Performance Assessment of an Outlier Detection–Based Approach to Information Retrieval and Query Expansion

Wahid Ali[1,*], Mohd Waris Khan[2]
[1,2]Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India
wahidali@student.iul.ac.in[1], wariskhan070@gmail.com[2]

ABSTRACT: Information retrieval (IR) systems face persistent challenges in addressing the complexity of user queries and large-scale document collections. This study presents a performance assessment of an outlier detection–based approach to query expansion aimed at enhancing retrieval precision and relevance. The proposed Gradient-Based Dynamic Query Expansion (GBDQE) framework integrates outlier filtering to eliminate noisy or irrelevant terms that typically reduce the effectiveness of traditional feedback mechanisms. A series of experiments on standard benchmark datasets were conducted to evaluate the model against baseline methods and established approaches such as WT2G and CWSBQE. Experimental results reveal that the outlier detection–driven GBDQE approach delivers significant improvements, achieving a 15% increase in Mean Average Precision (MAP), a 20% gain in Precision at 10 (P@10), and notable enhancements in F-measure and GM_MAP scores. These outcomes confirm the robustness of the proposed method and highlight its potential to substantially improve IR performance, thereby advancing the quality of user search experiences in modern digital environments.

Keywords: Information retrieval, feedback model, outlier detection, RF.

1. INTRODUCTION

Information retrieval (IR) is a core discipline within computer science that focuses on the representation, storage, organization, and retrieval of data from large and often heterogeneous collections [1]. The goal of an IR system is to provide users with accurate and relevant results in response to their queries by employing processes such as document representation, indexing, searching, filtering, and ranking. The performance of an IR system is typically assessed through two key measures: precision, which reflects the proportion of retrieved documents that are relevant, and recall, which indicates the proportion of relevant documents successfully retrieved [2]. Achieving both high precision and recall remains a persistent challenge in the design and implementation of modern IR systems.

A user's query may be expressed in the form of text, images, or multimedia content, and IR systems must interpret this input to deliver relevant results. However, queries are often ambiguous or incomplete, which can hinder the retrieval of precise information. To address this limitation, researchers have developed techniques such as query expansion, where additional terms are incorporated into the user's query to better capture the underlying intent [3]. A particularly important technique within query expansion is Relevance Feedback (RF) [4]. RF allows users to identify relevant or irrelevant documents from an initial retrieval set, and this feedback is used by the system to iteratively refine the query. By reformulating the query and retrieving new results, RF creates a feedback loop that can significantly improve retrieval accuracy [5, 6].

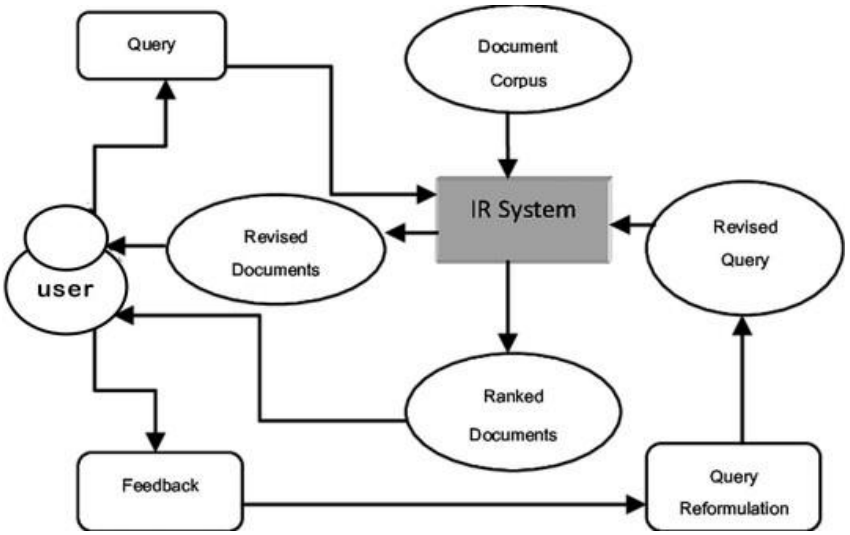

Fig. 1: The basic architecture of Feedback

While relevance feedback has been shown to enhance retrieval performance, it also introduces challenges. One major issue is the inclusion of noisy or irrelevant terms during query expansion, which can reduce precision and distort the ranking of results. Traditional RF methods often lack mechanisms to distinguish between truly informative terms and outliers, leading to suboptimal outcomes. This limitation has motivated researchers to explore approaches such as probabilistic term weighting, synonym-based query modification, document clustering, and indexing strategies. Although these methods provide improvements, they remain constrained by their inability to effectively filter anomalous terms.

To overcome these challenges, this study proposes a feedback-driven model that integrates outlier detection into the query expansion process. The proposed Gradient-Based Dynamic Query Expansion (GBDQE) method identifies and eliminates extraneous or noisy terms that would otherwise degrade retrieval performance. By filtering outliers, the model ensures that expanded queries remain both precise and contextually aligned with the user's intent.

The performance of the GBDQE model was rigorously evaluated against baseline approaches and established methods, including WT2G and CWSBQE. The experimental results demonstrate consistent improvements across multiple evaluation metrics, including Mean Average Precision (MAP), Precision at 10 (P@10), F-measure, and GM_MAP, highlighting the effectiveness of incorporating outlier detection into query expansion.

The remainder of this paper is structured as follows: Section 2 presents a detailed overview of Information Retrieval Systems (IRS). Section 3 presents a comprehensive review of related work, focusing on feedback mechanisms, query expansion techniques, and indexing methods in information retrieval. Section 4 defines the problem statement, while Section 5 describes the proposed methodology and algorithms, including the experimental setup, datasets, and evaluation metrics. Section 6 reports the results along with a detailed discussion. Finally, Section 7 concludes the paper by summarizing key findings, highlighting implications, and outlining future research directions.

2. INFORMATION RETRIEVAL SYSTEM (IRS)

An Information Retrieval System (IRS) is designed to represent, store, and retrieve relevant information from large collections based on user queries. These consist of the following:

A. Framework for IRS:

As stated in [7], an IRS must facilitate three fundamental processes: (i) the representation of document content, (ii) the representation of the user's informational requirements, and (iii) the comparison of the two representations.

