Journal of Computer Science

Research Article

An Approach to Query Reformulation in Cross Lingual Information Retrieval Emphasizing Term Placement

Amit Asthana and Sanjay K. Dwivedi

Department of Computer Sciences, Babasaheb Bhimrao Ambedkar University, Lucknow, India

Article history
Received: 08-03-2025
Revised: 06-08-2025
Accepted: 11-09-2025

Corresponding Author:
Amit Asthana
Department of Computer Sciences, Babasaheb Bhimrao Ambedkar University,
Lucknow, India
Email: aamitonline@gmail.com

Abstract: The effectiveness of retrieving relevant information is often hindered by ambiguous and short queries, compounded by often imprecise initial translation in Cross Lingual Information Retrieval (CLIR). These limitations are still a challenge for Query Reformulation (QR) techniques, which primarily focus on selecting effective expansion terms but generally neglect the impact of determining their optimal placement within the query. This paper introduces an approach to QR that not only focuses on identifying contextually relevant expansion terms but also effectively determines their optimal placement within the query to maximize retrieval performance. The method integra Continuous Bag of Words (CBOW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques to extract meaningful expansion terms from a snippet dataset. Co-occurrence-based term placement algorithm has also been proposed to find the optimal location for term placement. Further, we observed the improvements of 23.42 and 17.39% in the retrieval effectiveness when the expansion term added at optimal location and manually at the end, which underline the importance of both precise term selection and optimal term positioning in improving CLIR effectiveness.

Keywords: Web Query, CBOW, Query Reformulation

Introduction

Cross-Lingual Information Retrieval (CLIR) extends the capabilities of Information Retrieval (IR) to support retrieving documents written in different languages from the user's query language. CLIR involves retrieving information in a language different from the language of the user's query. This capability is essential in a globalized world where users seek information beyond language barriers, enabling access to a broader range of knowledge and resources.

The initial step of a CLIR system includes either the translation of user's query into the document's language or document translation into query language. The later approach can be complex and computationally expensive for large document collections. Therefore, query translation is generally preferred in CLIR using Machine Translation (MT) systems. However, MT may introduce translation errors that lead to ambiguity.

Moreover, users often input vague, ambiguous, and incomplete queries, making query reformulation important as well as challenging task for improving search results. Query expansion is a crucial technique in information retrieval aimed at improving IR system's effectiveness by adding expansion terms to the initial query (Azad and Deepak, 2019a). The expansion term should be appropriate enough that bridges the gap between user's intent and the actual intent of the query. Various methods have been proposed for implementing query expansion including utilizing external ontologies, entire document collection and top retrieved documents, which are then combined to create the final expanded query model (Carpineto and Romano, 2012). It is essential to carefully select expansion terms to avoid query drift, where poor expansion terms can result in irrelevant outcomes (Farhan et al., 2023).

Query expansion approaches involve selecting expansion terms from the documents retrieved by the original query to improve the retrieval effectiveness. Techniques like local context analysis focus on choosing expansion terms based on their co-occurrence with the query terms within the top-ranked documents ensuring the relevance of the expansion terms (Xu and Croft, 2000). Additionally, personalized query expansion integrates additional terms from individual user profiles to customize query results according to specific user preferences (Hameed, 2023).

Query expansion methods significantly enhance search engine results to improve search effectiveness (Yusuf et al., 2021). Query reformulation addresses the problem of term mismatch by expanding queries to get more relevant information through search engine (Nie

SCIENCE Publications

et al., 2016). Query expansion is a popular technique to improve the overall performance of search engines highlighting its significance in information retrieval (Jagerman et al., 2022).

Prominent methods to extract suitable expansion terms includes co-occurrence based method extracts terms that frequently appear next to the original query terms within a document collection. However, this method may sometimes overlook some contextual and semantic information as it primarily focuses on statistical connections. Mutual Information (MI), measures the strength of the relationship between query terms and terms within retrieved documents. Terms with higher MI scores are generally preferred for expansion. This approach provides a deeper understanding of term associations compared to the co-occurrence.

Word embeddings that utilize vector representations of words is an effective technique for query reformulation includes techniques such as Word2Vec, GloVe, and BERT that transform words into dense vectors that capture contextual similarities and semantic relationships. Although computationally intensive, word embeddings are favored for their advanced capability to handle complex linguistic phenomena, tasks including word-sense disambiguation, text classification, sentiment analysis, named entity recognition, question answering, and machine translation. These approaches when applied for query expansion processes, generally enhance the precision and relevance of the IR system. The basic idea behind many techniques, that employ word embeddings to improve user queries, is that words that appear together in the same context usually have similar meanings that is based on the distributional hypothesis. Two main categories of word embedding approaches are count-based and prediction-based. Prediction-based methods focus on making word predictions based on context, such as Continuous Bag of Words (CBOW) and Skip-Gram. Another one is the character n-gram trained such as GloVe and FastText that generate word vectors by using both the local context of a word and global co-occurrence of word to word.

