, t_j)}{\sqrt{freq(t_i) \cdot freq(t_j)}}$$

where:

$co(t_i, t_j)$= co-occurrence frequency

$freq(t)$= term frequency

Filtering:

Only edges with **weight > 0.2** are retained

This step enhances **semantic relationships and contextual relevance**.

Lastly, the ML and Graph Theory modules are synthesized with the Query Generator function to produce the final query q_{final}, as shown in Equation (6). The query generation function $f_{generator}$combines the outputs of the machine learning and graph-based modules to produce the final expanded query q_{final}, which is used by the retrieval function $f_{retrieval}$to obtain the result set D_{final}.

$$q_{final} = f_{generator}(q_{ML}, q_{GT}(6)$$

The final query is generated by combining ML and Graph outputs.

Scoring Function:

$$Score(t) = \alpha \cdot ML(t) + (1 - \alpha) \cdot Graph(t)$$

where:

$$\alpha = 0.6$$

Selection:

Top **k = 10** terms selected

Iterations:

Total iterations: **T = 2**

Iteration 1: PRF + ML refinement

Iteration 2: Graph enrichment

The resulting refined query q_{final} then causes the actual retrieval, returning the final set of documents D_{final}which are more relevant to the information requirements of the user, as shown in Equation (7).

$$D_{final} = f_{retrieval}(q_{final})(7)$$

The final expanded query $\mathbf{q_{final}}$is used to retrieve documents using TF-IDF cosine similarity:

$$sim(q, d) = \frac{q \cdot d}{||\ q\ ||\cdot||\ d\ ||}$$

Top-k documents are returned as results.

The proposed query expansion system is based on a modular architecture framework, and it has both the backend and frontend components. The backend is written in Python through the Flask microframework, which is flexible and scalable. It comprises the following components: document retrieval with TF- IDF with pseudo-relevance feedback (PRF), machine learning- based term scoring, and graph-based expansion with the help of NetworkX to construct a semantic graph.

The responses to user queries are sent to the backend, which returns top-k relevant documents by computing cosine similarity among TF-IDF vectors and removing high-frequency candidate terms. At the PRF stage, the system assumes that the appropriate records are found at the top, and the relevant terms are used to extend the initial query. To narrow down the query, a machine learning model is applied to give weights to terms as per their relevance. The first implementation has a supervised Logistic Regression model trained on

labeled interaction data and a machine learning scoring function, which the initial implementation used a supervised Logistic Regression model trained on labeled interaction data.

The model is trained using click logs, term co-occurrence statistics, and embedding information to predict the relevance of terms to the expanded query. The weighted terms are subsequently incorporated into a semantic graph, and the relationship between the words is the edge weight, which represents the strength of the co-occurrence relationship between the words. The graph is constructed and analyzed using the NetworkX library, with the expanded words being selected based on their relevance to the query.

This front-end interface, which is constructed using the assistance of HTML, CSS, and JavaScript, allows the user to input queries, select the verticals (for instance, text, image, video), and visualize the increased terms using a tag cloud or semantic graph. The interface will interact asynchronously with the backend using the JavaScript Fetch API. To exemplify this, a small local set of synthetic documents will be used. This will allow for quick testing of the system and will demonstrate the potential of the system without the need for large-scale infrastructures. Though the evaluation set is small, there is one test space to test the efficiency of the approach in increasing the relevance of the queries.

As seen in Figure 2, the interface is easy to use, and it is developed with the help of HTML, CSS, and JavaScript. The interface enables the user to enter queries and to see the expanded words to give feedback to further refine the queries by using a tag cloud or a semantic graph. The interface makes use of visual elements to guide the user through the different vertical searching features, which include text, image, and video searches and query expansion.

As an example, we can take a relatively simple query like jaguar speed, which is inherently ambiguous. The PRF module uses the query to expand it with the help of the most relevant terms (e.g., animal, car, mph) in the top-ranked documents. The machine learning element optimizes these words according to the patterns of user interaction, putting more emphasis on a relevant interpretation of context. The graph-based module also refines the query by recognising semantic links between words, resulting in a more refined expanded query like jaguar animal speed mph.