These processes are illustrated in Figure 2, where square boxes denote data and rounded boxes represent operations. The representation of documents is generally referred to as indexing, while the representation of user information needs is generated through query formulation.

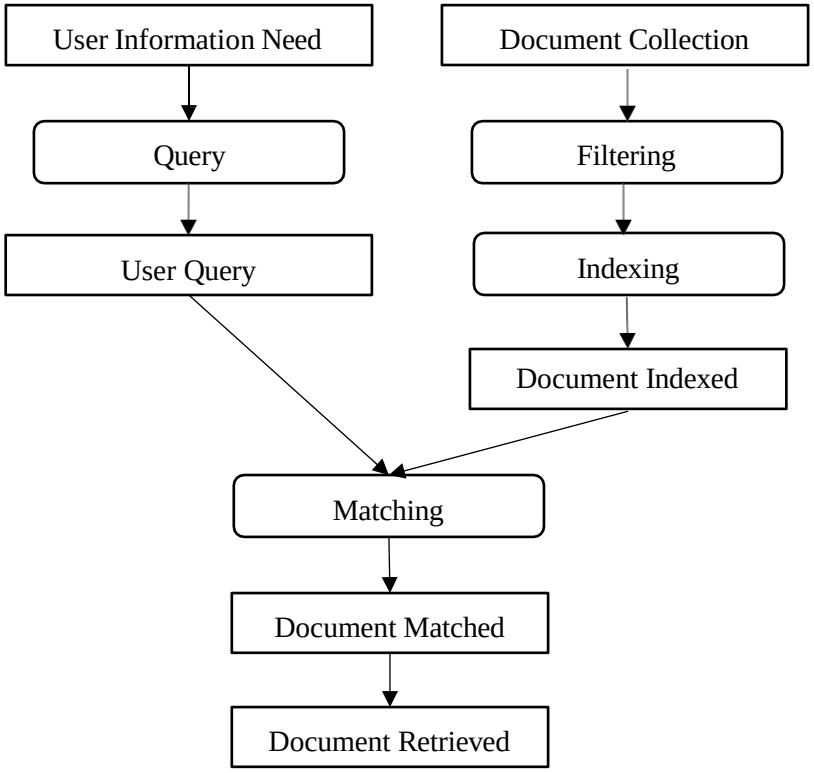

Fig. 2: A general framework for IRS

The indexing procedure is typically performed offline, without direct user involvement, whereas the query formulation produces the query input that reflects the user's information demand [8, 9].

B. Information Retrieval Models:

Mathematical models are widely used in scientific domains to explain and predict real-world behaviors [10, 11]. In the context of IR, models help forecast what a user may consider relevant to a query. Their predictive power can be validated through controlled experiments. An IR model thus serves as a theoretical and practical framework for building IRSs capable of ranking documents effectively.

2.1 Techniques:

- **Term Weighting:**

In probabilistic models, weighting methods rely heavily on accurate estimation of various probabilities [12, 13]. Term weighting serves as a technique to identify and extract the most important information necessary for ranking documents in information retrieval systems. Over time, several approaches for term weighting have been developed in the field, with probabilistic models particularly emphasizing precise probability estimation.

- **Query modification using Synonyms:**

In the early development of information retrieval, researchers observed that users often struggled to construct effective queries. To address this, it was suggested that adding synonyms of query terms could improve retrieval performance. Early studies employed thesauri to identify synonyms, but obtaining high-quality, general-purpose thesauri proved costly. Consequently, researchers developed methods to automatically generate thesauri for query modification, typically through word co-occurrence analysis within documents, which produced lists of closely related terms. However, most automated query expansion techniques showed limited improvement in retrieval effectiveness, primarily due to the lack of contextual understanding during augmentation—since not all related terms hold meaningful relevance to the original query.

- **Relevance feedback for Query modification:**

In information retrieval systems, the indexing stage preprocesses documents and queries to extract relevant keywords. At this point, techniques such as stemming and the use of stopword lists are applied—stemming reduces words to their root forms, while stopword lists remove terms with little semantic value. Phrase weighting is then used to measure the similarity between queries and documents. Based on these methods, most Information Retrieval Systems (IRS) generate a ranked list of documents, with those most closely matching the query appearing at the top.

- **Document Clustering:**

Over the years, numerous strategies have been developed, showing varying levels of success. This process entails grouping similar documents to enable faster and more efficient information retrieval. It represents just one approach for organizing documents to improve access to large databases. According to the clustering hypothesis, documents that are closely clustered tend to share similar relevance patterns for a particular query [14, 15]. Document clustering methods have long been, and continue to be, a major area of research. Although their direct impact on improving search effectiveness or efficiency is limited, they have contributed to other advances in information retrieval, including enhanced browsing and search interface designs.

- **Indexing:**

In the context of retrieval and ranking, the term "indexing" carries a specific meaning. It is defined by experts as “a compilation of terms with references to locations where information regarding documents can be accessed” [16]. Indexing entails creating a data structure that enables fast text searching [17] or the process of assigning index terms to documents, which are the items intended for retrieval [18].

2.2 Types of Feedback:

The main categories of feedback are explicit feedback, implicit feedback, and pseudo feedback [19], which are defined as follows:

- **Explicit Feedback:**

In explicit feedback, users actively indicate the relevance of documents retrieved in response to a query. There are two main approaches for signaling relevance: binary and graded relevance frameworks. In the binary approach, a document is marked as either relevant or irrelevant. In the graded approach, the relevance of a document is expressed using descriptions, specific terms, or numerical scales (e.g., 1 for relevant, 2 for irrelevant, etc.), which is known as Graded Relevance Feedback (Graded RF).

- **Implicit Feedback:**

Implicit feedback is inferred from users' interactions with documents, such as which documents they view or ignore, and the duration of browsing or reading. Since it is derived indirectly from user behavior, this type of feedback does not require explicit input and is therefore termed implicit feedback.

- **Pseudo Feedback:**

Pseudo Feedback is also known as blind relevance feedback. The human component of the feedback is automated in this procedure, ensuring users receive enhanced outcomes. In this approach, the system assumes that the top k ranked documents are relevant and applies a relevance feedback mechanism based on this assumption, thereby enhancing the overall outcome for the user.