Term Frequency-Inverse Document Frequency (TF-IDF) is another popular technique, which is a statistical method that evaluates the importance of a term within a document in a collection. TF-IDF combines two components (i.e. TF and IDF) to estimate term importance. TF measures the importance of a term based upon its frequency in a document whereas IDF examines the uniqueness of the term in the document collection.

Background Knowledge

This section outlines the foundational techniques used in the proposed approach, i.e. Word Embedding and TF-IDF, which are essential for understanding how expansion terms are selected and evaluated in the context of query reformulation.

Word Embedding

Word embedding is a popular technique used in the area of NLP which represents words as vectors in a multidimensional space. The semantic meaning and word association are utilized by categorizing related words together in the vector space. Word embedding actually refers to a group of techniques for creating vector representations of words that convey certain semantic information gathered from words in large text collections. These techniques include distributional hypothesis (Harris, 1954), which states that words that appear in similar positions typically have same meanings. Since word vector representations are determined by their context, words with similar meanings will have vector representations that are similar as well. A few examples of word embedding techniques are dimension reduction on the word co-occurrence matrix (Deerwester et al., 1990) probabilistic models (Blei et al., 2003) shallow neural network based techniques like word2vec, which is very effective on word analogy and similarity tasks and can also learn phrases’ vector representations (Mikolov et al., 2006), and recently, transformer, a type of neural network structure that pick the context through sequential data analysis and track relationships between sequence components using a set of modern mathematical techniques generally known as attention mechanism. These vector representations may be utilized to capture the semantic links between words by calculating the vectors' similarity.

Some popular word embedding algorithms are word2vec, GloVe, Fast Text and BERT. This work utilizes Word2Vec over other word embedding techniques due to its computational efficiency, lower resource requirements, and ability to capture local context.

Word2Vec provides efficient implementation of two types of neural network-based variations, i.e., Continuous Bag of Words (CBOW) and Skip Gram to calculate vector representations of words. There are several language processing activities that can be performed with these two representations. Skip-Gram architecture predicts context words around the current word whereas CBOW architecture predicts current words based on the surrounding context words as shown in Fig. 1.

Although algorithms like BERT are successful in capturing deep contextual understanding, they can be costly in terms of computational time and resources for the tasks like query expansion, where speed and efficiency are priority.

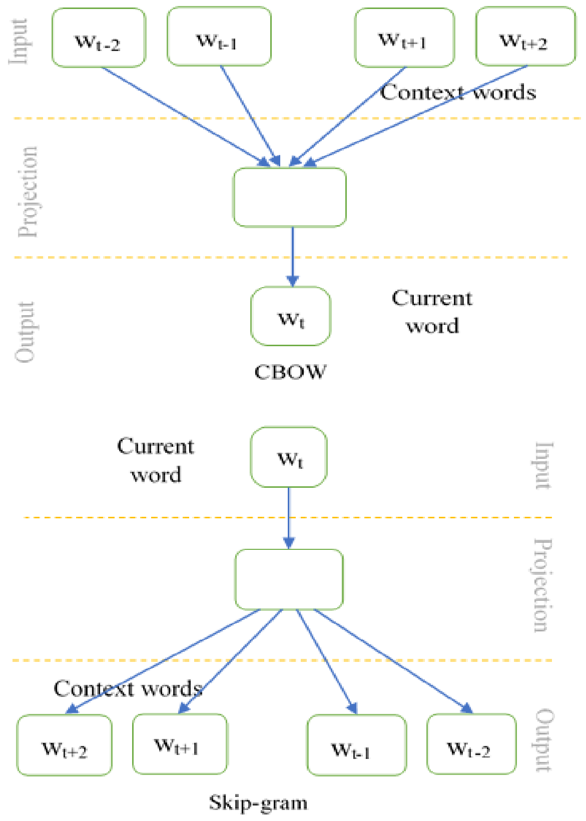

Fig. 1: CBOW and skip-gram architecture

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a simple and effective algorithm used in text mining to evaluate the importance of a word in a query within a document relative to a collection of documents. The relevance of a word within a document in relation to the collection as a whole is represented by the TF-IDF score, which includes both Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how often a word appears in a document whereas IDF represent the importance of a term within a document in the context of the entire collection and measures how uncommon a word is across the collection. It highlights words that are important in a specific document but not too common in the overall collection, making it useful for tasks like keyword extraction, document similarity and text classification (Azad and Deepak, 2019b; Wu et al., 2008). TF-IDF is calculated by multiplying the TF value of a term in a document by its IDF across all documents as:

$$\text{TF-IDF}(t,d,D)=\text{TF}(t,d)\times\text{IDF}(t,D)$$

Where:

$$\text{TF}(t,d)=\frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$\text{IDF}(t,D)=\log\left(\frac{\text{Total number of documents in the collection D}}{\text{Number of documents containing term t}}\right)$$

And *D* is the collection of documents *d* and *t* is the term present in the documents.