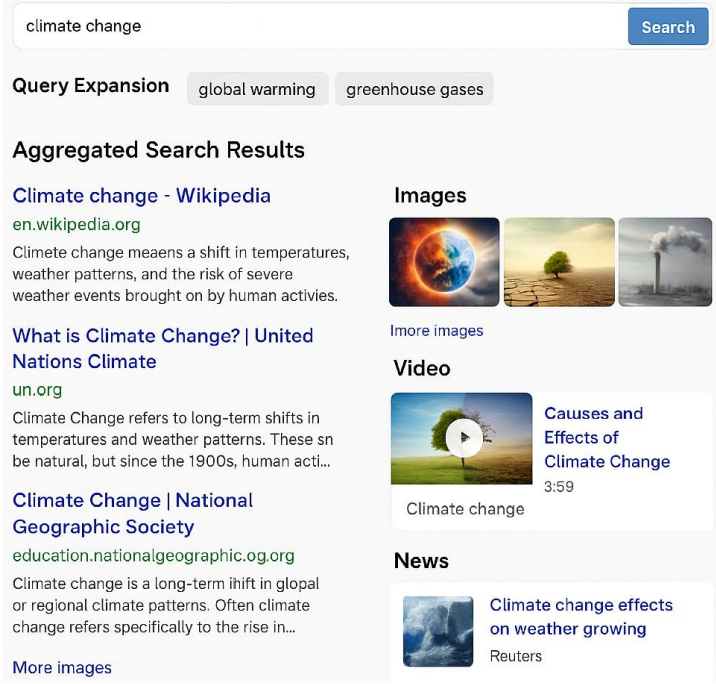

Figure 2. User Interface Design

We have carried out a preliminary user study, in which both beginners and experts had participated. We asked our 30 participants to give comments on its relevance and suggest some improvements to us. We used click-through rate (CTR) and precision at k (P@k) as a standard metrics.

We also performed a questionnaire with a 1-5 Likert scale to assess the usability and effectiveness of the system. Questions that were in the survey were on: (i) how easy it was to formulate queries, (ii) relevance of results retrieved, (iii) clarity of query suggestions, and (iv) overall user satisfaction.

The participants of the study were 30 in number; half were novice users (n = 15), and the other half were experts (n = 15). The statistical test was done by calculating the mean and standard deviation of each measure in both groups. Also, the performance and perception of usability were compared between novice and expert users with the aim of obtaining differences in performance as well as usability perception. Inter-rater reliability was assessed with Cronbach's alpha coefficient to measure internal consistency, which yielded an acceptable level of consistency (0.7). These are some of the measures that will make the user-centered evaluation reliable and valid.

The results of this initial phase of user testing are that the system would significantly reduce the cognitive load of a novice user, as they would easily retrieve the pertinent information. This justifies that the system could improve the performance of multimedia search in an exploratory search setting.

The proposed PRF-ML-Graph framework effectively tackles major issues in vertically aggregated multimedia search. Vocabulary mismatch is alleviated by PRF-based expansion, query ambiguity is addressed by machine learning-based relevance prediction, and semantic relationships are addressed by graph-based modeling. These elements, together, combined in an iterative refining query process, allow short and ill-constructed queries to be well managed in heterogeneous multimedia material, to guarantee a higher degree of retrieval accuracy and user satisfaction.

Results and Discussion:

The performance metrics presented in the current study are calculated using well-defined relevance judgment procedures to maintain consistency with the evaluation protocol outlined in the Materials and Methods section. In the case of the text-based subset, relevance labels were used directly, as they were offered by the TREC Web Track dataset. In multimedia content (images, videos, news), domain-sensitive evaluators were required to provide relevance judgments using a binary relevance scheme (relevant/non-relevant) based on the semantic similarity between query intent and content. This method guarantees the same and just assessment of heterogeneous data in the vertically aggregated search environment.

