scientific reports

OPEN

A multi-dimensional semantic pseudo-relevance feedback framework for information retrieval

Min Pan[1], Yu Liu[1✉], Jinguang Chen[2✉], Ellen Anne Huang[3] & Jimmy X. Huang[4✉]

Pre-trained models have garnered significant attention in the field of information retrieval, particularly for improving document ranking. Typically, an initial retrieval step using sparse methods such as BM25 is employed to obtain a set of pseudo-relevant documents, followed by re-ranking with a pre-trained model. However, the semantic information captured by pre-trained models from sentences or passages is usually only applied to document ranking, with limited use in query expansion. In fact, the semantic information within pseudo-relevant documents plays a critical role in selecting appropriate query expansion terms. Therefore, this paper proposes a novel approach that leverages pre-trained models to extract multi-dimensional semantic information from pseudo-relevant documents, offering more possibilities for query expansion. First, traditional sparse retrieval methods are used in the initial retrieval stage to ensure efficiency, and term-level weights are calculated based on statistical information. Then, the pre-trained model encodes both the query and the sentences and passages from the documents, extracting sentence-level and passage-level semantic similarities to the query. Finally, these semantic weights are combined with the term-level weights to generate an improved query for the second retrieval round. We conducted experiments on five TREC datasets and a medical dataset, showing improvements in official metrics such as MAP and P@10. The results demonstrate the effectiveness of utilizing multi-dimensional semantic information from pseudo-relevant documents to optimize query expansion. This study offers new insights into how the semantic information of pseudo-relevant documents can be effectively harnessed to enhance retrieval performance.

Information retrieval (IR) technology development is closely related to the human need for accessing information. As an integral component of intelligent information processing technology, IR and its product forms have evolved rapidly in recent years. The core of IR technology lies in finding documents relevant to the user's query from a large and unorganized dataset, typically returning a ranked list of documents based on relevance and user needs. The development of IR technology has been closely tied to the advancement of Internet technologies, and many product forms have been derived. Retrieving information from vast datasets and efficiently delivering results that accurately meet user needs have been the central focus and challenge of IR research.

Is the essence of the text retrieval task relevance matching or semantic matching? In IR, relevance matching is more concerned with exact term matching, and semantic matching is concerned with the semantic similarity of two texts as a whole. Early research in IR was predominantly based on keyword relevance matching. In existing studies, many retrieval models have been proposed, including vector space models[1], probabilistic models[2], and language models[3]. These models have been widely used with retrieval systems. The major idea of the vector space model is to simplify the processing of text content into vector calculations in vector space and to represent the similarity of text in terms of spatial similarity. The probabilistic model determines the relevance of a query to a document based on the probability of occurrence of a term in the document in questions. Probabilistic models describe the statistical information of the data in terms of probability distributions such as the BM25 model[4,5]. The essence of the language model is a probability-based retrieval model, Relevance Model 1 (RM1) and Relevance Model 2 (RM2)[3] are two relevance models designed according to the classical probability model, using only queries to estimate the probability of words in the relevant class and differentiated according to the

[1]College of Computer and Information Engineering, Hubei Normal University, Huangshi, China. [2]School of Electronic Information, Huzhou College, Huzhou, China. [3]Department of Computer Science, Western University, London, Canada. [4]Information Retrieval and Knowledge Management Research Lab, School of Information Technology, York University, Toronto, Canada. ✉email: liuyuuuu1110@gmail.com; jinguangchen100@163.com; jhuang@yorku.ca

1

way they are sampled. Relevance Model 3 (RM3) and Relevance Model 4 (RM4) are obtained by interpolating the original query language model with the RM1 and RM2 models.

With the development of artificial intelligence, frontier algorithms such as machine learning and deep learning have driven the evolution of IR systems. In the existing studies, modern search engines use sophisticated machine learning to search the documents with the highest relevance out of the query. Neural networks and deep learning are increasingly used in IR, particularly for document ranking[6–12]. The retrieval results achieved by combining deep contextual semantic representations pretrained on a large corpus[13–17] have shown significant progress compared to traditional information retrieval methods[18]. The bag-of-words model inherently overlooks the sequential order of words in a sentence, which can lead to completely different meanings depending on word arrangement. For instance, the sentences "The dog bit the man." and "The man bit the dog." contain the same words but convey entirely different messages due to the change in word order. Research has demonstrated that pre-trained models can better understand query intent and return more accurate query results. Semantic similarity has an essential function in natural language processing (NLP) tasks, and evaluating semantic relations of documents based on both sentence semantic relations[11,12,19] and paragraph semantic relations has been explored by many researchers. However, the meaning of a single sentence can shift depending on the surrounding context. For example, the phrase "I can't wait" may express excitement when anticipating an upcoming holiday, whereas in a context of time constraints or urgency, it can reflect impatience. Addressing the issue of semantic information, this paper focuses on sentence and paragraph semantics, incorporating semantic information of terms into the traditional pseudo-relevance feedback model. This is achieved by processing word elements with semantic information at both the sentence and passage levels, with the expectation of improving feedback document quality and, consequently, the performance of the retrieval model.

Pseudo-relevance feedback (PRF) is an essential technical branch of IR that enhances the efficiency of the retrieval system, producing search results that better satisfy the user's query requests[20]. In addition, it provides a method for automatic local analysis, which partially automates the manual operation of relevance feedback and improves the performance of IR without requiring additional user interaction. The method first performs the normal retrieval process by returning the most relevant documents to form the initial set and the top N documents are assumed to be relevant, then uses query expansion techniques to generate candidate expansion terms from these N documents, and finally performs relevance feedback to obtain retrieval results as before. Due to its convenience and efficiency, the PRF has rapidly become a technical research hotspot in IR once it was proposed. Query expansion (QE) technology is broadly used in IR because it can better solve the problems of mismatch between user query terms and document terms and incomplete user expressions by expanding and reconstructing the user's initial query[21]. In brief, query expansion refers to the retrieval system automatically expanding the synonyms or near-synonyms of the keywords in the user's original query to form a new query, utilizing an expanded list of terms prior to conducting the search.

Semantic information plays a crucial role in various NLP tasks[22]. However, most of the retrieval tasks take semantic information to help re-ranking and rarely use semantic information to improve query expansion. Based on the classical PRF approach, we integrate different levels of semantic information into the probabilistic model. This approach chooses expansion terms in the retrieved document by the BM25. Then we get the frequency and significance of each term at the term level. The sentences and passages in the feedback document are then encoded with the query by BERT to obtain the semantic similarity of the query to the sentences and passages individually. Finally, we combine term-level, sentence-level, and paragraph-level semantic weights to construct a new query for the second-round retrieval.

The main objective of this research paper is to incorporate semantic information into traditional text matching. The details can be described as follows:

1. Integrating sparse methods with pre-trained models for retrieval optimization: this study aims to explore the effective combination of sparse retrieval methods with pre-trained models to balance retrieval efficiency and precision. Traditional sparse methods specialize in term matching and computational efficiency, while pre-trained models can capture rich semantic information, making them suitable for advanced semantic analysis and document re-ranking. One of the primary goals of this research is to integrate these two techniques to provide optimized pathways for subsequent query expansion and document ranking.

2. Deeply mining multi-dimensional semantic information from pseudo-relevant documents: conventional query expansion methods fail to fully leverage the potential multi-dimensional semantic information contained in pseudo-relevant documents. By introducing pre-trained models to semantically encode these documents, this study aims to extract multi-dimensional semantic information at the sentence and passage-level, providing richer semantic cues for query expansion. The goal is to maximize the semantic expressive power of pre-trained models, overcoming the limitations of traditional methods.

3. Optimizing query expansion with sentence-level and passage-level semantic similarity: this study further utilizes pre-trained models to extract sentence-level and passage-level semantic similarities from pseudo-relevant documents. These similarity measures, combined with term-level weights, are applied to query expansion. Sentence-level and passage-level semantic similarities can more precisely capture the structure and content hierarchy of documents, offering a more comprehensive reference for generating expansion terms. By integrating different levels of semantic information, the research aims to produce more accurate queries, thereby enhancing retrieval performance.

4. Comprehensive application of multi-dimensional semantic information and term-level statistical weights: in the processes of query expansion and document re-ranking, relying solely on term statistical information may not accurately reflect the semantic characteristics of documents, while using only semantic information might overlook the importance of terms. Therefore, this study aims to integrate sentence and passage semantic similarities with term-level weights, balancing their roles in query expansion to optimize retrieval outcomes. The objective is to propose a more balanced approach to improve the effectiveness of query expansion.

The main contributions of this research are summarized as follows:

1. Proposed a query expansion framework integrating multidimensional semantic information: This paper presents a framework that incorporates multidimensional semantic information from pseudo-relevant documents into query expansion. This approach not only relies on term frequency statistics but also integrates sentence-level and passage-level semantic similarities to improve the selection of expansion terms. It allows the identification of terms that are semantically closer to the query, particularly in cases where multiple candidate expansion terms share similar frequencies.

2. Optimization of expansion term selection by combining term importance with semantic importance: This study is the first to systematically integrate term-level importance with multidimensional semantic information from pseudo-relevant documents. The proposed model captures a more comprehensive semantic structure of pseudo-relevant documents, allowing for the selection of expansion terms with stronger relevance to the query. This approach overcomes the limitations of traditional methods that rely solely on term matching and demonstrates significant performance improvements across several TREC datasets.