This paper utilizes word embedding for the extraction of expansion terms to create a term pool followed by TF-IDF utilized to select an appropriate expansion term for query reformulation. However, the effectiveness of this technique is influenced not just by the choice of expansion terms but also by their placement within the query. This paper also explores how the location of expansion term affects retrieval performance, providing insights that can inform the design of more effective IR systems.

Related Work

In CLIR system, Query reformulation increases the possibility of a lexical match with documents in the corpus by the reformulation of query terms that communicate the same information content. QR techniques have been the focus of many researches with early efforts focusing on relevance feedback mechanisms (Salton and Buckley, 1991). Techniques such as pseudo-relevance feedback and manual refinement have been used earlier to enhance search effectiveness. As the dependency on the search engine has rapidly increased, the need for automatic query reformulation also grown. As a result, recent advancements include the use of machine learning models to automatically suggest query modifications (Mitra and Craswell, 2018). Martinez et al. (2014) proposed an automatic query expansion method based on the Unified Medical Language System (UMLS), which enhanced the performance of a strong baseline in patient cohort searches by utilizing a graph representation of lexical units, concepts, and relationships within the UMLS meta-thesaurus with random walks across the graph initialized by the query terms. Their approach demonstrated improvements over the TREC Medical Record track in both the 2011 and 2012 datasets. Oh and Jung (2015) introduced a Cluster-Based Expansion Model (CBEEM) using external collections through a pseudo-relevance feedback approach. Extensive evaluation using three widely recognized biomedical collections, i.e. TREC CDS, CLEF eHealth, and OHSUMED for re-ranking purposes demonstrated that the proposed method outperformed representative expansion technique designed to utilize external collections. Diaz and Metzler (2006) found that external expansion, which involves using external corpora based on language modelling to extract suitable expansion terms, is more stable across different topics and improves mean average precision by up to 10%. Query expansion method based on Babel Net search and Word Embedding (Babel Net Embedding) aims to generate queries by exploring the semantic relationships within the original query to better understand query context (Maryamah et al., 2019). Candidate queries were created by identifying synonyms and measuring similarity using WordNet, Word Embedding across all Wikipedia articles, and Babel Net Embedding on Wikipedia Online articles. The proposed approach showed improved performance in comparison to existing semantic query expansion methods and achieved an accuracy of 89% in retrieving relevant Arabic documents. A framework utilizing domain-specific knowledge for ontology construction based on fuzzy ontology by leveraging the constructed fuzzy ontology proposed to identify the most semantically related terms to a query for query expansion (Jain et al., 2021). A fuzzy membership function is defined to capture various semantic relationships within the Global Ontology Concept Net. The framework evaluated across four popular search engines i.e. Google, Yahoo, Bing, and Exalead, resulting in 10% performance improvement.

Researches have shown that expansion terms, when carefully selected, can significantly improve retrieval performance (Carpineto and Romano, 2012). Sources for expansion terms include thesauri, ontologies, and user interaction data. However, most studies have concentrated on the selection of these terms rather than their placement within the query.

We could find very few studies those have addressed the impact of term placement within the query. Notably, Liu and Croft (2004) found that the positioning of terms can influence search results, but their work did not examine the relative effectiveness of queries by placing terms at different positions in de-tail. Chandra and Dwivedi (2024) proposed a location-based algorithm for the effective placement of expansion term in Hindi-English CLIR and found that the snippets produce the most effective improvements of 6.48% and 19.12% in terms of Mean Average Precision (MAP) over FIRE and nearest neighbor test collections respectively. This paper aims to fill this gap by systematically analyzing the impact of expansion term location on IR performance.

Proposed Approach

Proposed method consists two key components. The first component focuses on extracting an appropriate expansion term by utilizing the semantic enrichment of word embeddings while integrating frequency-based term extraction from TF-IDF to enhance query expansion. The second component addresses the optimal placement of this expansion term within the initial query. Together, these components aim to tackle the challenges of information retrieval by reformulating queries with relevant terms and positioning them effectively to boost retrieval performance. The steps involved in this approach are illustrated in Fig. 2.