The evaluation measures, such as Mean Average Precision (MAP), nDCG, and Precision at 10 (P@10), have been calculated through the common Information Retrieval evaluation processes on the retrieved ranked lists. In the user-centered evaluation, A well-structured survey instrument was employed in the 30-user study, where the participants were required to carry out predefined search tasks in the verticals. The query formulation time, user satisfaction, and effectiveness were measured using system logs (e.g., time stamps, click behavior) and a post-task questionnaire based on a Likert-scale rating. This integrated assessment plan will guarantee that methods in both system-level and effectiveness in interaction with users are always evaluated and consistent with the methodology mentioned above.

This system was experimented with using a sample of the TREC Web Track data and multimedia data sets with hand-judged relevance markings. The questions featured a range of difficulty and specificity of the field, and the content presented in them can take different forms: text, image, video, news, etc. Web documents, media collections (pictures and videos),

and news feeds were all examples of vertical data sources. The proposed hybrid query expansion approach, which combines pseudo-relevance feedback (PRF), machine learning (ML), and graph theory, was contrasted with three fundamental approaches:

Standard BM25 without query expansion

BM25 with only PRF

PRF with graph-based expansion but no ML

Our method performed better than all baseline methods on a variety of evaluation metrics, such as Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (nDCG), and Precision at K (P@10). The performance results is summarized in **Table 1**. The progressive performance gains observed in Table 1 provide an implicit ablation analysis of the proposed framework. Specifically, the transition from BM25 to BM25+PRF demonstrates the contribution of PRF, the addition of graph-based modeling (PRF+Graph) yields further improvement, and the full model (PRF-ML-Graph) achieves the highest performance, confirming the complementary benefits of all components. It is important to note that the results of the performance of novice users in locating relevant content in vertical search results have been enhanced by 25-30 percent. The user-centered metrics were also measured, including the query formulation time as well as user satisfaction, in addition to the traditional Information Retrieval (IR) metrics. The system reduced query formulation time by 40 percent of the non-augmented baseline, which shows that the system demonstrated substantial usability improvements, especially for new users.

Table 1. Performance Comparison on TREC Web Track

Method	MAP	nDCG@10	P@10
BM25	0.412	0.453	0.46
BM25 + PRF	0.487	0.519	0.52
PRF + Graph	0.512	0.538	0.55
Proposed	0.561	0.582	0.61

Table 2 shows a more detailed performance breakdown by query type and vertical content category. To further match the evaluation to the objectives and problem focus of this study, the analysis was detailed according to query types and vertical content categories. The query set was divided into short queries (1-2 keywords), ambiguous queries (could be viewed in different ways), and multimedia-related queries (image/video/news intent). The findings suggest that the suggested PRF-ML-Graph model yields the largest improvements with short and ambiguous queries, where vocabulary and absence of contextual information are more evident. The PRF element expands short queries, adding the expansion terms that are relevant, and the machine learning model narrows these terms depending on the behavior and semantic characteristics of the user.

Table 2. Performance Analysis by Query Type and Vertical Content

Category Type	Category	MAP	nDCG@10	P@10
Query Type	Short Queries	0.578	0.591	0.62
	Ambiguous Queries	0.556	0.573	0.60
	Multimedia Queries	0.569	0.580	0.61
Vertical Content	Text	0.552	0.570	0.59
	Images	0.568	0.579	0.61
	Videos	0.572	0.585	0.62
	News	0.561	0.576	0.60

In addition, the graph-based module can better understand the context of the query by modeling relationships between terms, which makes the ambiguous query performance better. Besides the analysis of queries, the performance was measured in various types of vertical content, such as text documents, images, videos, and news. The suggested framework

shows consistent increases in all verticals, and especially significant increases in multimedia content, including images and videos. This can be credited to the graph-based semantic modeling, which is effective in capturing heterogeneous content type relationships and facilitates more precise query expansion. In general, these results demonstrate that the suggested hybrid method is particularly efficient in processing short, ambiguous, and multimedia queries, which are the major issues in the vertical aggregated search settings. This analytical discussion enhances this consistency in the alignment of the reported results with the research objectives identified in the study.