3. Extended the semantic dimension of the Rocchio feedback model: This paper extends the Rocchio feedback model by incorporating both contextual semantics of single sentences and semantics across consecutive sentences, surpassing traditional sentence-level semantic analysis. By capturing the overall semantic representation of documents, the model enhances the selection of expansion terms within pseudo-relevant feedback, thereby significantly improving query expansion performance. The results confirm the effectiveness of integrating multidimensional semantic information in optimizing retrieval performance.

This paper is organized as follows. Section "Related work" will provide a review of related work. Section "Our proposed framework" will describe our proposed approach to combine term-level importance with sentence-level semantics and passage-level semantics for selecting query expansion terms. Section "Experimental setup" will present the experimental setup. Following that, Section "Results and analyses" will describe our experimental findings and conducts an analysis. In conclusion, Section "Summary and future work" will summarize our work and delve into the prospects for future research. Our code can be found at the following location: https://github.com/panminiii/MSROC.

Related work

Bidirectional Encoder Representations from Transformers (BERT)[14] uses a bi-directional transformers-based coding framework that captures contextual information better than previous pre-trained models. BERT incorporates bi-directional contextual information in modeling the representation of text and in this way, the trained model can capture the semantics of the context and attain more correct text prediction generation in handling question and answer or linguistic inference tasks. The combined contextual semantic information allows it to successfully handle NLP tasks. Since the widespread adoption of BERT in NLP tasks, researchers in IR have also begun applying BERT in the field of retrieval[23]. Due to BERT's limitation on input sequence length, researchers have shifted their focus to relevance at the sentence and passage levels.

Sentence level

Yang et al.[24] applied BERT in the field of IR for the first time, proposing that if a query is related to certain sentences within a document, it should also be considered related to the entire document. The validity of the local relevance principle was demonstrated by using high-scoring sentences instead of whole documents, thereby showcasing the feasibility of applying BERT in document reordering tasks. Yilmaz et al.[25] proposed that the task of a document relevance estimation is converted into a relevancy estimation task for single sentences, and then the obtained scores are aggregated. The relevance scores of documents in this scheme were derived from a union of the original candidate document score (e.g., from BM25) and the highest scoring sentence's contribution from the document as determined by the BERT model. The experimental results demonstrate the feasibility and validity of this idea. Reimers et al.[19] introduced Sentence-BERT, which uses a Siamese Network to output a fixed-length sentence embeddings with semantic information. The similarity is calculated by cosine distance, Manhattan distance, or Euclidean distance. Gao et al.[26] proposed SimCSE, a method that integrates contrastive learning into sentence embedding, constructing both positive and negative sample pairs for better sentence embedding representation. The unsupervised method constructs positive sample pairs by two standard dropout methods, and the supervised SimCSE integrates sentence pairs with annotations into contrast learning based on performing sentence embedding. Wang et al.[11] applied the BERT model to evaluate the relevance of the query to the sentences in the document, and then the semantic information is integrated into a PRF model for document re-ranking, and the results of the experiments show that the scheme can improve the performance of the retrieval system effectively. Pan et al.[12] used the BERT model to obtain the sentence-level semantic information within the document, incorporated this information into the query expansion process, and significantly improved the effectiveness of document reordering. The experimental results demonstrate the feasibility and effectiveness of the proposed scheme.

The aforementioned methods have improved the selection of query expansion terms by incorporating sentence-level semantic information or by evaluating semantic similarity to assess the relevance between queries and documents. These approaches have effectively utilized pre-trained language models for semantic matching. However, relying solely on sentence-level semantic information may have limitations, such as an inability to fully capture the continuous information and contextual relationships within a document. To address this issue, we consider the semantic information at the paragraph level, utilizing the continuity of sentences to enhance the semantic coherence between sentences and provide a more comprehensive understanding of the context.

Passage level

Nogueira et al.[9] ranked the final retrieval results again by calculating the relevance of the query and the passages in the document. The results of the experiment show that this scheme is useful in improving retrieval ranking. Wu et al.[27] divided the documents into overlapping passages and inferred the document-level relevance scores based on the relevance of each passage. Passage relevance was graded based on BERT-MaxP, and the experimental analysis indicated that highly relevant documents had longer passages and more text compared to less relevant documents. The scores of the top k passages were selected as the relevance scores of the documents and were improved in terms of validity. Li et al.[28] segmented long documents into a fixed number of equal passages and aggregated the $[CLS]$ form for each passage. The experimental results demonstrate that the aggregated passage representation is more effective than the pooled passage scores. Zheng et al.[29] proposed a method where BERT is employed for the first-stage ranking, and the top-scoring document chunks are used for query expansion, leading to improved results after a second-stage ranking. Yu et al.[30] introduced the ANCE-PRF model, where the main idea is to concatenate the top K ranked passages from the first round of dense retrieval with the original query to form a new query for subsequent retrieval stages. Wang et al.[31] developed the ColBERT-PRF model, where query expansion is achieved by clustering pseudo-relevant documents from the first-stage dense retrieval, demonstrating strong performance in dense retrieval frameworks.

PRF assumes that the top N documents ranked in the initial retrieval result with the query are relevant, these documents are called feedback documents or pseudo-relevance documents. The query expansions are obtained in the feedback documents obtained from the initial retrieval round, and the query terms for the subsequent retrieval round are improved by different principles of selecting expansions. Then the newly acquired query is utilized for the subsequent retrieval. Numerous studies have demonstrated that PRF methods can effectively improve retrieval results[32]. Two types of mainstream PRF methods: probability-based PRF methods[33] and language modeling-based PRF methods[34]. Most of the PRF methods rank the feedback documents by considering the query-document match only from the perspective of terms. Valcarce et al.[33] introduced a linear method (LiMe) for PRF tasks. The LiMe framework calculates the similarity generated in the query and pseudo-relevance sets. Subsequently, this similarity is employed to expand the original query with additional similarity information. Pan et al.[34] studied based on the fact that terms with higher co-occurrence with query terms are more likely to be related to the query topic. The evaluation results of the proposed KRoc model and KRM3 model based on nuclear co-occurrence of PRF framework for enhanced IR exceeded those of the Rocchio model, RM3 model, and outperformed the MAP and P@10 results of most PRF models at that time on most data sets.

QE also known as query optimization, facilitate finding more relevant documents by supplementing or reorganizing the query representation to make the query closer to the user's true intent. Research on query expansion techniques falls into two main types: the first type of approach is relevance feedback-based query expansion, which generally uses the relevant documents judged by the user from the first round of returned results to expand the query. The second type of approach uses query expansion techniques based on PRF[34], and this technique assumes that documents ranked high after the initial retrieval round are pertinent to the query. Miao et al.[35] introduced a proximity-based model named PRoc, the article suggests that the terms that are close to the query distance may be related to the subject of the original query, and three proximity measures are used to evaluate the relationship between the expansion term and the query term. However, most existing PRF methods usually slice the original query into separate terms for processing, and select query expansion terms based on term frequency, location or proximity information, while the contextual semantic information in the original query may be ignored in the process. Padaki Ramith and Dai[36] leveraged BERT's capability to comprehend long natural language and proposed two methods for expanding short keyword queries in a BERT-based reranker. The first method adds structural words to transform brief queries into coherent natural sentences, while the second introduces new concepts for query expansion. Their experiments demonstrated that combining these methods yielded better performance in BERT-based re-rankers. Naseri et al.[37] introduced Contextualized Embeddings for Query Expansion (CEQE), utilizing query-centric context embeddings, and showed that neural reranking combined with CEQE achieved strong results. Tekli et al.[38] designed a full-fledged semantic indexing and querying model specifically for seamless integration into traditional RDBMS to address semantic-aware querying, extending RDBMS capabilities to support semantic-level operations. Jiang et al.[39] proposed an ontology-based semantic search method for open government data, improving search precision by associating data with well-defined ontological concepts, particularly for heterogeneous and poorly described datasets. Ngo et al.[40] developed a deep learning-based semantic search system for large-scale clinical ontologies, achieving notable results in free-text and concept-to-concept retrieval tasks. Tekli et al.[41] proposed a new XML keyword search method that integrates semantic information, combining offline processing and online query parsing, improving the quality and ranking of search results by leveraging context and global disambiguation techniques. Experimental results demonstrated the method's superior performance in enhancing search quality.