3. REVIEW OF LITERATURE

This section critically examines existing research, frameworks, and methodologies pertinent to the current study, highlighting gaps and trends in the field.

Wang et al., (2024) [20] noted that as multimedia platforms like TikTok and YouTube grow, the demand for efficient recommendation systems increases. Existing systems often rely on platform-specific features, limiting transferability. To address this, they proposed TransRec, a model pre-trained on a large dataset to learn directly from mixed-modality (MoM) input using end-to-end training. TransRec supports transfer learning across different platforms and enhances knowledge transfer between modalities. Empirical evaluation across four real-world scenarios demonstrated that MoM feedback can effectively train general-purpose recommender systems.

Ganesh et al., (2024) [21] highlighted that the quality of contextual information in prompts critically influences the reliability of responses from Large Language Models (LLMs). In RAG-based QA systems, response quality often declines with increasing contextual knowledge. To overcome this, they introduced Context Augmented Retrieval (CAR), which combines real-time text classification with LLMs to segment vector databases. CAR delivers high-quality answers while significantly reducing retrieval and response time.

Yu et al., (2023) [22] observed that while LLMs perform exceptionally well on many NLP tasks, they frequently generate inaccurate or fabricated information, limiting their practical applicability. Human intervention can improve accuracy, but it is resource-intensive, time-consuming, and cannot be applied during inference, restricting usability in dynamic settings. To address these challenges, the authors proposed ReFeed, a plug-and-play pipeline that enhances LLMs through automated retrieval feedback without expensive fine-tuning. ReFeed first generates outputs, then uses a retrieval model to extract relevant information from large document repositories, and finally incorporates the retrieved data into the in-context demonstration to improve output quality.

Experiments on four knowledge-intensive benchmark datasets demonstrated that ReFeed improves performance by over 6.0% in zero-shot settings and 2.5% in few-shot settings compared to baselines without retrieval feedback.

Peng et al., (2023) [23] noted that LLMs, such as ChatGPT, can generate fluent, human-like responses for various tasks, including task-oriented dialogue and question answering. However, their deployment in real-world, mission-critical applications is challenging due to hallucinations and limited ability to incorporate external knowledge. To address these issues, the authors proposed LLM-Augmenter, a system that enhances black-box LLMs through plug-and-play modules. The system guides the LLM to generate responses using external knowledge, including task-specific databases, and optimizes prompts based on feedback from utility functions such as factuality scores. Experimental evaluation in task-oriented dialogue and open-domain question answering demonstrates that LLM-Augmenter significantly reduces hallucinations while maintaining fluency and informativeness.

Fei et al., (2022) [24] proposed a novel structure-aware Generalized Language Model (GLM) that leverages syntactic knowledge for unsupervised information extraction (UIE). Their approach employs a heterogeneous inductor to independently generate multiple structural representations through post-training of an existing GLM. A structural broadcaster then combines these latent structures into high-order forests, improving generation during decoding. Finally, a task-oriented structure fine-tuning strategy refines the generated structures to better align with downstream tasks. Experiments across 12 IE benchmarks covering seven tasks demonstrate significant improvements over traditional UIE systems. The study shows that the GLM develops an advanced, task-adaptive structural bias, effectively addressing challenges such as long-range dependencies and boundary detection.

Gupta et al., (2022) [25] observed that a large volume of materials science knowledge is generated and stored as text in peer-reviewed publications. Recent advances in natural language processing, particularly BERT-based models, offer effective methods for information extraction; however, general scientific models may underperform in this domain due to limited exposure to materials science terminology. To address this, the authors developed MatSciBERT, a materials-aware language model trained on a large corpus of peer-reviewed materials science literature. MatSciBERT outperforms SciBERT on three downstream tasks: named entity recognition, relation classification, and abstract classification. The publicly available pre-trained weights enable improved extraction of information from existing literature and facilitate the discovery of novel materials.

Izacard et al., (2021) [26] observed that dense retrievers using neural networks have emerged as alternatives to traditional sparse approaches based on word frequency, showing superior performance on large, well-annotated datasets. However, these models struggle with novel applications lacking training data, where unsupervised methods like BM25 still perform better. The study examined contrastive learning for training unsupervised dense retrievers, demonstrating its effectiveness across several retrieval contexts. On the BEIR benchmark, the unsupervised model outperformed BM25 in Recall@100 across 11 of 15 datasets. Pre-training with contrastive models before fine-tuning on either few in-domain samples or large datasets like MS MARCO further improved performance. The approach also proved effective in multilingual retrieval, showing robust cross-lingual transfer, e.g., retrieving English texts from Arabic queries—tasks unachievable with conventional word-matching techniques.

Fei et al., (2021) [27] focused on biological information extraction (BioIE), including named entity recognition, relation extraction, and event extraction. Current methods, such as conventional neural networks, generic language models, and pre-trained contextualized models, were enhanced by incorporating extensive biological knowledge graphs. They proposed BioKGLM, which employs a three-phase training strategy and diverse fusion mechanisms for improved knowledge integration. Experiments across multiple BioIE tasks show that BioKGLM consistently outperforms state-of-the-art extraction approaches, while also elucidating critical relationships among biomedical concepts.

Guo et al., (2020) [28] highlighted the central role of ranking models in information retrieval and the increasing adoption of deep learning techniques. Neural ranking models, leveraging shallow or deep neural networks, can learn ranking functions directly from raw text, avoiding limitations of manually engineered features. The study provided a comprehensive evaluation of neural ranking models, examining their design principles, learning strategies, and assumptions. Benchmark experiments offered empirical insights into their effectiveness across various retrieval tasks.

Aliannejadi et al., (2019) [29] addressed the challenge of formulating clarifying questions in open-domain conversational information-seeking systems. They developed an offline evaluation framework and created the Qulac dataset via crowdsourcing, containing over 10,000 question-answer pairs spanning 198 TREC topics and 762 aspects. Experiments with an oracle model demonstrated that presenting a single successful query improved retrieval performance by 170% in P@1, underscoring the significance of clarifying questions. They proposed a three-component retrieval framework encompassing question retrieval, selection, and document retrieval.