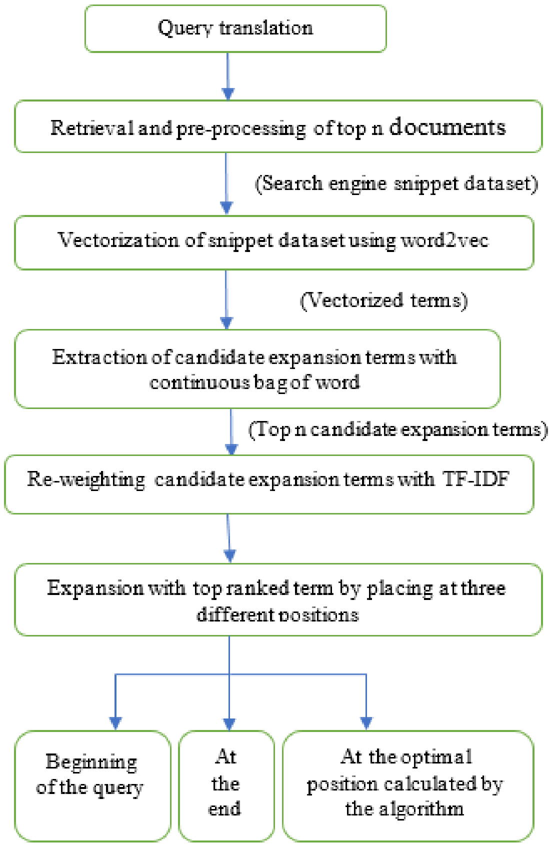
Query translation
Retrieval and pre-processing of top n documents
(Search engine snippet dataset)
Vectorization of snippet dataset using word2vec
(Vectorized terms)
Extraction of candidate expansion terms with continuous bag of word
(Top n candidate expansion terms)
Re-weighting candidate expansion terms with TF-IDF
Expansion with top ranked term by placing at three different positions
Beginning of the query
At the end
At the optimal position calculated by the algorithm

Fig. 2: Flow of proposed methodology

Significance and Novelty

- The proposed approach introduces a hybrid technique combining CBOW and TF-IDF, which effectively selects contextually relevant and statistically important terms for query expansion
- Unlike previous methods that only focus on term selection, this study also emphasizes the placement of expansion terms, a largely overlooked aspect in existing works
- The use of a co-occurrence based placement algorithm uniquely enhances retrieval effectiveness by integrating semantic proximity
- Compared to Liu and Croft (2004); Chandra and Dwivedi (2024), our approach shows greater consistency across diverse query types (FIRE 2012 dataset)
- Evaluation across 50 queries demonstrates 23.42% improvement in MAP, highlighting the practical impact of both term selection and its positioning

Figure 2 provides a schematic overview of the proposed approach. It outlines the key steps including query translation, snippet collection, embedding vectorization, extraction and selection of expansion terms, and final reformulation at three positions. Each component is detailed in the following subsections.

Query Selection and Translation

Query translation is initial and the most important step in CLIR as it has higher probability of introducing ambiguity in queries. A set of 50 Hindi queries from FIRE (Forum of Information Retrieval Evaluation) 2012 dataset are collected and translated into English using Google MT tool as we found it the most efficient for query translation among free online MT systems (Savoy and Dolamic, 2009).

Snippet Dataset

The first step in developing a QR system is to identify a suitable source from which the candidate expansion terms are obtained. In this work, translated query is provided as input to the search engine, specifically using Google. We collected top 20 (i.e. n = 20) search results (Vaughan, 2004) along with their corresponding snippets, utilizing a tool developed with python and Google API. After gathering the data, pre-processing takes place for the purpose to reduce noise and structure the data into a format that it further gets transformed into vector representation using word embeddings.

Expansion Term Extraction

Snippet dataset (Asthana and Dwivedi, 2023) has been embedded into vector using word2vec model. Then CBOW modal applied to predict the target terms based on the context words residing in the query and the set of candidate expansion terms are extracted by the estimation of probability distribution in a way to predict the target word from the query text in presence of surrounding context words through wort-to-word association of word2vec. Table 1 shows five sample queries along with the candidate expansion terms using CBOW approach.

