Multimedia Tools and Applications (2024) 83:10393–10407
https://doi.org/10.1007/s11042-023-15831-x

Understanding the impact of query expansion on federated search

Adamu Garba[1] · Shah Khalid[2] · Irfan Ullah[3]

Received: 19 January 2022 / Revised: 27 March 2023 / Accepted: 10 May 2023 /
Published online: 21 June 2023

Abstract

Query expansion (QE) has been studied extensively in traditional search settings due to its efficacy in improving retrieval performance. However, the level of performance achieved in the traditional settings has not been reported in the literature on the federated search. Some of the possible reasons include the lack of complete information regarding the corpus statistics of the databases and their diverse content. Nevertheless, several studies have experimented with different QE approaches and reported mixed results. This paper extends the findings of these publications by studying the impact of using a different source for selecting terms to be used in QE on the federated search. Specifically, the expansion terms are extracted from uniform resource locators (URLs) of the documents returned by each database. The retrieval experiments with TREC 2013 FedWeb dataset demonstrates that the expanded query using the proposed approach performs better in many instances than the unexpanded query.

Keywords Distributed Information Retrieval · Federated Search · Results Merging · Query Expansion · Uncooperative Environments

All authors contributed equally to this work.

✉ Shah Khalid
shah.khalid@seecs.edu.pk

Adamu Garba
yakasai6@yahoo.com

Irfan Ullah
irfan@sbbu.edu.pk

[1] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

[2] Department of Computing, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan

[3] Department of Computer Science, Shaheed Benazir Bhutto University, Upper Dir, Sheringal 18050KP, Pakistan

1 Introduction

Due to the vast number of specialized databases and information repositories on the Web, web users often rely on traditional search engines to address their information needs [39]. However, the sheer size of the web makes indexing all its contents almost impossible [23]. Moreover, a considerable among of content is unreachable to web crawlers due to proprietary or commercial reasons [15], resulting in valuable information resources going undiscovered by traditional web search engines. These issues can be addressed in two ways: either each database offers a search user interface (UI) or a single UI is designed that serves unified access to all databases collectively. The latter is referred to as federated search [15, 27, 42, 48], which deals with providing a unified searchable interface to access databases without necessarily indexing their contents. To achieve efficiency and effectiveness in the federated settings, a subset of the databases considered to be relevant for a given user query is selected and searched. The results returned by these databases are fused into a single ranking list and presented to the user [42]. The decades of research carried out on the core components of the federated search which include: resource representation [3], resource selection [4, 19, 27], and results merging [16, 20]. The results of the research led to the development of a variety of federated search systems on the Web, including California Digital Library (CDL)[1] and EEXCESS.[2]However, satisfying users' information needs by any information retrieval (IR) system depends on how the users formulate their requests in the form of a search query. Sometimes, a large number of relevant documents are not matched to the user's query due to a vocabulary mismatch and the level of this mismatch widens as users use fewer terms to express their information needs [13].

Several techniques have been proposed to solve the mismatch problem. These include document representation, query reformulation, weighting/ranking schemes[47] and QE [25]. These techniques were successful in increasing the retrieval effectiveness of centralized search settings. However, the same level of success has not been reported in the literature on federated search. There can be several reasons. First, unlike centralized settings that control their corpus index, in a federated search environment, the central broker has little or no control over the indexes of the databases. Second, since the databases are autonomous, they may employ different strategies for document processing and retrieval. Finally, the size and, in some instances, the content composition of the databases vary (i.e., some are text-only, and others include text, images, and videos). These issues make query reformulation and expansion challenging in finding the right source to select the expansion terms from in a federated environment.

Nevertheless, some prior studies that investigated QE in federated search have used either feedback documents, lexical dictionaries, or large vocabulary corpora to select expansion terms from. For example, the studies [29, 41] selected the expansion terms to augment the query from the feedback documents. However, they reported a decrease in the performance of most of their expanded query results compared to the unexpanded ones. The performance decrease is caused by the expansion terms being selected from top-ranked feedback documents, which are mostly from the same database. Thus, the expanded query could not be generalized across all the selected databases. Another study [31] selected the expansion terms from external large vocabulary corpora. Regrettably, there is

[1] One of the biggest online US libraries to catalog educational resources, available at: https://cdlib.org/

[2] Hosted by the European Union, offers federated search to literary works and heritage collections, available at http://eexcess.eu/

no significant improvement in the performance of the expanded queries compared to the unexpanded ones due to topic drift.

The results reported from the aforementioned studies motivate us to explore other sources to select the expansion terms from, which may not cause either topic drift or be biased to some databases. As such, this work exploits the URLs that appear in the result snippets of each database as a source for selecting the expansion terms. Our aim in this paper is to find out: (i) Are the terms in documents' URLs suitable to select expansion terms from in a federated environment? (ii) To what extent does QE improve the retrieval effectiveness of federated search? To achieve these aims, we make the following contributions:

1. We propose the use of URLs as the source to select the expansion terms for QE in federated search.
2. Considering the nature of the URL, we propose a robust term weighting function that selects the most appropriate terms for QE.
3. We conducted an extensive set of experiments using the TREC 2013 FedWeb dataset. The experimental results show that in some instances QE improves the retrieval effectiveness of federated search.

The rest of the paper unfolds as follows: We discuss related works on federated search and QE in Section 2. The proposed approach, along with the experimental setup, is described in Section 3. The results and discussions are presented in Section 4. We conclude the paper and present some future directions in Section 5 followed by references.

2 Related works

Both federated search [42] and QE [41] have been studied in depth for the last three decades. In this section, we summarize the most relevant previous works and highlight their similarity and differences with the proposed study.