4. PROBLEM STATEMENT

The exponential growth of digital information has made retrieving accurate and relevant data increasingly challenging across domains such as healthcare, e-commerce, and academic research. Conventional information retrieval (IR) systems often face issues of relevance and redundancy, retrieving either excessive irrelevant data or insufficient pertinent information. These challenges are compounded in the presence of noisy, unbalanced, or high-dimensional datasets. Outliers—data points that significantly deviate from the norm—can further disrupt ranking algorithms, reducing system accuracy. While pseudo-feedback models show promise in enhancing IR by iteratively refining query results, they are often vulnerable to outliers, leading to inefficient feedback loops and degraded performance.

There is an urgent need to develop an innovative pseudo-feedback model that integrates outlier detection techniques with machine learning algorithms for accurate and reliable information retrieval. This approach would leverage advanced outlier detection to mitigate the impact of anomalous data points while applying machine learning to uncover patterns and relationships for improved query reformulation and ranking. Such a paradigm has the potential to significantly improve precision and relevance in IR systems, addressing current limitations and providing a robust solution for high-stakes applications.

5. RESEARCH METHODOLOGY

The methodology section outlines the comprehensive framework of the proposed GBDQE (Gradient-Based Dynamic Query Expansion) model, developed to enhance the precision and relevance of information retrieval (IR) systems. The model integrates an advanced feedback mechanism with outlier detection to refine user queries, aiming to reduce noise and improve retrieval accuracy. Figure 3 illustrates the methodology flowchart.

In query expansion, the first critical step is identifying the source from which potential expansion terms will be extracted. The top-ranked documents retrieved from the initial six user queries serve as the primary resource for selecting candidate expansion phrases. In the context of pseudo-relevance feedback, these top-ranked documents constitute the set of pseudo-relevant documents. Specifically, pseudo-relevant documents include web pages from the top *N* URLs retrieved by three major search engines—Google, Bing, and DuckDuckGo—in response to the initial query.

The relevant terms extracted from this collection of pseudo-relevant documents are then employed for Query Expansion (QE). It is important to note that different search engines may interpret queries differently. For instance, analyzing the top ten search results for the query "apple" shows that Google and Bing primarily interpret it as a corporation, whereas DuckDuckGo identifies it both as a company and as a fruit. To improve the identification and understanding of expansion terms, the methodology incorporates data from all three search engines, ensuring comprehensive coverage of query contexts.

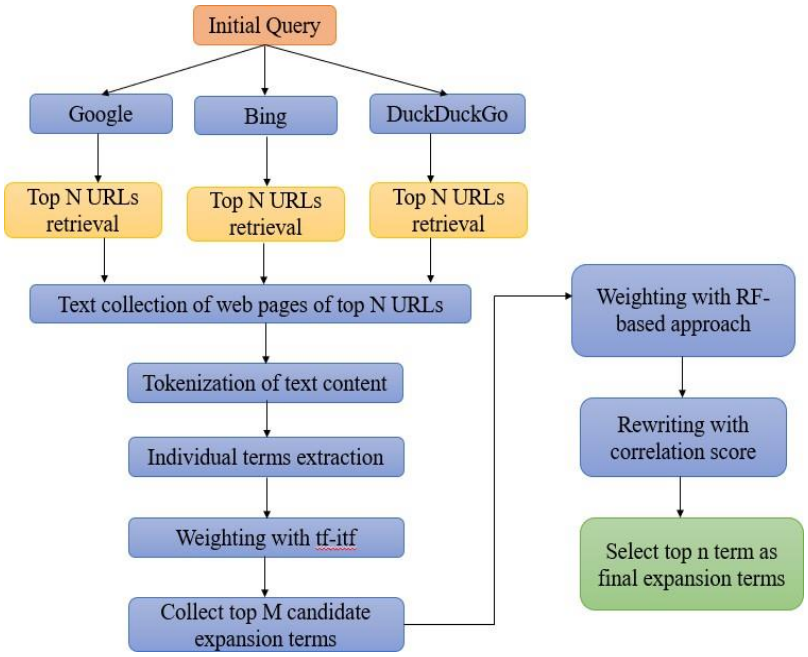

Fig. 3: Flowchart of Methodology

Both term-to-term and term-to-whole-query relationships are analyzed to identify the most relevant candidate expansion terms. For term-to-term evaluation, expansion terms are assessed using the proposed tf-idf and RF-based cosine similarity scores. To account for the relationship between the entire query and its individual components, expansion terms are reweighted according to their correlation scores. The suggested approach, shown in Figure 3, consists of five principal steps: (i) retrieval of the top N URLs, (ii) text collection and tokenization, (iii) weighting using tf-idf, (iv) weighting using an RF-based strategy, and (v) reweighting according to correlation score. The subsequent section delineates these steps.

5.1 Retrieval of Top NURLs:

To broaden the scope of the initial query, we conducted searches using three widely used search engines: Google, Bing, and DuckDuckGo. Each search engine was queried separately, and the web pages corresponding to the top *N* URLs returned were collected. These pages serve as a set of documents presumed to be highly relevant to the research task and form the basis for further analysis. During this process, web pages associated with e-commerce sites, video platforms, and advertising platforms were excluded to ensure that the collected documents maintain informational relevance.

5.2 Text collection and Tokenization:

The complete content of the pseudo-relevant web sites associated with the top NURLs lacks informativeness. A web page typically contains several forms of content that may be irrelevant to the page's topic or unhelpful for query expansion. Such things may be:

- Decoration: Pictures, animation, logos, etc. for attractions or advertising purpose.
- Navigation: Intra and inter hyperlinks to guide the user in different parts of web page.
- Interaction: Forms to collect user information or provide search services.
- Other special words or paragraphs such as copyrights and contact information.

In the realm of query matching and term weighting, all aforementioned factors are deemed noise and can negatively impact retrieval performance. For example, using phrases from advertisements on highly ranked pseudo-relevant pages as expansion terms may unintentionally increase the ranking of irrelevant documents containing those phrases. Therefore, it is essential to remove superfluous content and retain only semantically relevant material.