In statistical information-based PRF methods, query expansion has shown certain advantages, particularly in handling direct matching between queries and documents. However, these methods may overlook semantic information in the contextual environment and have limitations in identifying synonyms and semantic relationships. In recent years, the application of pre-trained language models in PRF has significantly addressed this issue. These methods have proven instrumental in capturing the semantic relationships between queries and documents, thus enhancing the alignment of user intent with relevant content in information retrieval systems. However, the information in pseudo-relevant documents should not be overlooked when selecting query expansion terms. Therefore, we propose to integrate semantic information into the Rocchio model based on the traditional PRF framework to help select query expansion terms more semantically similar to the query topic. The difference from the above studies is that the proposed PRF approach mainly focuses on obtaining multi-dimensional semantic information, and integrating the semantic information into the PRF for query expansion.
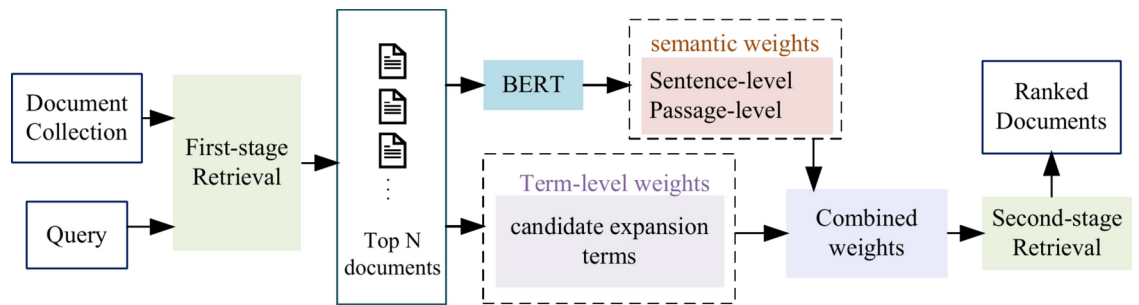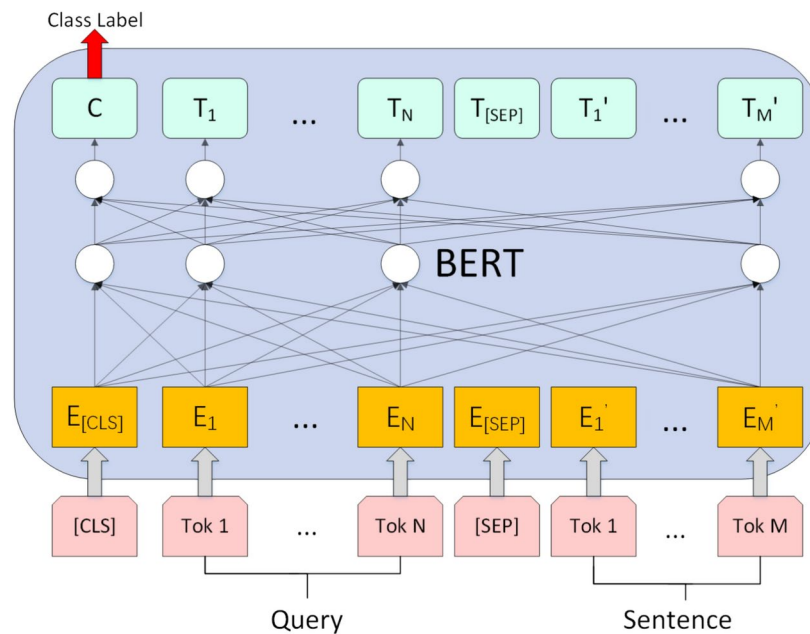
Fig. 1. PRF framework for multi-dimensional semantics.

Fig. 2. BERT model to get semantics.

Our proposed framework

In this section, we introduce the proposed model with a framework for sentence ranking and passage ranking. Especially, in our approach, we initially acquire the exact relevance weights of queries and documents at the lexical level using the relevant matching method. The initial ranking of documents utilized the scores computed through the traditional BM25 approach, and the top N documents were chosen for subsequent re-ranking. Subsequently, utilizing the relevance between sentences and queries and the relevance between passages and queries to identify documents relevance in which the sentence is located to the query. In particular, in our approach, we apply BERT to obtain the semantic message in addition to the traditional PRF model, thus obtaining semantic similarity between queries and documents. The detailed framework structure is illustrated in Fig. 1.

Sentence level

Because of the length limitation of BERT on the input sequence, we consider slicing and dicing the top-N documents obtained from the first retrieval stage. Simply speaking, the documents are divided into sentences, and then the query and each sentence are stitched together and sent to the BERT model, and the output gets the relevance scores of the query and the sentences. Based on the principle of local relevance, a document is considered relevant to a query if the sentences in the document are relevant to the query.

The first use of the BERT model to achieve the relevance of sentences $S_{i,j}$ and query Q judgment with the BERT model to calculate the semantic similarity of query and sentence is shown in the following Fig. 2.

$S_{i,j}$ denotes the j-th sentence of the i-th document in the document collection D;

Q is the user input query. The BERT model set in practice uses BERT-Base, including 12 sub-layers with the same structure. The query is connected with the sentence input, and the specific input sequence format is $[[CLS], Q, [SEP], S_{i,j}, [SEP]]$, and the final output of the model is at the 0-th character position, corresponding to the semantic similarity score of the input query and the sentence within the document.

Based on the principle of local relevance if some sentences within a document are semantically relevant to the query, then the terms within that sentence are also considered relevant to the original query Q. The weight of the terms t within the sentence is $w_t\prime$, the weights of terms within a sentence are expressed as in Eq. (1):

$$w_t\prime = R\left(S_{i,j}, Q\right) \tag{1}$$

Passage level
The segmentation of the documents in the document set D_1 obtained by BM25 is performed, in particular, by observing the distribution of passages in the dataset, where we use the natural segments of the documents in the dataset, and the p-th passage of document i is denoted as S_i^P. The semantic similarity score of each passage of document i with query Q is denoted as $R\left(S_i^p, Q\right)$ by the BERT method. The weights of the terms t within a segment are expressed as shown in Eq. (2).

$$w_t^{''} = R\left(S_i^p, Q\right) \tag{2}$$

Integration
As a classical probability model, BM25 can help calculate the frequency of all terms in PRF documents. Many scholars have researched and proved the effectiveness of the BM25 model. The formula for the weight of the term t obtained by the BM25 method is given in Eq. (3).

$$w_t^{'''} = \sum_{t \in Q} \log\left(\frac{N' - df_t + 0.5}{df_t + 0.5}\right) \times \frac{(k_1 + 1) \times tf(t,d)}{tf(t,d) + K} \times \frac{(k_3 + 1) \times tf(t,Q)}{k_3 + tf(t,Q)} \tag{3}$$

where N' denotes the number of total documents in the index, $tf(t,d)$ denotes the number of occurrences of term t in document d, $tf(t,Q)$ denotes the number of occurrences (frequency) of term t in a document Q, and df_t denotes the number of documents in which term t appears; and K is equal to $k_1 \times \left(1 - b + b \times \frac{dl}{avgdl}\right)$

, where k_1 and b are adjustment factors to balance the effect of document length, and k_3 is an adjustment parameter for adjusting the frequency of terms in the query. The term weights obtained by BM25 are expressed as $w_t^{'''}$. The weights of terms combining raw term frequency and semantic information are expressed as Eq. (4) where α and β are the adjustment coefficients to regulate the weights.

$$w_t = \alpha w_t^{'''} + (1 - \alpha)\left(\beta w_t^{''} + (1 - \beta) w_t\prime\right) \tag{4}$$

The semantic information is integrated into the traditional Rocchio model to obtain the new query terms, represented as shown in Eq. (5).

$$Q\prime = \mu Q_0 + (1 - \mu) Q_1 \tag{5}$$

The original query is denoted as Q_0, for Q_1 is the query expansion term and the new query is $Q\prime$, μ is a constant taking the value between 0 and 1, which is used to adjust the weighting ratio between the original query and the query expansion terms. The new query terms constructed were passed through BM25 for a second round of searching to obtain the final search results.

Experimental setup
Data sets and evaluation metrics
We selected TREC standard datasets, including AP90, AP88-89, DISK4&5, WT2G, and WT10G. These datasets, which vary in size and type, enable the evaluation of our proposed model, specifically, the BERT model is fine-tuned on the MS MARCO passage dataset. To facilitate the comparison and evaluation of our model's effectiveness with other PRF methods, we also conducted an assessment on the 2016 TREC Clinical Support Medicine dataset. Specifically, for the parameters of the BERT model, we choose the BERT-Base. The BERT-Base model has 12 sub-layers, each of which has the same structure. AP90 is an Associated Press article published in 1990, and AP88-89 is for articles published from 1988 to 1989. DISK4&5 contains newsletter articles from different sources, and these text data are usually considered as noisy and high-quality data. WT2G collection is a general crawl of web documents with a total of 2 g of uncompressed data for TREC8 web. The WT10G collection is a medium-sized crawl of web documents and is used in the TREC 9 and TREC 10 web tracks. It contains 10 g of uncompressed data. The TREC Clinical Decision Support Track collections consist of 1.25 million articles published on PubMed Central in 2016, including PMC-00, PMC-01, PMC-02, and PMC-03. Figure 3 shows the details of number 12 in the test topic file for the PMC dataset. First, the EHR admission note (only the HPI section, which is the "case"). Next, a more layman-friendly "description" similar to previous tracks, which removes much of the jargon and replaces clinical abbreviation for better readability. Finally, a "summary" similar to previous tracks, which is a 1–2 sentence summary of the description. Table 1 describes the details of the dataset and the query. Multiple sizes and types of datasets facilitate evaluation of model performance. For each topic, only short keywords are used for retrieval, and this is similar to the actual retrieval, and the QE method is more effective for short queries.