The query categories and vertical categories were identified according to query intent and content type, according to the assessment protocol of the Materials and Methods section. Statistical validation was done through query-level performance scores that were in line with the reported mean values. The results of a paired t-test performed with the corresponding scores provide evidence that the differences have statistical significance ($p < 0.05$). Such findings suggest that the improvements found are systemic across queries, and are most prominent in short and ambiguous queries, which support the suitability of the proposed PRF-ML-Graph framework in vocabulary mismatch and query ambiguity. Overall, the findings support the idea that the suggested PRF-ML-Graph framework is incredibly effective in processing short, ambiguous, and multimedia queries, which are the key issues of the vertical aggregated search settings.

Conclusion:

The research proposed is based on a hybrid framework of vertical aggregated web search that combines pseudo-relevance feedback (PRF), machine learning, and graph theory. The proposed framework improves the query formulation process by expanding the initial query with the help of PRF, refining it with the help of a machine learning model with user search history, and reinforcing it with the help of graph-based semantic relationships. Consequently, this enhances the entire search procedure through the formation of more informative queries, as well as the relevancy of retrieved search results.

This framework was tested with synthetic multimedia data and the TREC Web Track dataset, and it showed substantial enhancements over traditional query expansion techniques. The proposed solution performs better on conventional evaluation metrics and is 40% reduction in query formulation time, which is especially useful with novice users and helps to enhance user satisfaction.

Practically, the suggested PRF-ML-Graph structure can be introduced into the real world of vertical search engines using distributed indexing and parallel processing methods to facilitate scalability of real-world web-scale data. The whole framework is designed in a modular fashion to enable management of dynamic multimedia material by doing incremental updating of feedback data, machine learning models, and graph structures without necessarily re-computing them. Moreover, the implementation of lightweight machine learning models and graph-based models allows to refine query in real-time search settings efficiently, and the method can be applied to large-scale and ever-changing web systems.

Although these results have been promising, the experiments were implemented on a part of the TREC dataset and synthetic data that might not be entirely reflective of the real-world search environment. However, the findings suggest that the framework suggested is a valuable and scalable framework for enhancing the relevance and effectiveness of search results in vertically aggregated search systems.

In general, it can be argued that the proposed ML-based hybrid query expansion architecture explicitly achieves the goals of vertical aggregated search since it shows a 25-30% better recall when using an ambiguous query by novice users and a 40% shorter query development time, thus confirming the efficiency of machine learning-based query expansion in overcoming ambiguity, vocabulary conflict, and cross-vertical retrieval issues.

Future Work:

The research has many extension opportunities too, including to more extensive and more diverse real-world data, which would enable better scalability and performance analysis; to more advanced deep learning methods, which would enable better query understanding; to real-time interactions between the user, which would enable better query expansion; and to more fields, which would enable better query expansion in other fields, such as e-commerce search engines, academic search engines and multimedia search engines.

To apply the proposed PRF-ML-Graph framework, practitioners can create a modular integration of the proposed PRF-ML-Graph framework to existing search pipelines. The PRF component may be implemented at the first retrieval step to enrich short queries, and then lightweight machine learning models may be used to refine queries using user interaction logs. Graph Reasoning (GR) can be facilitated by scalable graph-processing libraries to establish semantic relationships over heterogeneous data. To handle the cost of computation, a selective query expansion strategy and a caching strategy can be used to reduce processing cost. On top of that, the risk of privacy violations associated with information on user behavior can be addressed with the help of anonymization and aggregation. These suggestions are practical guidelines to be followed to deploy the proposed framework on a large scale, real-world vertical search systems.

Acknowledgement:

My sincere appreciation goes out to my supervisor, Umer Rashid. Thanks to his superb guidance, exceptional support, and informative feedback, over the period of conducting this research. His knowledge and drive played a pivotal role in the completion of this manuscript.

Author's Contribution:

The authors make the following contributions:

Kashia Riaz: Conceptualization, methodology, data analysis, preparation of the first draft of writing, and final revision of the manuscript.