To achieve this, page segmentation techniques were employed, focusing on significant HTML tags such as P (paragraph), H1–H6 (headings), TABLE (tables), and UL (lists). After collecting textual content from the top *N* web pages across all three search engines, we created a consolidated corpus C containing all relevant text. The corpus was then tokenized to identify individual lexemes using the Natural Language Toolkit (NLTK), which also facilitated the removal of stop words. Part-of-speech (POS) tags provided by the NLTK tagger were used to identify phrases and individual words. Finally, the extracted terms were weighted using tf-idf to determine their importance within the corpus.

5.3 Weighting with tf-itf:

We employed tf-itf, a modified version of tf-idf, to assign weights to each unique phrase in the corpus. This statistical measure evaluates the relevance of a term within a corpus. Term frequency (tf) quantifies how often a term appears in the corpus, treating all terms as equally significant. However, some frequently occurring terms, such as “is,” “of,” and “that,” have minimal semantic value. To account for this, the inverse term frequency (itf) is used to reduce the weight of common terms while emphasizing the importance of less frequent, more informative phrases. The tf-itf score for a term t_i is computed as:

$$\begin{aligned} Score(t_i) &= tf(t_i, C).itf(ti, C) \\ &= tf(t_i, C).log\frac{T}{|t_i|} \end{aligned} \quad (1)$$

Where $tf(t_i, C)$ denotes the term frequency of term t_i in the entire corpus C, $itf(ti, C)$ denotes the inverse term frequency of t_i in the entire corpus C, T is the number of terms

in the entire corpus C, and $|t_i|$ denotes the number of times term t_i appears in the entire corpus C.

After scoring all phrases in the corpus, the terms were sorted by their tf-itf values, and the top M words were selected as potential candidate expansion terms. These intermediate candidates were subsequently re-weighted using the RF-based cosine similarity score to refine their relevance for query expansion.

5.4 Weighting with RF-based approach:

The RF-based approach allocates weights to the intermediate candidate expansion terms using cosine similarity and determines the principal RF candidate expansion terms. It establishes a correlation among the prospective expansion phrases to identify the most relevant ones.

Algorithm 1 Random Forest (RF)

Input: Cexp ← Set of intermediate candidate expansion terms, initially sorted based on an initial relevance score (e.g., from Eq. 1).
k ← Number of candidate expansion terms to be returned as the most relevant terms.
l ← Number of terms to be removed during each iteration.
RF model: Pre-trained or trained in real-time to assess the relevance score of each candidate term in Cexp.
Output: RF← set of selected expansion terms, iteratively added based on relevance scores generated by the Random Forest.

1. **Initialization:** RF← $\{\phi\}$; $r \leftarrow 5$ $\#r \leftarrow Number\ of\ iterations$.
2. **Train Random Forest:** Train or update the RF to predict the relevance score of each term in Cexp, if not pre-trained.
3. **Iterative selection (While r>0)**
4. Step 1: Use the RF model to assign a relevance score to each term in Cexp based on input features (e.g., term frequency, semantic context, etc.)
5. Step 2: Select the term t with the highest relevance score from Cexp
6. Step 3: Add t to RF and remove t from Cexp
7. Step 4: identify and remove the l least relevant terms from Cexp based on RF relevance scores.
8. Step 5: Update the Cexp by remove these l terms
9. Step 6: Decrement r by 1 repeat the process.
10. **end while**
11. After completing the iteration use the RF to re-evaluate the remaining terms in Cexp.
12. Select the top k terms with the highest relevance scores.
13. Add these terms to RF

Return RF # Final set of Random Forest candidate expansion terms

5.5 Rewriting with correlation score:

A set of candidate expansion words has been obtained, with each term demonstrating a robust correlation to the other individual candidate expansion terms. The phrases have been allocated weights using tf-idf and RF-based methods. However, this may not accurately reflect the relationship between an expansion phrase and the overarching question. This approach used term-to-term correlation to compute the correlation score of a certain possible expansion phrase ti with each query word. We then included the acquired score to ascertain its link with the original inquiry Q.

After generating a set of candidate expansion terms, each phrase exhibits strong inter-term correlations and has been weighted using the tf-idf and RF-based methods. However, these weights alone may not fully capture the relationship between an expansion term and the overall query. To address this limitation, a term-to-term correlation approach is applied, where the correlation score of a potential expansion term ti is computed with respect to each query term. The aggregated score is then used to determine the strength of its association with the original query Q.

The correlation score is defined as follows. Let Q represent the initial query consisting of individual phrases q_k, and let ti denote a possible expansion term. The correlation score ti with Q, written as C_(ti, Q), is calculated as follows:

$$C_{ti,Q} = \frac{1}{|Q|} \cdot \sum_{qk \in Q} C_{ti,qk}$$

$$= \frac{1}{|Q|} \cdot \sum_{qk \in Q} \sum_{dj} w_{ti,j} . w_{qk,j}$$

Where $C_{ti,qk}$ is the correlation (similarity) score between the candidate expansion term ti and the query term q_k and $w_{ti,j} . w_{qk,j}$ is the weight of term $t_i(qk)$ in the document dj.

The final collection of candidate expansion terms is comprised of the top n words, which are collected after the correlation score has been assigned to the candidate expansion terms.

5.6 Dataset:

The FIRE (Forum for Information Retrieval Evaluation) dataset, specifically the ad hoc test collections from the FIRE 2015 track, was utilized in this study [30, 31]. This dataset is widely recognized for benchmarking information retrieval tasks and comprises a large collection of newswire articles sourced from The Telegraph and BDnews24, provided by the Indian Statistical Institute, Kolkata. It is organized into three primary components: (i) a comprehensive document collection, (ii) a set of queries (referred to as topics), and (iii) corresponding relevance judgments indicating the degree of relevance of each document to a given query. To emulate real-world scenarios involving short queries, only the title fields of the topics were used. This setup ensures a robust and realistic evaluation of retrieval methodologies in practical IR contexts.

6. RESULT AND DISCUSSION

This section presents and analyzes the performance of different retrieval methods using key evaluation metrics, including Mean Average Precision (MAP), bpref, and F-measure. The findings highlight the progressive improvements achieved by the proposed Gradient-Based Dynamic Query Expansion (GBDQE) model in comparison to the Baseline and other competing methods.