In the indexing and querying process, we performed word separation for each word item in all datasets used[42] and used the standard 418 stopwords[43]. We chose the main TREC evaluation metrics MAP and NDCG to

```
<topic number="12" type="test">
<note>
66 yo female pedestrian struck by auto. Unconscious and
unresponsive at scene. Multiple fractures and complication
secondary to the primary injury. S/p embolization of the avulsed
second branch of brachial artery, complicated by exp lap secondary
to suspicion of abdominal compartment syndrome. Not much of the
response after weaning the sedation but with minimal improvement
with CT of the head showing with extensive interparenchymal
hemorrhages throughout Tmax: 34.4 C (93.9 T current: 34.4 C (93.9
HR: 71 (71 - 88) bpm BP: 171/82(120) {158/74(113) - 171/
83(122)} mmHg RR: 24 (10 - 24) insp/min SPO2: 99% Heart
rhythm: SR (Sinus Rhythm)
</note>
<description>
66 yo female pedestrian struck by auto. Unconscious and
unresponsive at scene. Multiple fractures and complication
secondary to the primary injury. S/p embolization of the avulsed
second branch of brachial artery, complicated by exp lap secondary
to suspicion of abdominal compartment syndrome. Not much of the
response after weaning the sedation with CT of the head showing
extensive interparenchymal hemorrhages throughout.
</description>
<summary>
66 yo female pedestrian struck by auto. Unconscious and
unresponsive at scene. Multiple fractures and head CT showing
extensive interparenchymal hemorrhages.
</summary>
</topic>
```

Fig. 3. The example of topic.

Collection	Queries	Number of queries	Number of documents	Size
AP90	51–100	50	78,321	0.23 Gb
AP88-89	51–100	50	164,597	0.50 Gb
DISK4&5	301–450	150	528,155	1.86 Gb
WT2G	401–450	50	247,491	2.14 Gb
WT10G	451–550	100	1,692,096	10.00 Gb
PMC-00	1–30	30	263,175	16.90 Gb
PMC-01	1–30	30	240,347	15.80 Gb
PMC-02	1–30	30	389,431	21.20 Gb
PMC-03	1–30	30	357,047	19.60 Gb

Table 1. Data set and detailed description.

evaluate the validity of our proposed model; the MAP and NDCG of the top 1000 documents were chosen not only to consider the fairness of the evaluation metrics but also to take into account the problem that users only pay attention to the top-ranked documents returned in practical applications. Therefore, we also chose P@10 at some point to evaluate the validity of our proposed model.

Hyperparameter
The hyperparameters we use have a wide range of applications in the field of information retrieval, where the BM25 model takes b in the interval from 0 to 1 with a step size of 0.05; and the values of k_1 and k_3 are 1.2 and 0.8, which is consistent with the settings proposed by Robertson et al.[44], and the query expansion terms take values in the range of 10–50[35,45]. Based on previous work on PRF[46,47], the values of α, β and μ range from 0 to 1. To evaluate the proposed method on the medical dataset, we employed five-fold cross-validation[29]. The parameters learned from the training set were applied to the validation set for assessment.

Results and analyses
Experimental results compare with baseline model
To better evaluate our proposed model, the BM25 + Rocchio model and RM3 model are firstly selected as the strong baseline model for a comprehensive comparison with the proposed model, and the experimental results are analyzed in detail. In most cases, these two methods achieve better retrieval performance. Therefore, it is justified to use them as strong baseline models for PRF. The BM25 weighting computer mechanism, also known as the Okapi weighting computer mechanism, which focuses more on considering factors like term frequency and document length than the original BIM model. In the Rocchio framework, the TF-IDF model is usually utilized to select the top ranked candidate expansion terms in importance from the pseudo-relevance documents.

Model	AP90	AP88-89	DISK4&5	WT2G	WT10G
BM25 + Rocchio	0.2892	0.2940	0.2330	0.3254	0.2076
RM3	0.3041	0.3135	0.2561	0.3326	0.2122
MSRoc	**0.3317** (+14.69%)	**0.3210** (+9.18%)	**0.2612** (+12.10%)	**0.3430** (+5.41%)	**0.2146** (3.37%)

Table 2. Results are compared with those of the baseline model on MAP metrics. The best results obtained for each dataset are shown in bold. The percentage values in parentheses represent the enhancement of the proposed model with respect to the Rocchio-based model for certain data sets.

Model	AP90	AP88-89	DISK4&5	WT2G	WT10G
BM25 + Rocchio	0.4442	0.4566	0.4235	0.4925	0.3082
RM3	0.4421	0.4572	0.4277	0.4820	0.3084
MSRoc	**0.4681** (+5.38%)	**0.4918** (+7.71%)	**0.4333** (+2.31%)	**0.5060** (+2.74%)	**0.3194** (3.63%)

Table 3. Results are compared with those of the baseline model on P@10 metrics. The best results obtained for each dataset are shown in bold. The percentage values in parentheses represent the enhancement of the proposed model with respect to the Rocchio-based model for certain data sets.

The Rocchio model is known as one of the famous classical feedback models. Our focus is on incorporating sentence-level and passage-level semantic similarity into the Rocchio-based model, and therefore it is essential to compare our approach with the baseline of some collections.

The results in Tables 2 and 3 show that the proposed model improves on both MAP and P@10 evaluation metrics of the five standard TREC datasets, especially on MAP metrics, significant improvement on AP90 dataset and DISK4&5 dataset. The improvement was 14.69% on the AP90 dataset and 12.10% on the DISK4&5 dataset. On P@10 metrics AP90 dataset and AP88-89 dataset are optimized significantly. The improvement was 5.38% on the AP90 dataset and 7.71% on the AP88-89 dataset. The improvement in MAP and P@10 observed in the experimental results can be attributed to BERT's ability to encode latent semantic relationships between queries and documents, building on the initial round of exact matching. The outcomes of the above evaluation metrics prove the feasibility and validity of our proposed model.

Experimental results compared with PRF retrieval model

To better evaluate the feasibility and validity of our proposed model, our model is compared with advanced query expansion retrieval models on five standard TREC datasets in terms of MAP evaluation metrics of experimental results. Markov random field (MRF)[48] selects query expansion terms based on the language model. Hyperspace semantic model (HAL)[49] considers the idea of query expansion terms based on spatial distance, and terms that are closer in spatial distance to the original query terms are considered to be semantically closer to the original query as well. IF&FB[49] selects query expansion terms through the language model's framework, which are mainly evaluated by the context of the terms. For the proximity-based QE model, we selected only one of them, PRoc3[35], in the comparison of experimental results because of its low sensitivity. TF-PRF[50] is a query expansion term selection based on the co-occurrence information of terms based on the proximity-based model PRoc2. The following Table 4 shows the detailed results of comparing our approach with the corresponding QE model. The HAL-based Rocchio's model (HRoc)[51] was proposed based on the proximity relationships of terms and introduced three normalization methods to incorporate term proximity information into query expansion. In addition, we also compared the results of the 2016 TREC Clinical Support Medicine dataset with the results obtained using the PRF method. We present the experimental results of the 2016 TREC Clinical Support Medicine dataset with the PRF method in Table 5. The traditional baseline model BM25 results are also included to facilitate a clearer comparison of the experimental outcomes.

As shown in Table 4, among the MAP metric results of the compared models, our model achieves more desirable results on the datasets AP90 and AP88-89, with less improvement on the WT2G and WT10G datasets. Our models achieve optimal results compared to the compared models on the AP90 dataset and AP88-89 dataset, reaching 0.3317 on AP90 and 0.3201 on AP88-89. In addition to this, the improvement on DISK4&5 is also significant, with an increase of 15.68%. The enhancement of the traditional PRF retrieval method demonstrates the efficacy of our proposed model.

In our comparison with a range of PRF methods on the 2016 TREC Clinical Support Medicine dataset, we found that our proposed model consistently outperformed the contrasted models in terms of the average results for the MAP, P@10, and P@20 evaluation metrics. When compared to BM25, our model exhibited an average improvement of 81.69%, 18.89%, and 27.00% in the three evaluation metrics, respectively. The most significant enhancement was observed in the MAP metric, with a value of 0.0814. Our proposed model demonstrated a significant advantage over the compared models on the 2016 TREC Clinical Support Medicine dataset. The positive performance of the MSRoc model on the 2016 TREC Clinical Support Medicine dataset can be clearly observed through Table 5 and Fig. 4.

Model	AP90	AP88-89	DISK4&5	WT2G	WT10G
BM25	0.2738	0.2882	0.2258	0.3192	0.2050
MRF	0.2920	0.3088	0.2579	0.3380	0.2214
	(+6.65%)	(+7.15%)	(+14.22%)	(+5.89%)	(+8.00%)
HAL	0.2810	0.2916	0.2363	0.3285	0.2158
	(+2.63%)	(+1.18%)	(+4.65%)	(+2.91%)	(+5.27%)
IF&FB	0.2886	0.2971	0.2565	0.3301	0.2180
	(+5.41%)	(+3.09%)	(+13.60%)	(+3.41%)	(+6.34%)
PRoc3	0.3181	0.3179	0.2575	**0.3534**	0.2256
	(+16.18%)	(+10.31%)	(+14.04%)	(+10.71%)	(+10.05%)
TF-PRF	0.3074	0.3190	**0.2699**	0.3448	**0.2350**
	(+12.27%)	(+10.69%)	(+19.53%)	(+8.02%)	(+14.63%)
MSRoc	**0.3317**	**0.3201**	0.2612	0.3430	0.2146
	(+21.15%)	(+11.07%)	(+15.68%)	(+7.46%)	(+4.68%)

Table 4. Results of the proposed model are compared with a state-of-the-art PRF method using query expansion based on five TREC datasets in terms of MAP metrics. The values in parentheses are augmentations to the baseline BM25. Bold values indicate the best results for each set.