Table 1: Queries with extracted candidate expansion terms using CBOW

Query	First cricketer to take 700 test wickets	Steve Irwin death	Guwahati 2008 bombing damage	Chamunda Temple stampede	Adarsh Housing Society scam resignation
Candidate expansion terms extracted by CBOW	biography	inevitable	lasso	post	upmarket
	bear	part	barpeta	hilltop	state
	never	ago	suspect	police	become
	wickets	killed	city	handful	bench
	indies	documentary	markets	year	drove
	comments	attack	sketch	near	symbol
	first	mandelaeffect	industry	disaster	truth
	completed	filming	ambulance	rises	thehindu
	single	2006	namely	officer	trouble
	elite	died	jumpto	2017	political
	finest	massive	database	important	accepted
	three	warrior	edu	200	spectrum
	fine	more	elections	television	resign
	jul	september	occurred	showed	corruption
	ece	great	woman	religious	difficult
	control	walk	iopscience	central	environment
	whole	one	compensation	vikram	fire
	again	internationally	main	died	series
	bowler	stalker	20edition	hindus	maharashtra
	different	tragically	economic	steep	narayan
	get	revenge	information	during	available
	premier	media	subs	pathway	apartments
	record	statements	gender	saturday	offer
	australia	massive	chandigarh	human	rhyme
	11akkdu	stephen	1088	reached	find
	thought	bindi	case	hardly	sunday
	thailand	voice	ssb	wall	act
	suffering	star	dispur	wikipedia	people
	demise	thurmond	ministry	climbing	governor
	name	says	damaged	precincts	western
	occasion	years	15	festivals	2015
	rights	63	connection	ago	commonwealth
	story	shooting	already	report	role
	older	tragic	guides	hill	housing
	near	yesterday	fall	panic	biggest
	across	glitch	doc	cause	chavan
	2007	day	results	special	government
	spin	wikipedia	rules	place	oct
	bowl	saying	rediff	september	unearthed
	photos	com	corporate	floor	probe

Selection of Suitable Expansion Term

After the candidate expansion terms pool is created, the next task is to fetch the suitable expansion term. The candidate terms are reordered based on the TF-IDF ranking. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It combines term frequency i.e. how often a word appears in a document and inverse document frequency i.e. how rare the word is across documents. The top ranked term is selected as the expansion term. Candidate expansion term having highest TF-IDF score for sample queries have been shown in Table 2.

Table 2: Queries with candidate expansion term having highest TF-IDF score

Query	Candidate Expansion Term
First cricketer to take 700 test wickets	record
Steve Irwin death	tragic
Guwahati 2008 bombing damage	connection
Chamunda Temple stampede	disaster
Adarsh Housing Society scam resignation	maharashtra

Placement of the Expansion Term at Optimal Location

In order to compare and analyze the impact on the retrieval effectiveness the effects of placing the expansion terms at different positions in the query, we placed the expansion term at three distinct positions, i.e.,

Case 1: At the beginning
Case 2: At the end
Case 3: At the optimal location calculated by the proposed co-occurrence based term location algorithm

The co-occurrence based term location algorithm determines the most suitable query position for the expansion term based on ordered co-occurrence frequencies. It analyzes how frequently the expansion term appears before and after each query word, then inserts it where the semantic association is highest.

The co-occurance based term location algorithm to place the expansion term for case 3 is given below:

Co-Occurance Based Term Location Algorithm

```
query = {input query}
expansion_term = {extracted expansion term}
n = number of query terms

int co_occurrence_before[n]
int co_occurrence_after[n]

for i = 0 to n - 1 do
    //number of ordered co-occurrence when expansion term followed by query term
    co_occurrence_before[i] = ordered_cooccurrence(expansion_term, query[i])
    //number of co-occurrence when query term followed by expansion term
    co_occurrence_after[i] = ordered_cooccurrence(query[i], expansion_term)
end for

// Find max and positions
highest_before = co_occurrence_before[0]
highest_before_pos = 0
highest_after = co_occurrence_after[0]
highest_after_pos = 0

//highest co-occurrence value and its index
for i = 1 to n - 1 do
    if co_occurrence_before[i] > highest_before then
        highest_before = co_occurrence_before[i]
        highest_before_pos = i
    end if
    if co_occurrence_after[i] > highest_after then
        highest_after = co_occurrence_after[i]
        highest_after_pos = i
    end if
end for

if highest_after > highest_before then
    Place expansion_term after query[highest_after_pos]
else
    Place expansion_term before query[highest_before_pos]
end if
```

In the above algorithm, initially two array variables co-occurrence_before and co-occurrence_after are used to store the co-occurrence count for each term with the expansion term placed before each query term and after each query term respectively. Ordered co-occurrence of expansion term followed by each query term is stored in co-occurrence_before and ordered co-occurrence for each query term followed by expansion term has been stored in co-occurrence_after. Then highest value of co-occurrence_before and co-occurrence_after with the position of term in the query is stored in highest_before_pos and highest_after_pos respectively. Then if the co-occurrence value in highest_after is greater than highest_before, the expansion term is placed after the query term having the highest co-occurrence value otherwise, the expansion term is placed before the query term having the highest co-occurrence value. Table 3 shows the initial query and the reformulated query by placing the expansion term at the beginning, at the end and at the optimal position according to the proposed location-based algorithm, i.e. for case 1, case 2 and case 3 respectively.