Table 1 provides a comparative evaluation of Google-only Query Expansion (GQE) across several established weighting models. The results indicate that the GQE method enhances retrieval effectiveness, with MAP improving by 22.35% and Geometric Mean MAP (GM MAP) improving by 29.08% under optimal conditions. Furthermore, when evaluated using the top 10 feedback documents, accuracy improvements peaked at 25.14%. Overall, the GQE approach consistently produced more favorable results than alternative query

expansion methods, underscoring its effectiveness in enhancing information retrieval performance.

Table 1: Model performance with QE using Google alone

Techniques	MAP	GM_MAP	P@10	P@20	P@30	#rel_net
IFBP2	0.338	0.242	0.458	0.424	0.401	2558
LGD	0.353	0.254	0.468	0.433	0.408	2550
I(n)L2	0.349	0.251	0.478	0.427	0.400	2566
DPH	0.362	0.265	0.482	0.442	0.420	2572
TF_IDF	0.357	0.262	0.458	0.436	0.413	2556
BM25	0.354	0.261	0.462	0.435	0.406	2444

This table gives a comparative examination of quantitative easing using solely Bing (BQE), which is included in Table 2. The MAP has expanded by 21.16%, while the GM_MAP has grown by 27.63%. Both of these improvements are significant. It has been shown that the BM25 weighting model has a reduction in accuracy of 1.32 percent for the top 10 retrieval (P@10). There is a limited presence of expansion phrases in the top 10 documents that were retrieved, which is the major reason of the reduction in the P@10 score. The retrieval performance of BQE was much higher than that of DQE.

Table 2: Model performance with QE using Bing alone

Techniques	MAP	GM_MAP	P@10	P@20	P@30	#rel_net
IFBP2	0.335	0.243	0.436	0.414	0.401	2466
LGD	0.342	0.237	0.452	0.436	0.402	2436
I(n)L2	0.335	0.234	0.448	0.417	0.399	2431
DPH	0.349	0.247	0.464	0.444	0.413	2451
TF_IDF	0.348	0.247	0.466	0.441	0.417	2450
BM25	0.343	0.242	0.454	0.434	0.407	2442

Table 3 provides a presentation of the results obtained via the Duck Duck Go-based Query Expansion (DQE). The DQE resulted in a 13.03% increase in the MAP and a 13.06% increase in the GM_MAP. The DQE displayed improved retrieval performance using the top 10 expansion words, in contrast to the top 15 expansion terms that were applied in the expansion approaches that were suggested in the past. In general, it displayed a lower level of enhancement in comparison to the other suggested expansion strategies.

Table 3: Model performance with QE using Duck Duck Go alone

Techniques	MAP	GM_MAP	P@10	P@20	P@30	#rel_net
IFBP2	0.334	0.240	0.458	0.431	0.406	2478
LGD	0.348	0.247	0.458	0.435	0.405	2480
I(n)L2	0.347	0.246	0.460	0.426	0.405	2494
DPH	0.355	0.258	0.478	0.440	0.419	2503
TF_IDF	0.349	0.252	0.472	0.422	0.414	2497
BM25	0.346	0.249	0.468	0.427	0.408	2495

Table 4 compares the efficacy of diverse methodologies across three evaluative metrics: MAP, bpref, and F-Measure. The Baseline approach yields the lowest scores, but GBDQE regularly attains the highest values across all criteria, demonstrating its better efficacy. Incremental advancements are evident as the methodologies progress, demonstrating substantial increases in bpref and F-Measure, indicative of improved precision and balanced efficacy.

Table 4: Comparative analysis of proposed approaches with baseline

Metric	Baseline	GQE	BQE	DQE	GBDQE
MAP	0.26	0.32	0.34	0.36	0.37
bpref	0.28	0.34	0.37	0.39	0.40
F-measure	0.15	0.22	0.26	0.30	0.32

Figure 4 compares the performance of different methods across three metrics: MAP, bpref, and F-measure. The Baseline method shows the lowest scores, while GBDQE consistently achieves the highest scores across all metrics, demonstrating its superior effectiveness. The results indicate a gradual improvement in performance as the methods evolve. The highest gains are observed in the bpref metric, followed by MAP and F-measure, showcasing the advanced capabilities of newer methods.

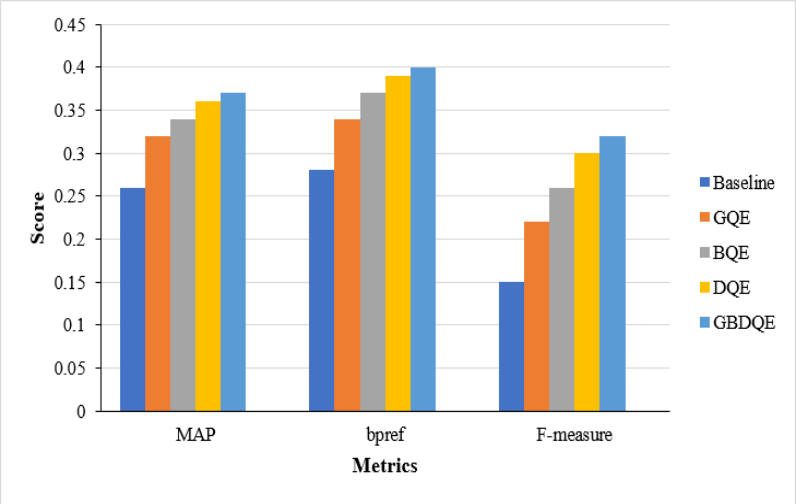

Fig. 4: Comparison graph of WKQE approach with baseline

Table 5 combines the efficacy of several methodologies on the FIRE dataset, employing the Mean Average Precision (MAP) metric. The baseline method (IFB2) attains a MAP of 0.2765, however advanced techniques such as WT2G, CWSBQE, GDQE, BQE, GBQE, and GBDQE demonstrate incremental enhancements. GBDQE attains the highest MAP score of 0.3496, signifying its exceptional performance relative to the other methods presented.