Model	MAP	P@10	P@20
BM25	0.0448	0.2467	0.2100
BM25+Rocchio	0.0490	0.2533	0.2167
RM3	0.0540	0.2467	0.2233
PRoc2	0.0600	0.2533	0.2417
PRoc3	0.0593	0.2300	0.2167
TF-PRF	0.0580	0.2467	0.2317
HRoc1	0.0651	0.2733	0.2350
HRoc2	0.0647	0.2667	0.2317
HRoc3	0.0642	0.2317	0.2317
MSRoc	**0.0814** (+81.69%)	**0.2933** (+18.89%)	**0.2667** (+27.00%)

Table 5. Results of this model were compared with the query expansion-based PRF methods on the 2016 TREC Clinical Support Medicine dataset. Values in parentheses indicate the increment over the baseline BM25. Bold values represent the best results for each collection.
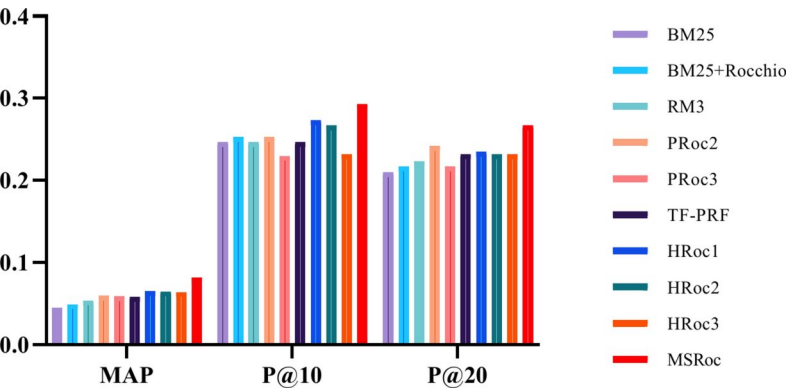

Fig. 4. Results for 10 models on the 2016 TREC clinical support medicine dataset.

Experimental results compared with neural information retrieval model

To better estimate the feasibility and validity of our proposed model, we compare several models based on neural IR methods with our proposed model in respect to canonical evaluation metrics. The neural IR models include the Deep Structured Semantic Model (DSSM)[52], which converts a sparse term vector represented by a one-hot vector into a dense term vector, constructs a representation of query and document using a fully connected

feedforward network, and assesses the document's relevance to the query by calculating the cosine similarity between the two vectors. Convolutional Latent Semantic Model (CDSSM)[53] is derived from an expansion of DSSM, specifically, it uses convolutional neural networks to improve the preservation of local term order information while acquiring the contextual information of document queries, and uses a maximum pooling strategy to filter semantic concepts. Deep Relevance Matching Model (DRMM)[6] uses a matching histogram approach that considers term information. Deep text matching model (MatchPyramid)[54] can effectively identify significant information by constructing a matching matrix to obtain the similarity between terms. And Transformer-Kernel pooling model for long text (TKL)[55] uses a local self-attention approach to move over the document using a fixed-length hole-in-the-ground window to consider the input of text by decreasing the calculation complexity of the self-attention mechanism, with query direct interaction. In addition to this, the proposed model is also compared with the PRF models SRoc[12] using deep learning, SRoc is the sentence-level semantic information obtained through the BERT model.

As shown in Table 6 and Fig. 5, the feasibility and validity of our proposed model can be better assessed on the four widely used evaluation metrics of the five standard TREC datasets. In the AP90 dataset, MAP, NDCG and MRR showed better results for the compared models, although the advantages in MAP and NDCG results were not so obvious, but the MRR results reached 0.6628 outperforming the BM25 model by 20.05%. In the AP88-89 dataset, P@10, NDCG and MRR all show better performance than the compared models, even though the improvements in P@10 and NDCG are smaller, the MRR result reached 0.6895. Different degrees of improvement were observed in MAP, NDCG and MRR metrics on DISK4&5 datasets, and the most significant

Model	AP90	AP88-89	DISK4&5	WT2G	WT10G
MAP					
BM25	0.2738	0.2882	0.2258	0.3192	0.2050
DSSM	0.1364	0.1425	0.1247	0.1649	0.1136
CDSSM	0.1069	0.1137	0.0970	0.1399	0.1039
DRMM	0.2770	0.2994	0.2558	0.3237	0.2155
MatchPyramid	0.2858	0.3022	0.2523	0.3329	0.2256
TKL	0.2998	0.2988	0.2532	0.3347	0.2309
SRoc	0.3271	0.3197	0.2598	**0.3444**	**0.2374**
MSRoc	**0.3317**	**0.3201**	**0.2612**	0.3430	0.2146
P@10					
BM25	0.4468	0.4531	0.4313	0.4760	0.3061
DSSM	0.2334	0.2417	0.2251	0.2399	0.1579
CDSSM	0.2135	0.2243	0.2058	0.2052	0.1343
DRMM	0.4595	0.4642	0.4459	0.4901	0.3283
MatchPyramid	0.4679	0.4686	0.4402	0.4956	0.3464
TKL	**0.4721**	0.4768	0.4525	0.5097	**0.3585**
SRoc	0.4634	0.4729	**0.4542**	**0.5130**	0.3551
MSRoc	0.4681	**0.4918**	0.4333	0.5060	0.3194
NDCG					
BM25	0.6579	0.6709	0.6678	0.7058	0.5819
DSSM	0.3342	0.3411	0.3479	0.3624	0.2966
CDSSM	0.2896	0.2637	0.3196	0.3352	0.2784
DRMM	0.6682	0.6693	0.6584	0.6943	0.6022
MatchPyramid	0.6732	0.6734	**0.6685**	0.7053	0.6038
TKL	0.6779	0.6828	0.6618	0.6972	0.6037
SRoc	0.6844	0.6801	0.6566	**0.7195**	**0.6174**
MSRoc	**0.6860**	**0.6876**	0.6639	0.7170	0.5777
MRR					
BM25	0.5521	0.5378	0.5608	0.6623	0.5191
DSSM	0.3744	0.3598	0.3352	0.3784	0.2946
CDSSM	0.3525	0.3392	0.2884	0.3393	0.2658
DRMM	0.5824	0.5436	0.5786	0.6783	0.5239
MatchPyramid	0.5758	0.5583	0.5885	0.6828	**0.5339**
TKL	0.5954	0.5604	0.5835	0.6994	0.5283
SRoc	0.6391	0.5639	0.6060	0.7140	0.5338
MSRoc	**0.6628**	**0.6895**	**0.6242**	**0.7473**	0.5299

Table 6. Experimental results of our proposed model are compared with those of the neural retrieval model for different metrics on five standard TREC datasets. Significant values are given in bold.
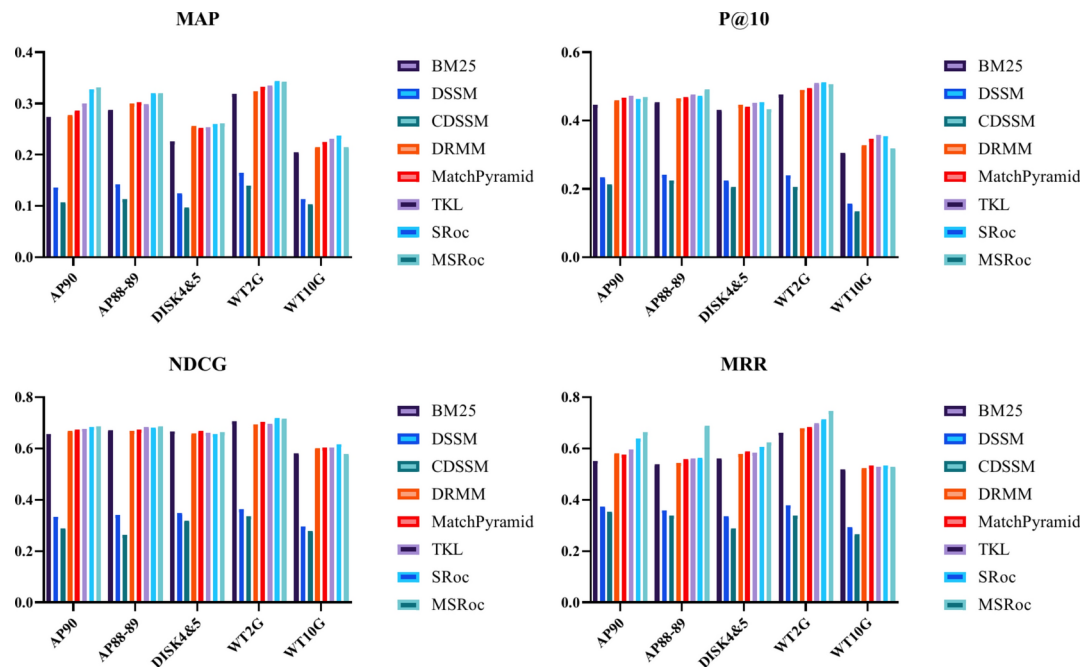
Fig. 5. Performance of 8 models in MAP, P@10, NDCG and MRR metrics.