Retrieval and Evaluation

After placing the expansion term in the query, retrieval results have been collected for all three positions based upon the location of the expansion terms.

Table 3: Initial query and expanded query by placing expansion term at case 1, case 2, and case 3 position

Initial Query	Expended query for case 1	Expended query for case 2	Expended query for case 2
First cricketer to take 700 test wickets	record first cricketer to take 700 test wickets	first cricketer to take 700 test wickets record	first cricketer to take 700 test wickets record
Steve Irwin death	tragic steve irwin death	steve irwin death tragic	steve irwin tragic death
Guwahati 2008 bombing damage	connection guwahati 2008 bombing damage	guwahati 2008 bombing damage connection	guwahati 2008 bombing damage connection
Chamunda Temple stampede	disaster chamunda temple stampede	chamunda temple stampede disaster	chamunda temple stampede disaster
Adarsh Housing Society scam resignation	maharahtra adarsh housing society scam resignation	adarsh housing society scam resignation maharahtra	adarsh housing society maharahtra scam resignation

Performance of query reformulation is assessed using the precision and MAP, i.e. the proportion of relevant documents retrieved and a measure of precision across all relevant documents respectively.

Results and Discussion

In this section, the performance evaluation result of the proposed methodology is discussed. The proposed methodology is applied on a set of 50 queries collected from FIRE 2012 dataset. FIRE dataset has been used as it provides a description along with detailed narration stating what exactly the intent of the query and what output the query should provide. This work provides a way to improve the retrieval effectiveness by reformulating web query with combined approach of word embedding and TF-IDF that takes place after the query translation. Semantic connections between words in the vector space is utilized through the use of Word2Vec that allows it to predict contextually relevant terms from snippet dataset to build a candidate expansion term pool. TF-IDF prevents the selection of frequent but less significant words by preserving the uniqueness and accuracy of terms, and thus selects the statistically suitable expansion term from the term pool.

The experimental results reveal a significant variation in retrieval performance based on the position of expansion terms. The co-occurrence based location algorithm suggests a variety of appropriate positions for expansion term placement. We observed that the location-based algorithm suggested adding the expansion term at the end in 52% of the queries, at the beginning in 6% and somewhere in the middle positions for 42% of the queries. The range of positions in the query for placing the expansion term demonstrates the effectiveness of algorithm in determining the optimal placement of query terms. Placing expansion terms at the beginning of the query resulted in lower precision. This suggests that the search system retrieved a broader set of documents, including more irrelevant ones. Placing expansion terms at the end typically increased precision, indicating a narrower focus on documents closely matching the original query terms. When the expansion term is placed at the end of the query, the primary intent of the original query is preserved. Search engines typically prioritize the terms at the beginning of a query as the most important for matching relevant documents. In order to investigate the impact of expansion term placement, we have calculated and analyzed the retrieval effectiveness by placing expansion term for all the three cases as shown in Table 4.

Table 4: Precision and average precision for initial query and expanded query for case 1, case 2, and case 3 category

Query#	Precision(P@10)				Average Precision (AP@10)			
	Initial Query	Case 1	Case 2	Case 3	Initial Query	Case 1	Case 2	Case 3
1	0.4	0.4	0.5	0.5	0.310	0.260	0.349	0.438
2	0.5	0.4	0.6	0.6	0.369	0.293	0.478	0.478
3	0.1	0.4	0.2	0.5	0.033	0.260	0.054	0.215
4	0.6	0.5	0.6	0.7	0.501	0.364	0.505	0.620
5	0.3	0.6	0.6	0.6	0.188	0.339	0.339	0.339
6	0.8	0.8	0.9	0.9	0.648	0.725	0.879	0.879
7	0.7	0.4	0.5	0.6	0.579	0.344	0.304	0.575
8	0.8	0.7	0.9	0.9	0.700	0.608	0.790	0.900
9	0.6	0.6	0.8	0.8	0.538	0.501	0.789	0.789
10	0.7	0.9	0.9	0.8	0.624	0.790	0.852	0.659
11	0.4	0.3	0.3	0.3	0.204	0.117	0.154	0.154
12	0.7	0.6	0.8	0.8	0.550	0.414	0.725	0.725
13	0.5	0.5	0.5	0.5	0.332	0.315	0.360	0.360
14	0.6	0.6	0.7	0.7	0.440	0.546	0.616	0.616
15	0.8	0.7	1.0	0.9	0.800	0.673	1.000	0.900
16	0.5	0.4	0.6	0.6	0.324	0.210	0.401	0.401