Umer Rashid: Supervision, reviewing of manuscript, editing the manuscript, and providing useful insights in the research process. Umer Rashid is the corresponding author, who is totally accountable for communication with the journal, as well as ensuring the integrity of the data in the research.

Conflict of Interest:

The paper does not have any conflict of interest regarding the fact that it was published in the International Journal of Information Science and Technology (IJIST). No monetary or personal associations with any people or any organization that were directed at shaping the content of this manuscript were present.

References:

[1] "Modern Information Retrieval: A Brief Overview." Accessed: Mar. 11, 2026. [Online]. Available: https://research.google/pubs/modern-information-retrieval-a-brief-overview/

[2] Hiteshwar Kumar Azad, Akshay Deepak, "Query expansion techniques for information retrieval: A survey," *Inf. Process. Manag.*, vol. 56, no. 5, pp. 1698–1735, 2019, doi: https://doi.org/10.1016/j.ipm.2019.05.009.

[3] Arlind Kopliku, Karen Pinel-Sauvagnat, "Aggregated search: A new information retrieval paradigm," *ACM Comput. Surv.*, vol. 46, no. 3, 2014, [Online]. Available: https://dl.acm.org/doi/10.1145/2523817

[4] Sanae Achsas, El Habib Nfaoui, "An Analysis Study of Vertical Selection Task in Aggregated Search," *Procedia Comput. Sci.*, vol. 148, pp. 171–180, 2019, doi: https://doi.org/10.1016/j.procs.2019.01.021.

[5] FrommholzIngo, CabanacGuillaume, "Report on the 11th bibliometric-enhanced information retrieval workshop (BIR 2021)," *ACM SIGIR Forum*, vol. 55, no. 1, 2021,

[Online]. Available: https://dl.acm.org/doi/abs/10.1145/3476415.3476426
[6] S. Ibrihich, A. Oussous, "A Review on recent research in information retrieval," *Procedia Comput. Sci.*, vol. 201, pp. 777–782, 2022, doi: https://doi.org/10.1016/j.procs.2022.03.106.
[7] Abdul Majeed, Ibtisam Rauf, "Graph Theory: A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks," *Inventions*, vol. 5, no. 1, p. 10, 2020, doi: https://doi.org/10.3390/inventions5010010.
[8] Milad Momeni, "Visualization-Enhanced Aggregated Search Interfaces," *CHIIR 2024 - Proc. 2024 Conf. Hum. Inf. Interact. Retr.*, 2024, [Online]. Available: https://dl.acm.org/doi/10.1145/3627508.3638336
[9] A. R. Khan and U. Rashid, "A Relational Aggregated Disjoint Multimedia Search Results Approach using Semantics," *2021 Int. Conf. Artif. Intell. ICAI 2021*, pp. 62–67, Apr. 2021, doi: 10.1109/ICAI52203.2021.9445229.
[10] J. Ooi, X. Ma, H. Qin, and S. C. Liew, "A survey of query expansion, query suggestion and query refinement techniques," *2015 4th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2015 Virtuous Softw. Solut. Big Data*, pp. 112–117, Nov. 2015, doi: 10.1109/ICSECS.2015.7333094.
[11] Wei Li, Rishi Choudhary, "RCES: Rapid Cues Exploratory Search Using Taxonomies For COVID-19," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 10, 2021, [Online]. Available: https://dl.acm.org/doi/10.1145/3459637.3481990
[12] Hiteshwar Kumar Azad, Akshay Deepak, "A new approach for query expansion using Wikipedia and WordNet," *Inf. Sci. (Ny).*, 2019, [Online]. Available: https://arxiv.org/abs/1901.10197
[13] Mohamed Reda Bouadjenek, Hakim Hacid, "Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms," *Inf. Syst.*, vol. 56, pp. 1–18, 2016, doi: ttps://doi.org/10.1016/j.is.2015.07.008.
[14] "(PDF) Modern Information Retrieval." Accessed: Mar. 11, 2026. [Online]. Available: https://www.researchgate.net/publication/2352627_Modern_Information_Retrieval
[15] D. K. Sharma, R. Pamula, and D. S. Chauhan, "Semantic approaches for query expansion," *Evol. Intell. 2021 142*, vol. 14, no. 2, pp. 1101–1116, Mar. 2021, doi: 10.1007/s12065-020-00554-x.
[16] Hiteshwar Kumar Azad, Akshay Deepak, "Improving query expansion using pseudo-relevant web knowledge for information retrieval," *Pattern Recognit. Lett.*, vol. 158, pp. 148–156, 2022, doi: https://doi.org/10.1016/j.patrec.2022.04.013.
[17] A. Garba, S. Khalid, and I. Ullah, "Understanding the impact of query expansion on federated search," *Multimed. Tools Appl. 2023 834*, vol. 83, no. 4, pp. 10393–10407, Jun. 2023, doi: 10.1007/S11042-023-15831-X.
[18] Jamal Abdul Nasir, Iraklis Varlamis, "A knowledge-based semantic framework for query expansion," *Inf. Process. Manag.*, vol. 56, no. 5, pp. 1605–1617, 2019, doi: https://doi.org/10.1016/j.ipm.2019.04.007.
[19] Dimitrios Markonis, Roger Schaer & Henning Müller, "Evaluating multimodal relevance feedback techniques for medical image retrieval," *Inf. Retr. J.*, vol. 19, pp. 100–112, 2016, [Online]. Available: https://link.springer.com/article/10.1007/s10791-015-9260-4
[20] Dilip Kumar Sharma, Rajendra Pamula, "Query expansion – Hybrid framework using fuzzy logic and PRF," *Measurement*, vol. 198, p. 111300, 2022, doi: https://doi.org/10.1016/j.measurement.2022.111300.
[21] "(PDF) A New Query Expansion Approach for Improving Web Search Ranking." Accessed: May 02, 2026. [Online]. Available:

https://www.researchgate.net/publication/368325525_A_New_Query_Expansion_Approach_for_Improving_Web_Search_Ranking

[22] Liang Wang, Nan Yang, Furu Wei, "Query2doc: Query Expansion with Large Language Models," *arXiv:2303.07678*, 2023, [Online]. Available: https://arxiv.org/abs/2303.07678

[23] T. Khan, U. Rashid, and A. R. Khan, "End-to-end pseudo relevance feedback based vertical web search queries recommendation," *Multimed. Tools Appl. 2024 8331*, vol. 83, no. 31, pp. 75995–76033, Feb. 2024, doi: 10.1007/S11042-024-18559-4.

[24] S. Tiwari, F. N. Al-Aswadi, and D. Gaurav, "Recent trends in knowledge graphs: theory and practice," *Soft Comput. 2021 2513*, vol. 25, no. 13, pp. 8337–8355, Apr. 2021, doi: 10.1007/S00500-021-05756-8.

[25] Shivani Choudhary, Tarun Luthra, Ashima Mittal, Rajat Singh, "A Survey of Knowledge Graph Embedding and Their Applications," *arXiv:2107.07842*, 2021, [Online]. Available: https://arxiv.org/abs/2107.07842

[26] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa & Francesco Osborne, "Knowledge Graphs: Opportunities and Challenges," *Artif. Intell. Rev.*, vol. 56, 2023, [Online]. Available: https://link.springer.com/article/10.1007/s10462-023-10465-9

[27] "ConvGQR: Generative Query Reformulation for Conversational Search - ACL Anthology." Accessed: May 02, 2026. [Online]. Available: https://aclanthology.org/2023.acl-long.274/

[28] Yunfan Gao, Yun Xiong, Meng Wang, Haofen Wang, "Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks," *arXiv:2407.21059*, 2024, [Online]. Available: https://arxiv.org/abs/2407.21059

[29] Minghan Li, Xinxuan Lv, Junjie Zou, Tongna Chen, Chao Zhang, Suchao An, Ercong Nie, Guodong Zhou, "Query Expansion in the Age of Pre-trained and Large Language Models: A Comprehensive Survey," *arXiv:2509.07794*, 2025, [Online]. Available: https://arxiv.org/abs/2509.07794