Collection	Metric	Rocchio	PRoc2	KRoc	MSRoc
AP90	NDCG	0.6682	0.6689	0.6810	0.6824
	MRR	0.6106	0.5777	0.6259	0.6337
AP88-89	NDCG	0.6745	0.6716	0.6812	0.6855
	MRR	0.5337	0.5415	0.5534	0.6571
DISK4&5	NDCG	0.6564	0.6580	0.6701	0.6639
	MRR	0.5903	0.5989	0.6037	0.6242
WT2G	NDCG	0.7085	0.7093	0.7138	0.7170
	MRR	0.6830	0.6907	0.6973	0.7473
WT10G	NDCG	0.5778	0.5838	0.5787	0.5777
	MRR	0.5207	0.5194	0.5286	0.5299

Table 7. Experimental results of our proposed model on MRR and NDCG metrics compared with a strong baseline model on five standard data sets.

improvement was observed in MRR metrics in particular. The feasibility of our proposed model is demonstrated in the WT2G dataset for four metrics, and the MRR metric reaches 0.7473 and is better than all the comparative models. The models compared in this section all leverage different deep learning frameworks to compute the semantic relationships between queries and documents. These latent semantic relationships can effectively improve the ranking of retrieval results. The improvement of the proposed model can be attributed to the fact that single-sentence semantics may not fully capture the intended meaning, whereas contextual information from consecutive sentences provides better semantic understanding. However, for datasets such as DISK4&5, WT2G, and WT10G, the reason why some evaluation metrics did not achieve optimal results may be due to the presence of irrelevant information within the relevant passages, which introduces semantic noise and affects the model's performance.

As shown in Table 7, in terms of NDCG and MRR metrics, contrast the corresponding Rocchio model-based approach, our model has improved to different degrees from the compared models on each data set, especially the improvement in MRR metrics is more obvious. The reason for analyzing the better MRR is that our model ranks the relevant documents in the ranking results of the top 1000 document sets more highly than the previous models, it can meet the user requirements in practical applications to some extent.

Experimental results compared with SOTA retrieval model

To further validate the advancements of the proposed method, this section presents a comparative analysis between our approach and several representative models currently utilized in the field of information retrieval. Query Description[36] enhances the original query by incorporating structural words and additional terms that

contribute to the construction of coherent natural language sentences, thereby introducing new concepts. BERT-QE[29], a prominent retrieval model based on BERT, is characterized by its approach of selecting the highest-scoring text blocks for re-ranking after the first round of BERT sorting. Given that this study employs the BERT-base model, we specifically choose to compare it with the BERT-QE-LBL model. Additionally, Contextualized Embedding for Query Expansion (CEQE)[37] utilizes query-centric contextualized embedding vectors. Among the CEQE models, CEQE-MaxPool demonstrates superior performance by optimizing the selection of expanded terms through the MaxPool method, significantly enhancing the effectiveness of query expansion and improving the overall performance of the retrieval system. In the dense retrieval methods, the ColBERT-PRF[31] model identifies a set of pseudo-relevant documents through the first round of dense retrieval, subsequently employing the K-means method to select representative feedback embeddings, which are then integrated into the query representation for re-ranking. Another dense retrieval method, ANCE-PRF[30], connects the top K ranked paragraph texts from the first round of retrieval feedback with the original query to create a new query input, followed by another round of dense retrieval and re-ranking. To facilitate a clearer visualization of the results, we utilize bar charts to present the comparative outcomes, with specific information detailed in Fig. 6. The datasets selected for this study are Robust04[56] and TREC DL 2020[57], with the chosen metrics aimed at maintaining consistency with prior research efforts to ensure the comparability and validity of the results.

In Fig. 6a, we present the comparative results of our model against advanced models on the Robust04 dataset. According to the MAP metric, the proposed model demonstrates a certain level of competitiveness. However, in comparison to the CEQE-MaxPool model, the performance of our model is relatively inferior. This phenomenon may be attributed to the CEQE model's effective utilization of high-dimensional embedding representations of terms and their contextual information when selecting documents. The semantic information contained within these embeddings may be more comprehensive than that derived from the weighted semantic information in the Rocchio method, thus optimizing the selection for query expansion. The results in Fig. 6b further indicate that the proposed model possesses a degree of comparability in terms of performance. Considering that we employed a sparse retrieval method in the first round of retrieval, the dense retrieval models, ColBERT-PRF and ANCE-PRF, have assessed the semantic relevance between queries and documents when selecting the set of pseudo-relevant feedback documents. This semantic-level evaluation likely provides more effective support for subsequent retrieval processes, significantly enhancing the results compared to the pseudo-relevant document sets obtained solely through sparse retrieval methods. Therefore, we believe that future research should place greater emphasis on integrating contextual information and semantic relevance in the document selection process to further optimize the effects of query expansion and re-ranking. This approach will contribute to improving the overall performance of information retrieval systems, better meeting users' retrieval needs.

Weighting adjustment factor

Analyzing the reasons that may affect the robustness of our proposed model, it is considered that one of the important parameters regulating the ratio of original query to query expansion term weights is the parameter μ. μ is the main parameter that affects the ratio of expansion term weights. A smaller μ value implies that more sentence-level and passage-level semantic information is taken into account in the selection of new query terms. To analyze the specific performance of μ, we varied the μ value from 0 in steps of 0.1 to 1 in our experimental setup. The detailed outcomes of the proposed model and the comparison model on MAP evaluation metrics are depicted in Fig. 7.

We utilize MAP value result change curves to analyze the best performing weight parameter value and the best number of query expansion terms $|Tf|$ ($|Tf| = \{10, 20, 30, 50\}$) for each dataset, and the conclusion can be drawn in Fig. 7. According to Fig. 7, we observed that the MAP metrics performed better on the used datasets when μ was taken from 0.3 to 0.6. In general, for most of the datasets, MAP results improved with increasing semantic information in a certain range, especially for the WT2G dataset when was taken from 0.1,
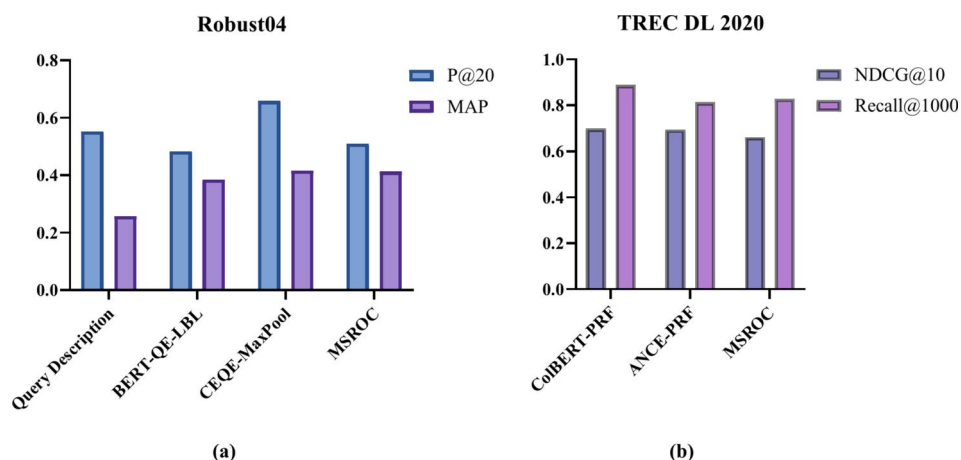

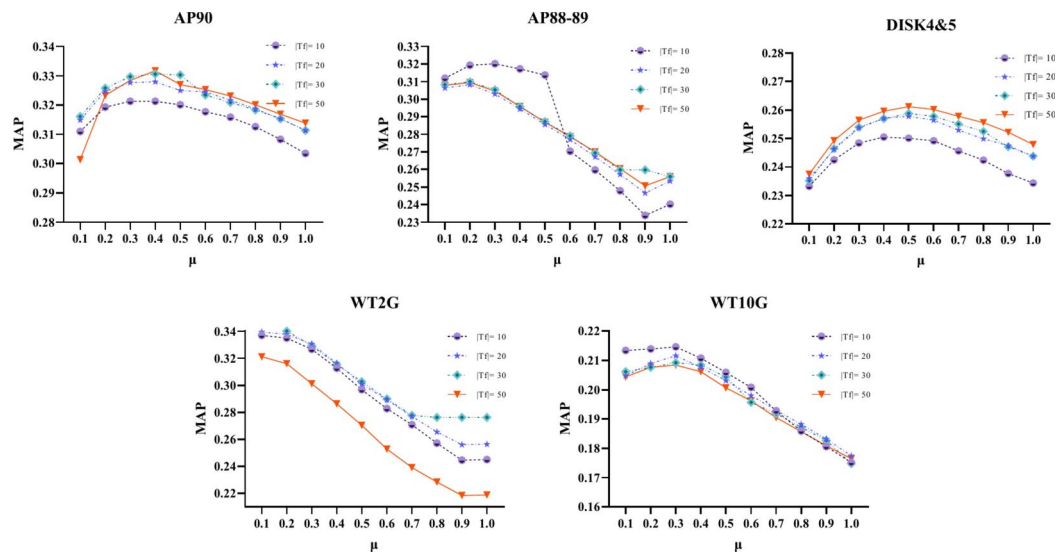**Fig. 6**. Comparison Results Showing Superiority Over 5 SOTA Models on the Robust04 and TREC DL 2020 Datasets.