Table 4: Continued

17	0.2	0.1	0.2	0.2	0.117	0.011	0.167	0.200
18	0.7	0.7	0.8	0.8	0.608	0.602	0.78	0.780
19	0.6	0.5	0.6	0.6	0.371	0.28	0.414	0.414
20	0.9	0.8	1.0	1.0	0.879	0.733	1.000	1.000
21	0.5	0.6	0.5	0.5	0.287	0.366	0.327	0.327
22	0.9	0.9	0.9	1.0	0.900	0.835	0.900	1.000
23	0.6	0.6	0.7	0.7	0.505	0.546	0.673	0.673
24	0.9	0.7	0.9	0.9	0.900	0.700	0.890	0.900
25	0.9	0.8	1.0	1.0	0.866	0.768	1.000	1.000
26	0.7	0.7	0.7	0.7	0.483	0.500	0.500	0.500
27	0.5	0.5	0.3	0.6	0.256	0.250	0.143	0.323
28	0.2	0.6	0.6	0.6	0.167	0.478	0.326	0.586
29	0.8	0.7	0.8	0.8	0.626	0.588	0.646	0.646
30	0.4	0.4	0.5	0.5	0.230	0.260	0.310	0.310
31	0.7	0.6	0.6	0.8	0.636	0.396	0.461	0.727
32	0.5	0.6	0.8	0.7	0.395	0.489	0.671	0.616
33	0.7	0.9	0.9	0.9	0.700	0.900	0.900	0.900
34	0.4	0.5	0.6	0.6	0.267	0.332	0.460	0.460
35	0.7	0.6	0.7	0.8	0.620	0.478	0.608	0.727
36	0.5	0.6	0.6	0.6	0.382	0.423	0.442	0.456
37	0.6	0.8	0.8	0.8	0.473	0.684	0.762	0.762
38	0.6	0.1	0.1	0.1	0.473	0.100	0.100	0.013
39	0.3	0.7	0.8	0.8	0.233	0.541	0.692	0.692
40	0.7	0.6	0.7	0.7	0.583	0.560	0.624	0.700
41	0.4	0.5	0.5	0.5	0.284	0.304	0.343	0.343
42	0.6	0.7	0.8	0.7	0.406	0.656	0.743	0.656
43	0.5	0.6	0.6	0.6	0.405	0.575	0.575	0.549
44	0.7	0.8	0.9	0.9	0.678	0.682	0.879	0.879
45	0.8	0.9	0.9	0.9	0.800	0.900	0.900	0.900
46	0.8	0.8	0.8	0.9	0.800	0.800	0.800	0.890
47	0.9	0.9	0.9	1.0	0.879	0.890	0.900	1.000
48	0.3	0.3	0.3	0.3	0.188	0.188	0.210	0.300
49	0.9	0.9	0.9	0.9	0.890	0.879	0.879	0.879
50	0.8	0.9	0.9	0.9	0.646	0.815	0.815	0.790
Total	30.2	30.7	34	35	25.073	25.268	29.435	30.946
	Mean Average Precision (MAP)				0.50146	0.50536	0.5887	0.61892

By placing the original query terms at the end, search engine focuses on retrieving documents that are highly relevant to the user's initial intent, and the expansion term acts as a secondary filter to refine the results. Placing the expansion terms at the position having the highest co-occurrence value with the query term, resulted in highest precision, indicating an optimal integration of additional context without changing the intent of original query.

Conclusion

This work highlights the importance of strategic placement of expansion terms in query reformulation and intends to improve the retrieval effectiveness by applying QR through hybrid approach of word embedding and TF-IDF. This work focuses not only the extraction of a suitable expansion term but also the impact of expansion term placement. The findings from our study demonstrate an increase in retrieval effectiveness through the combined utilization of CBOW followed by TF-IDF for query expansion. The proposed approach of QR by appending expansion term at an optimal location in the query shows a significant improvement of 23.42% in the retrieval effectiveness in terms of MAP i.e. higher when the expansion term added at the beginning and at the end, i.e. 0.78 and 17.39% respectively. CBOW extracts the expansion terms that are contextually relevant to the initial query whereas TF-IDF refines these candidate expansion terms by focusing on the statistical aspect and the combined approach results in improved retrieval effectiveness.

It has also been observed that the location of the expansion term plays a significant role in retrieval effectiveness. Placing expansion term at an inappropriate location generally degrade the retrieval effectiveness. Further research and refinement of this approach hold substantial promise for advancing the query reformulation through the utilization of hybrid approach that uses two or more techniques with the inclusion of location aspect.

Future Work

Future studies should explore adaptive placement algorithms that dynamically analyze user intent and context to determine optimal term placement based on query characteristics. Integration with transformer-based embeddings may yield further improvements.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to reports.