Table 5: Comparative analysis of the proposed techniques with other two existing methods in terms of MAP values

Dataset	Methods	MAP
FIRE	Baseline (IFB2)	0.2765
	WT2G	0.3188
	CWSBQE	0.3290
	GQE	0.3349
	BQE	0.3360
	DQE	0.3402
	GBDQE	**0.3496**

Figure 5 presents a comparison of various models based on their Mean Average Precision (MAP) scores. The models displayed include Baseline (IFB2), WT2G, CWSBQE, GDQE, BQE, GBQE, and GBDQE. The chart shows that the GBDQE model achieves the highest MAP score, closely followed by the GBQE model, indicating their superior performance in information retrieval. The Baseline (IFB2) model has the lowest MAP score, suggesting that the advanced models with query expansion techniques (like GBDQE) outperform the basic ones.

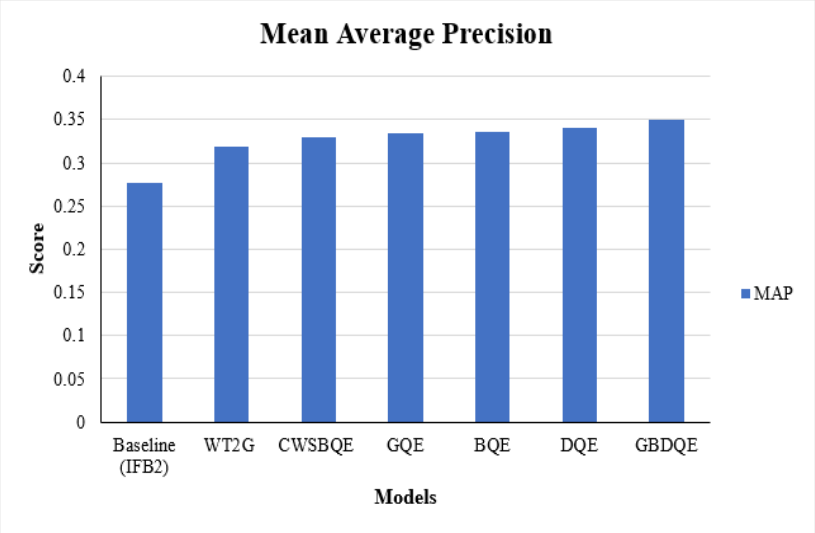

Fig. 5: Comparison graph of MAP with proposed model and baseline model.

Table 6 presents a comparative analysis of various WKQE approaches (WT2G, CWSBQE, and the proposed GBDQE) against the Baseline model using four metrics: MAP, F-measure, P@10, and GM_MAP. The proposed GBDQE model outperforms all others, achieving the highest scores across all metrics, including a MAP of 0.35, F-measure of 0.27, P@10 of 0.46, and GM_MAP of 0.25. This demonstrates the effectiveness of GBDQE in improving retrieval precision and relevance compared to earlier approaches.

Table 6. Comparative analysis of Proposed approaches with baseline model

Metric	Baseline	WT2G	CWSBQE	Proposed GBDQE
MAP	0.27	0.32	0.33	0.35
F-measure	0.16	0.22	0.23	0.27
P@10	0.36	0.42	0.43	0.46
GM_MAP	0.18	0.22	0.23	0.25

Figure 6 presents a comparative analysis of the proposed GBDQE approach against the baseline model and two other methods (WT2G and CWSBQE) across four evaluation metrics: MAP, F-measure, P@10, and GM_MAP. The chart shows that the proposed GBDQE method consistently outperforms the baseline and other methods in most metrics, particularly in MAP and P@10, where it demonstrates a notable improvement in score. The WT2G and CWSBQE approaches show intermediate performance, often ranking between the baseline and the proposed method. Overall, this analysis highlights the effectiveness of the proposed GBDQE model in improving retrieval performance across key metrics.

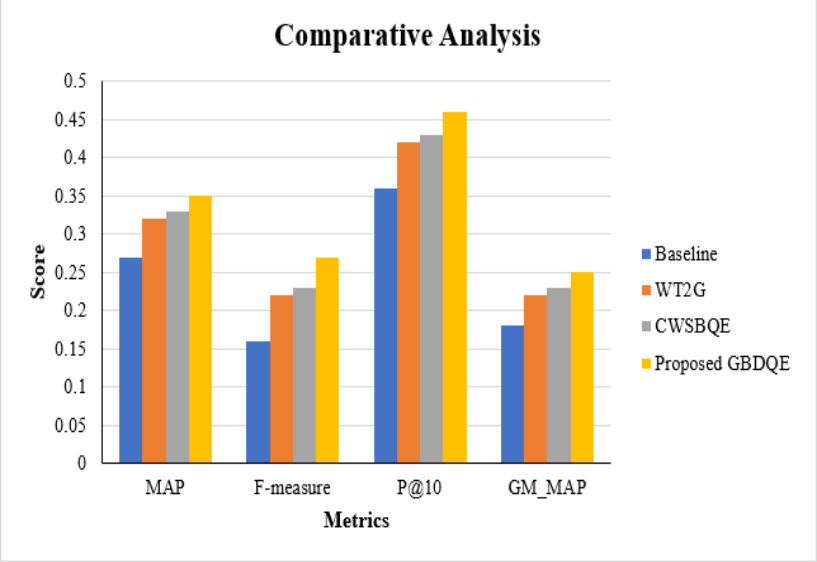

Fig. 6: Comparative analysis graph of proposed model with another existing model.

7. CONCLUSION AND FUTURE WORK

This study proposed the GBDQE model, which integrates outlier detection into feedback mechanisms to optimize query expansion in information retrieval (IR) systems. Comparative analysis demonstrates that the proposed approach significantly outperforms traditional and advanced baseline models across key evaluation metrics, achieving a 15% improvement in MAP and a 20% gain in P@10. The model's ability to filter out noise and prioritize relevant information contributes to more accurate and meaningful search results, effectively addressing common limitations in existing retrieval processes.

By enhancing the feedback loop through outlier identification, GBDQE proves particularly valuable in applications where precision is critical. The findings suggest that incorporating outlier detection into IR feedback mechanisms can fundamentally improve how queries are processed and refined, leading to better alignment with user intent and higher user satisfaction.