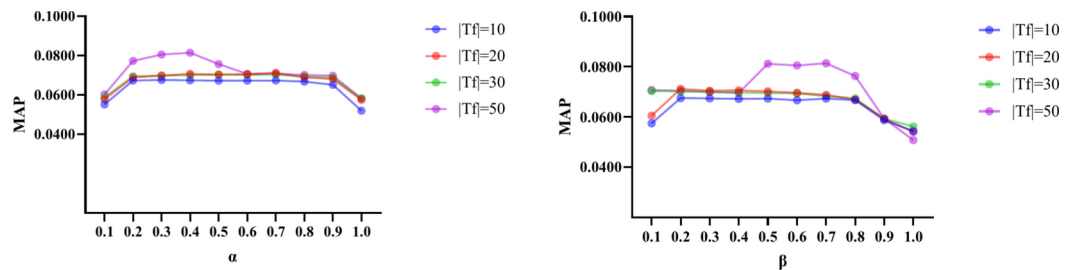**Fig. 7**. Sensitivity of the weighting parameter for different $|Tf|$.

Fig. 8. Sensitivity of α and β.

which means that semantic information improved the retrieval performance. We further found that passage semantic information improved MAP results by helping to obtain the overall document semantics. To obtain better results, we recommend that the value of μ should be in the range of 0.2–0.6 for most of the datasets, and when we consider the effect of query expansion $|Tf|$, we recommend $|Tf| = 30$ for most of the datasets, but we should also consider the results with $|Tf| = 10$ and $|Tf| = 50$.

To analyze the importance of term-level and semantic information, we also examined the impact of the parameters on retrieval performance using the 2016 TREC Clinical Support Medicine dataset, as detailed in Fig. 8. An increase in the α indicates that less semantic information is incorporated when selecting query expansion terms. The experimental results show that when selecting 50 candidate expansion terms, the model achieves relatively superior retrieval performance. This suggests that moderately incorporating semantic information is crucial for improving retrieval effectiveness. When semantic information is completely disregarded, the model's performance significantly degrades, further validating the critical role of semantic information in enhancing retrieval accuracy. Moreover, we analyzed the β impact on sentence-level and paragraph-level information. The results indicate that the best performance is achieved when the β is in the range of 0.5–0.7. This implies that both single-sentence information and the information from consecutive sentences are equally important in determining the relevance between a query and documents. Therefore, considering both local sentence-level information and contextual paragraph-level information is essential for effectively improving the performance of query expansion. These findings provide strong theoretical support for optimizing information retrieval systems and highlight the importance of integrating both semantic and term-level information.

Case study

To comprehensively evaluate the effectiveness of the proposed method in selecting expansion terms, we selected topic 53 "Leveraged Buyouts" from the AP90 dataset for detailed analysis. Leveraged Buyouts, also known as leveraged acquisitions or debt-financed acquisitions, are a financial strategy in which a company or individual uses the target company's assets as collateral in order to secure debt to finance the acquisition.

The document shown in Fig. 9 is ranked 38th by the ROC method and 17th by the MSRoc method.

Through an analysis of the detailed document information in Fig. 9, we found that the drugstore chain Revco, during its bankruptcy restructuring, planned to lay off employees and sell some of its stores. Revco announced that it would sell 221 of its 1873 stores to Reliable Holdings Corp., based in Fort Worth, Texas. The underlying

```
<HEAD>Revco Agrees to Sell 221 Stores</HEAD>
<DATELINE>TWINSBURG, Ohio (AP) </DATELINE>
<TEXT>
   Revco D.S. Inc. said Friday it agreed to
sell 221 of its 1,873 stores to Reliable Holdings Corp. of Fort
Worth, Texas, as part of its restructuring under bankruptcy court
protection.
   Revco, one of the largest drugstore chains in the country with a
work force of 26,800 people, announced Jan. 26 that it planned to
sell 712 of its stores.
   The price of the deal with Reliable Holdings was not disclosed.
   Weighed down by debt incurred in a takeover, the chain had filed
for protection from creditors under Chapter 11 of the federal
bankruptcy law in July 1988. Its restructuring involves selling
stores in the western United States and emphasizing its operations
in the East, where it gets most of its sales.
   Revco will sell to Reliable Holdings all its stores in western
Tennessee and in Illinois, Iowa, Indiana, Kansas, Kentucky,
Michigan and Missouri, Revco said.
   Reliable Holdings Corp. was formed by Acadia Partners LP, and
the management of Reliable Drug Stores Inc., a 112-store chain
based in Indianapolis.
   ``We anticipate a very smooth transition of these stores to
Reliable,'' said Boake A. Sells, Revco chairman and chief executive
officer.
   Friday's announcement is the first involving buyers for the
stores. Any sales are subject to bankruptcy court approval.
   Revco officials in January said they hoped that those working in
stores targeted for sale would be hired by eventual buyers.
   The company said its business in 10 Eastern states accounts for
64 percent of the chain's sales.
   The states are Georgia, Maryland, New York, North Carolina,
Ohio, Pennsylvania, South Carolina, eastern Tennessee, West
Virginia and Virginia.
   Other Revco stores are in Alabama, Arizona, Arkansas, Colorado,
Florida, Louisiana, Michigan, Mississippi, New Mexico, Oklahoma and
Texas.
   Revco has annual sales of more than $2.7 billion.
   The company filed for Chapter 11 protection from creditors in
July 1988, after it failed to make an interest payment on
outstanding ``junk bonds'' used to finance its $1.5 billion,
management-led leveraged buyout the previous year.
   In a leveraged buyout, the buyer of the company borrows the
purchase money with the idea of repaying it through asset sales and
earnings.
</TEXT>
```

Fig. 9. Document example of query no. 53: "Leveraged Buyouts".

reason for this restructuring was the substantial debt Revco accumulated during a prior leveraged buyout. Although the exact term "Leveraged Buyout" appears only twice in the document, there is a strong semantic relevance between the query and the document.

Multi-dimensional semantic information played a crucial role in this case by uncovering deeper semantic connections between the query and the document. For instance, while Revco's restructuring did not explicitly reiterate the details of the leveraged buyout, information related to debt restructuring, asset sales, and other financial actions is semantically linked to "Leveraged Buyouts." By leveraging multidimensional semantic information, the retrieval system can more precisely capture these latent semantic relationships, thereby improving both the accuracy and relevance of the search results.

This section compares the expansion terms generated by the MSRoc model with those from the baseline model Rocchio. Table 8 presents the top 15 expansion terms and their corresponding weights generated by both models, with all terms undergoing preprocessing steps such as stop words removal, stemming, and normalization.

Through a thorough comparative analysis, we found that both methods can generate effective expansion terms relevant to the initial query. However, the MSRoc model, which is designed to emphasize the extraction of deep semantic information from pseudo-relevant documents, demonstrated a significant advantage in identifying expansion terms that are highly semantically aligned with the query. For instance, terms closely related to the "Leveraged Buyouts" theme, such as "debt" and "financ" received significantly higher weights in the MSRoc model. Additionally, the MSRoc model successfully introduced some new expansion terms like "firm" and "deal," which are also highly relevant to the query topic, thereby enriching the semantic expression of the query.

This case study not only validates the superior performance of the MSRoc model in selecting expansion terms but also underscores the importance of leveraging pre-trained models to deeply mine multi-dimensional semantic information from pseudo-relevant documents. This approach enhances the flexibility and adaptability

Rocchio		MSRoc	
buyout	1.1553	buyout	1.0000
leverag	1.1011	leverag	0.9619
revco	0.1249	revco	0.1188
zaretski	0.0673	debt	0.0845
debt	0.0636	kohlberg	0.0820
campeau	0.0543	Lbo	0.0791
safewai	0.0540	campeau	0.0787
bankruptci	0.0531	junk	0.0726
kohlberg	0.0511	financ	0.0711
financ	0.0450	zaretski	0.0710
examin	0.0448	kravi	0.0705
lbo	0.0418	bankruptci	0.0700
store	0.0409	takeov	0.0696
junk	0.0407	**doskocil**	0.0655
gottstein	0.0407	**borrino**	0.0601
takeov	0.0396	**makovski**	0.0601
kravi	0.0340	southland	0.0599

Table 8. Rocchio model and MSRoc model on query topic 53 "Leveraged Buyouts" for the top 15 query expansion terms and corresponding weights. Terms highlighted in bold represent new terms selected by the MSRoc model.

of the model, providing a richer and more diverse semantic perspective for the selection of expansion terms. By doing so, the MSRoc model can more accurately capture the latent intent of the query, leading to a significant improvement in overall information retrieval performance. This research offers valuable insights and guidance for future advancements in query expansion technology.

Summary and future work

In this paper, we propose a novel framework that incorporates the multidimensional semantic information from pseudo-relevant documents to enhance query expansion term selection. Our approach not only leverages term frequency data but also integrates multidimensional semantic similarity to identify optimal expansion terms. When multiple candidate expansion terms share similar frequencies, our method can more accurately select those that are semantically closer to the query from pseudo-relevant documents. Furthermore, our study demonstrates the effectiveness of combining term-level importance with semantic-level importance, showing that the integration of multidimensional information significantly improves query expansion performance. The proposed model goes beyond single-sentence semantics by also considering the contextual semantics of consecutive sentences, thereby capturing the overall semantic representation of documents more comprehensively and facilitating the identification of multidimensional semantic expansion terms.