Authors Contributions

The authors contributed equally to this study.

Ethics

This article is original and has not been published elsewhere. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Asthana, A., & Dwivedi, S. K. (2023). Exploring Snippets as a Dataset to Overcome Challenges in CLIR. *ITM Web of Conferences*, *54*, 01012. https://doi.org/10.1051/itmconf/20235401012

Azad, H. K., & Deepak, A. (2019a). A new approach for query expansion using Wikipedia and WordNet. *Information Sciences*, *492*, 147–163. https://doi.org/10.1016/j.ins.2019.04.019

Azad, H. K., & Deepak, A. (2019b). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, *56*(5), 1698–1735. https://doi.org/10.1016/j.ipm.2019.05.009

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082

Chandra, G., & Dwivedi, S. K. (2024). Query Expansion Using Proposed Location-Based Algorithm for Hindi–English CLIR: Analyzing Three Test Collections. *International Journal of Pattern Recognition and Artificial Intelligence*, *38*(05). https://doi.org/10.1142/s0218001424590018

Carpineto, C., & Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, *44*(1), 1–50. https://doi.org/10.1145/2071389.2071390

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Diaz, F., & Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 154–161. https://doi.org/10.1145/1148170.1148200

Farhan, Y. H., Mohd Noah, S. A., Mohd, M., & Atwan, J. (2023). Word-embedding-based query expansion: Incorporating Deep Averaging Networks in Arabic document retrieval. *Journal of Information Science*, *49*(5), 1168–1186. https://doi.org/10.1177/01655515211040659

Hameed, A. (2023). Personalized Query Expansion. *International Journal of Information Systems and Computer Technologies*, *2*(1). https://doi.org/10.58325/ijisct.002.01.0043

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Jagerman, R., Qin, Z., Wang, X., Bendersky, M., & Najork, M. (2022). On Optimizing Top-K Metrics for Neural Ranking Models. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2303–2307. https://doi.org/10.1145/3477495.3531849

Jain, S., Seeja, K. R., & Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. *International Journal of Information Management Data Insights*, *1*(1), 100009. https://doi.org/10.1016/j.jjimei.2021.100009

Liu, X., & Croft, W. B. (2004). Cluster-based retrieval using language models. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 186–193. https://doi.org/10.1145/1008992.1009026

Martinez, D., Otegi, A., Soroa, A., & Agirre, E. (2014). Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *Journal of Biomedical Informatics*, *51*, 100–106. https://doi.org/10.1016/j.jbi.2014.04.013

Maryamah, M., Arifin, A., Sarno, R., & Morimoto, Y. (2019). Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents. *International Journal of Intelligent Engineering and Systems*, *12*(5), 202–213. https://doi.org/10.22266/ijies2019.1031.20

Mikolov, T., Sutskeve, I., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Com-positionality. *Neural Information Processing Systems*. Curran Associates, Inc., Red Hook, New York, USA. https://doi.org/10.48550/arXiv.1310.4546

Mitra, B., & Craswell, N. (2018). An Introduction to Neural Information Retrieval in *Foundations and Trends® in Information Retrieval*, *13*(1), 1–126. https://doi.org/10.1561/1500000061

Nie, L., Jiang, H., Ren, Z., Sun, Z., & Li, X. (2016). Query Expansion Based on Crowd Knowledge for Code Search. *IEEE Transactions on Services Computing*, *9*(5), 771–783. https://doi.org/10.1109/tsc.2016.2560165

Oh, H.-S., & Jung, Y. (2015). Cluster-based query expansion using external collections in medical information retrieval. *Journal of Biomedical Informatics*, *58*, 70–79. https://doi.org/10.1016/j.jbi.2015.09.017

Salton, G., & Buckley, C. (1991). Global Text Matching for Information Retrieval. *Science*, *253*(5023), 1012–1015. https://doi.org/10.1126/science.253.5023.1012

Savoy, J., & Dolamic, L. (2009). How effective is Google's translation service in search? *Communications of the ACM*, *52*(10), 139–143. https://doi.org/10.1145/1562764.1562799

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, *40*(4), 677–691. https://doi.org/10.1016/s0306-4573(03)00043-8

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, *26*(3), 1–37. https://doi.org/10.1145/1361684.1361686

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, *18*(1), 79–112. https://doi.org/10.1145/333135.333138

Yusuf, N., Mohd Yunus, M. A., Wahid, N., Mustapha, A., & Mohd Salleh, M. N. (2021). A Survey of Query Expansion Methods to Improve Relevant Search Engine Results. *International Journal on Advanced Science, Engineering and Information Technology*, *11*(4), 1352. https://doi.org/10.18517/ijaseit.11.4.8868