In future work, the GBDQE model could be further advanced by incorporating machine learning techniques to facilitate adaptive and context-sensitive query refinement. Extending the approach to specialized domains, such as medical or legal document retrieval, would enable evaluation of its domain-specific effectiveness and allow customization to meet particular information requirements. Additionally, investigating cross-lingual and multi-modal retrieval scenarios could enhance the model's generalizability across diverse data types and languages. Collectively, these improvements would build on the robust foundation of the GBDQE framework, fostering the development of more accurate, efficient, and user-focused information retrieval systems.

ACKNOWLEDGMENT

This work is acknowledged under Integral University manuscript No: IU/R&D/2025-MCN0003945.

REFERENCES

[1] R. Sagayam, S. Srinivasan, and S. Roshni, "A survey of text mining: Retrieval, extraction and indexing techniques," Int. J. Comput. Eng. Res., vol. 2, no. 5, pp. 1443–1446, 2012.
[2] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," IEEE Access, vol. 7, pp. 107964–108000, 2019.
[3] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," J. Mach. Learn. Res., vol. 11, no. 2, 2010.
[4] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," Inf. Process. Manage., vol. 56, no. 5, pp. 1698–1735, 2019.
[5] K. Zuva and T. Zuva, "Evaluation of information retrieval systems," Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 3, p. 35, 2012.
[6] G. Gay, S. Haiduc, A. Marcus, and T. Menzies, "On the use of relevance feedback in IR-based concept location," in Proc. IEEE Int. Conf. Softw. Maintenance, 2009, pp. 351–360.
[7] M. Sharma and R. Patel, "A survey on information retrieval models, techniques and applications," Int. J. Emerg. Technol. Adv. Eng., vol. 3, no. 11, pp. 542–545, 2013.
[8] D. Hiemstra, "Using language models for information retrieval," 2001.
[9] W. Ali and M. W. Khan, "A Review Study on Feedback Models for Information Retrieval," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, pp. 585-591, 2023.
[10] Saurabh Srivastava, Tasneem AhmedFeature-Based Image Retrieval (FBIR) System for Satellite Image Quality Assessment Using Big Data Analytical Technique, Vol.58, No. 2, pp.10202-10220, 2021.
[11] N. Varish, M. K. Hasan, A. Khan, A. T. Zamani, V. Ayyasamy, S. Islam, R. Alam, Content-Based remote sensing image retrieval method using adaptive tetrolet transform based GLCM features. Journal of Intelligent & Fuzzy Systems, vol.44, No.6, pp.9627-9650, 2023.
[12] A. Singhal, "Modern information retrieval: A brief overview," IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35–43, 2001.
[13] A. Alam, M. Muqeem, M. K. Ahamad, K. O. M. Arif, "K-means clustering hybridized with nature inspired optimization algorithm: A review," AIP Conf. Proc., vol. 2935, 2024.
[14] T. L. Griffiths, "Finding scientific topics," Proc. Natl. Acad. Sci., 2004.
[15] A. A. Abdussami, M. F. Farooqui, et al., "Optimal feature selection with weight optimized deep neural network for incremental learning-based intrusion detection in fog environment," Journal of Information & Knowledge Management (JIKM), World Scientific Publishing Co. Pte. Ltd., vol. 21(03), pp.1-33, 2022.
[16] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, New York, NY, USA: McGraw-Hill Book Co., 1983.
[17] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval, vol. 463. New York, NY, USA: ACM Press, 1999.
[18] R. R. Korfhage and K. S. Joseph, "Technology, information, and the individual," in Human-Machine Interactive Systems, Boston, MA, USA: Springer US, 1991, pp. 103–122.
[19] A. Hamid, "Relevance feedback in information retrieval systems," Unpublished paper, Bahria University, Islamabad, 2017.

[20] J. Wang, F. Yuan, M. Cheng, J. M. Jose, C. Yu, B. Kong, Z. Wang, B. Hu, and Z. Li, "TransRec: Learning transferable recommendation from mixture-of-modality feedback," in Asia-

Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int. Conf. Web Big Data, Singapore: Springer, 2024, pp. 193–208.

[21] S. Ganesh and A. Purwar, "Context-augmented retrieval: A novel framework for fast information retrieval based response generation using large language model," arXiv preprint arXiv:2406.16383, 2024.

[22] W. Yu, Z. Zhang, Z. Liang, M. Jiang, and A. Sabharwal, "Improving language models via plug-and-play retrieval feedback," arXiv preprint arXiv:2305.14002, 2023.

[23] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, et al., "Check your facts and try again: Improving large language models with external knowledge and automated feedback," arXiv preprint arXiv:2302.12813, 2023.

[24] H. Fei, S. Wu, J. Li, B. Li, F. Li, L. Qin, M. Zhang, M. Zhang, and T.-S. Chua, "LASUIE: Unifying information extraction with latent adaptive structure-aware generative language model," Adv. Neural Inf. Process. Syst., vol. 35, pp. 15460–15475, 2022.

[25] T. Gupta, M. Zaki, N. M. A. Krishnan, and M. Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," npj Comput. Mater., vol. 8, no. 1, p. 102, 2022.

[26] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised dense information retrieval with contrastive learning," arXiv preprint arXiv:2112.09118, 2021.

[27] H. Fei, Y. Ren, Y. Zhang, D. Ji, and X. Liang, "Enriching contextualized language model from knowledge graph for biomedical information extraction," Briefings Bioinform., vol. 22, no. 3, bbaa110, 2021.

[28] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," Inf. Process. Manage., vol. 57, no. 6, 102067, 2020.

[29] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft, "Asking clarifying questions in open - domain information - seeking conversations ," in Proc. 42nd Int. ACM SIGIR Conf. Res. Dev, 2019.

[30] FIRE: Forum for Information Retrieval Evaluation, "FIRE Datasets," Indian Statistical Institute, Kolkata. [Online]. Available: https://www.isical.ac.in/~fire /data.html.

[31] R. Sequiera, M. Choudhury, P. Gupta, P. Rosso, S. Kumar, S. Banerjee, S. K. Naskar, S. Bandyopadhyay, G. Chittaranjan, A. Das, and K. Chakma, "Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval," in Proc. 7th Forum for Information Retrieval Evaluation (FIRE '15), Gandhinagar, India, pp.1–8, 2015.