In future work, we plan to further optimize the computational efficiency of the model by exploring lighter-weight architectures to reduce resource consumption. Our proposed framework will also be applied for more applications (such as biomedical and chemical IR)[58–60]. Additionally, we aim to investigate how to retrieve more relevant pseudo-relevant documents in the initial retrieval phase, thereby enhancing the overall performance of the retrieval system.

Data availability

References

1. Salton, G., Wong, A. & Yang, C. S. A vector space model for automatic indexing. *Commun. ACM* **18**, 613–620 (1975).
2. Robertson, S. & Jones, K. S. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **27**, 129–146 (1976).
3. Zhai, C. & Lafferty, J. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)* 403–410 (ACM Press, 2001).
4. Robertson, S. E. & Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Confer-ence on Research and Development in Information Retrieval (SIGIR'94)* 232–241 (1994).
5. Jian, F., Huang, X., Zhao, J., Ying, Z. & Wang, Y. A topic-based term frequency normalization framework to enhance probabilistic information retrieval. *Comput. Intell.* **36**, 486–521 (2020).
6. Guo, J., Fan, Y., Ai, Q. & Croft, W. B. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Manage-ment (CIKM'16)* 55–64 (ACM, 2016).

7. Hofst, S., Zlabinger, M. & Hanbury, A. Interpretable & time-budget-constrained contextualization for re-ranking. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI'20)* (2020).
8. Mitra, B., Diaz, F. & Craswell, N. Learning to match using local and distributed representations of text for web search (2016).
9. Nogueira, R. & Cho, K. Passage Re-ranking with BERT. (2019).
10. Xiong, C., Dai, Z., Callan, J., Liu, Z. & Power, R. End-To-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)* 55–64 (Association for Computing Machinery, Inc, 2017).
11. Wang, J. et al. A Pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Inf. Process Manag.* **57** (2020).
12. Pan, M. et al. A probabilistic framework for integrating sentence-level semantics via BERT into pseudo-relevance feedback. *Inf. Process Manag.* **59**, (2022).
13. Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inf. Process Syst.* (2020).
14. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Conference of the North Ameri-can Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT'19)* 4171–4186 (2019).
15. Lewis, M. et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://huggingface.co/transformers.
16. Radford, A. et al. Language Models Are Unsupervised Multitask Learners. https://github.com/codelucas/newspaper.
17. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer (2019).
18. Pan, M. et al. SPRF: A semantic Pseudo-relevance Feedback enhancement for information retrieval via ConceptNet. *Knowl. Based Syst.* **274** (2023).
19. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP'19) 3982–3992 (2020).
20. Ye, Z., Huang, J. X. & Lin, H. Finding a good query-related topic for boosting pseudo-relevance feedback. *J. Am. Soc. Inf. Sci. Technol.* **62**, 748–760 (2011).
21. Huang, X., Zhong, M. & Si, L. York University at TREC 2005: Genomics Track. in *TREC* (2005).
22. Hristea, F. & Colhon, M. The long road from performing word sense disambiguation to successfully using it in information retrieval: An overview of the unsupervised approach. *Comput. Intell.* **36**, 1026–1062 (2020).
23. Wang, J. et al. Utilizing BERT for information retrieval: Survey, Applications, resources, and challenges. *ACM Comput. Surv.* **56**, 1–33 (2024).
24. Yang, W., Zhang, H. & Lin, J. Simple Applications of BERT for Ad Hoc ocument Retrieval. (2019).
25. Yilmaz, Z. A., Yang, W., Zhang, H. & Lin, J. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Lan-guage Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 3488–3494 (Association for Computational Linguistics, 2019).
26. Gao, T., Yao, X. & Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 6894–6910 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021).
27. Wu, Z. et al. Leveraging Passage-Level Cumulative Gain for Document Ranking. In Proceedings of the Web Conference 2020 2421–2431 (Association for Computing Machinery, New York, NY, USA, 2020).
28. Li, C., Yates, A., MacAvaney, S., He, B. & Sun, Y. PARADE: Passage representation aggregation for document reranking. arXiv:abs/2008.09093 (2020).
29. Zheng, Z. et al. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In Findings of the Association for Computational Linguistics: EMNLP 2020 4718–4728 (Association for Computational Linguistics, Online, 2020).
30. Yu, H., Xiong, C. & Callan, J. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management 3592–3596* (Association for Computing Machinery, New York, NY, USA, 2021).
31. Wang, X., Macdonald, C., Tonellotto, N. & Ounis, I. Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* 297–306 (Association for Computing Machinery, New York, NY, USA, 2021).
32. Huang, J. X., Miao, J. & He, B. High performance query expansion using adaptive co-training. *Inf. Process Manag.* **49**, 441–453 (2013).
33. Valcarce, D., Parapar, J. & Barreiro, Á. Document-based and term-based linear methods for pseudo-relevance feedback. *Appl. Comput. Rev.* **18**, 5–17 (2019).
34. Pan, M. et al. A simple kernel co-occurrence-based enhancement for pseudo-relevance feedback. *J. Assoc. Inf. Sci. Technol.* **71**, 264–281 (2020).
35. Miao, J., Huang, J. X. & Ye, Z. Proximity-based Rocchio's model for pseudo relevance feedback. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in In-formation Retrieval (SIGIR'12)* 535–544 (ACM Press, 2012).
36. Ramith, P., & Dai, Z. Rethinking query expansion for BERT reranking. In *Advances in Information Retrieval* (ed. Jose Joemon M. and Yilmaz, E. and M. J. and C. P. and F. N. and S. M. J. and M. F.) 297–304 (Springer International Publishing, Cham, 2020).
37. Naseri, S., Dalton, J., Yates, A. & Allan, J. CEQE: Contextualized Embeddings for Query Expansion. in 467–482 (2021).
38. Tekli, J. et al. Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS. *Data Knowl. Eng.* **117**, 133–173 (2018).
39. Jiang, S., Hagelien, T. F., Natvig, M. K. & Li, J. Ontology-Based Semantic Search for Open Government Data. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* 7–15 (2019).
40. Ngo, D.-H., Kemp, M., Truran, D., Koopman, B. & Metke-Jimenez, A. Semantic Search for Large Scale Clinical Ontologies. (2022).
41. Tekli, J., Tekli, G. & Chbeir, R. Combining offline and on-the-fly disambiguation to perform semantic-aware XML querying. *Comput. Sci. Inf. Syst.* **20**, 423–457 (2023).
42. Porter, M. F. An algorithm for suffix stripping. *Program* **40**, 211–218 (2006).
43. Callan, J. P., Croft, W. B. & Broglio, J. TREC and TIPSTER experiments with inquery. *Inf. Process Manag.* **31**, 327–343 (1995).
44. Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M. & Payne, A. Okapi at TREC-4. In *Proceedings of the 4th Text REtrieval Conference (TREC'95)* 73–96 (Gaithersburg, MD: NIST, 1995).
45. Ye, Z. & Huang, J. X. A simple term frequency transformation model for effective Pseudo Relevance Feedback. In *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval 323–332* (Association for Computing Machinery, 2014).
46. Lavrenko, V. & Croft, W. B. Relevance-based language models. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)* 120–127 (Springer, 2001).
47. Li, C. *et al.* NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. In *Proceedings of the 23rd Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)* 4482–4491 (2018).
48. Metzler, D. & Croft, W. B. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)* 472–479 (2005).

49. Bai, J., Song, D., Bruza, P., Nie, J. Y. & Cao, G. Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM International Conference on Infor-mation and Knowledge Management (CIKM'05)* 688–695 (2005).

50. Zhao, J., Huang, J. X. & Ye, Z. Modeling term associations for probabilistic information retrieval. *ACM Trans. Inf. Syst.* **32**, 1–47 (2014).

51. Pan, M. et al. An adaptive term proximity based rocchio's model for clinical decision support retrieval. *BMC Med. Inform Decis. Mak.* **19**, 251 (2019).

52. Huang, P.-S. et al. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management (CIKM'13)*, pp. 2333–2338 (2013).

53. Shen, Y., He, X., Gao, J., Deng, L. & Mesnil, G. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, pp. 373–374 (2014).

54. Pang, L. et al. Text matching as image recognition. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, pp. 2793–2799 (2016).

55. Hofstätter, S., Zamani, H., Mitra, B., Craswell, N. & Hanbury, A. Local self-attention over long text for efficient document retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Re-search and Development in Information Retrieval (SIGIR'20) 2021–2024* (ACM, 2020).

56. Voorhees, E. M. Overview of the TREC 2004 robust track. In *Thirteenth Text Retrieval Conference* (2004).

57. Craswell, N., Mitra, B., Yilmaz, E. & Campos, D. Overview of the TREC 2020 deep learning track. (2021).

58. Huang, X. & Hu, Q. A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* 307–314 (2009).

59. Lupu, M., Huang, J., Zhu, J. & Tait, J. TREC-CHEM: Large scale chemical information retrieval evaluation at TREC. *Acm Sigir. Forum* **43**, 63–70 (2009).

60. Lupu, M., Piroi, F., Huang, X., Zhu, J. & Tait, J. Overview of the TREC 2009 chemical IR track. In *TREC* (2009).

Acknowledgements

Author contributions

M. P.: design, conception of the study, algorithms design. Y. L.: algorithms design, data collection, data analysis, data interpretation, manuscript drafting. J.C.: validation, data analysis. E. H.: Writing—Review & Editing. J. H.: Supervision, Project admin-istration, Funding acquisition.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.L., J.C. or J.X.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.