2.1 Federated search

Federated search, also called distributed information retrieval (DIR), deals with the unification of a search interface for concurrent searches across multiple databases [14]. In the federated settings, the central broker mediates between the users of the search systems and the databases [42]. As such, the broker receives the user query, selects a few relevant databases to search, and merges the results returned by them before presenting them to the users.

The prior work in federated search is categorized based on two environments, that is, cooperative and uncooperative [42]. The environment in which databases provide the central broker with their metadata information using standardized protocols is called a cooperative environment. Early models like START [18] and SDLIP [30] proposed protocols for this environment. However, in the real-world web environment, most databases are uncooperative in nature. As they only respond to broker queries, without providing information regarding their corpus indexes. As such, in this environment, the broker obtains the corpus statistics of the databases using query-based sampling [3] or other variants like adaptive query-based sampling [2]. The documents sampled by the broker from the databases are indexed in the centralized sample database, which serves as an agglomeration index of all the sampled documents. This index helps the broker determine which databases are most relevant to a particular user query. Additionally, it uses the sampled index documents to estimate the merging scores for documents returned by the databases. The uncooperative environment assumption has been used by several studies [10, 27, 48] due to its commonality with real-world federated search systems.

Recently, a snippet-based result merging for the federated search was proposed in [15]. Their model uses only the information the databases provide at query time to merge the multiple result lists. Furthermore, a federated search system that targets sports-related websites was discussed in [7]. The system was developed by creating four separate indexes in which each index contains one of these lists: list of competitions, names of the teams, names of the managers, and names of the players. For a given user query, the query is delimited into terms and each term is sent to an appropriate index. The returned results are merged into a single ranking list. Similarly, a knowledge-based method for resource selection in federated search is proposed in [19]. A learning to rank method that extracts multiscale features such as terms matching, central sample index, and topic features for resource selection was proposed in [50].

2.2 Query expansion methods

In most search systems, users express their information needs using keywords as queries. Sometimes a problem arises when users construct their query with terms different from the ones the search systems have indexed their documents with. This situation is often referred to as a vocabulary or terminology-mismatch problem [12, 46] and it makes retrieving relevant documents challenging. This terminology mismatch is often mitigated by adding an additional set of terms to the initial query, a process generally known as query expansion. The premise is that the set of documents this new query would retrieve is more likely to satisfy the user's information need. For decades, researchers have proposed a variety of sources to select expansion terms from that will not result in topic drift. The most widely used among them include pseudo-relevant feedback (PRF) documents [24, 44], external sources such as lexical dictionary [17], trained vocabulary corpus using word embedding techniques [9, 38], and query log [6].

The underlying assumption of the PRF approach is that most of the top-ranked documents returned by the initial user's query are relevant [22]. Therefore, the most frequent terms in those documents are considered to be a good set of candidate expansion terms. For each candidate term, its weight is computed based on its frequency in the top-ranked documents. Next, the terms are ranked based on their weights, and then the top-ranked ones are selected as expansion terms. Finally, the selected terms are added to the initial query, and the new query is used to retrieve the final set of documents to show to the users. However, if the top retrieved documents contain many non-relevant ones, the terms extracted from them would be unable to improve retrieval performance. To address this problem, Keikha et al. [21] used Wikipedia as a source for selecting the expansion terms. In [33], the score distribution was used to automatically select the right number of PRF documents that are more likely to have a good set of expansion terms for each of the given queries.

Alternatively, the expansion terms can also be selected directly from the lexical dictionaries. The commonly used dictionary to select the expansion terms in the literature is WordNet [1, 45], which is a built-in external conceptual dictionary that groups words into synonyms (i.e., synsets) together with the semantic relationships among them. With different methodologies, several studies [1, 17, 32] used WordNet for QE. For instance, Azad and Deepak [1] selected the expansion terms based on their similarity scores with the query obtained from WordNet and their occurrences in the Wikipedia articles. Another study [32] considered the use of the expansion terms from both WordNet and PRF documents.

Recently, word embedding has received unprecedented attention from researchers due to its ability to capture terms' similarity in low-dimensional vector space compared to the corpus vocabulary size. Word2Vec [28] and Glove [34] use a neural network to learn the vector representation of terms together with their synthetic and semantic similarities in large corpora. The success achieved by these word embedding techniques has opened a new paradigm in natural language processing and IR-related tasks. The models proposed in [11, 36] use various word embedding techniques for automatic QE. In summary, the aforementioned studies reviewed in this section attempted to address the vocabulary mismatch problem in a centralized setting. The next section discusses QE approaches in federated search.

2.3 Query expansion in federated search

The first known study on QE in the distributed environment [51] assumed that either the broker has access to the documents' index of the databases or can sample the documents in proportion to the size of the databases. Neither of the above assumptions is likely to be achievable in real-world scenarios as most of the databases are uncooperative, therefore, accessing their documents index or knowing their sizes in advance would be infeasible [29]. Based on this observation, Ogilvie and Callan [29] studied QE by sampling an even number of documents from the databases to set up the centralized sample index and selecting the expansion terms from them. They issued the same expanded query to all the selected databases, an approach generally known as the global query approach. Surprisingly, there was no significant performance difference between the results of expanded queries compared to the unexpanded ones. Several factors may contribute to the poor performance of the expanded queries' results. The major ones include selecting the expansion terms from the central database documents that have heterogeneous content compared to the individual databases and issuing the same query (i.e., global query) to all the databases.

Shokouhi et al. [41] postulated that the global query approach is only one of the multiple ways to execute QE in the federated environment. Other ways include local, fused, and cluster approaches [41]. In the local approach, the expansion terms for each database are selected from the documents sampled from that database. In the cluster approach, the expansion terms are selected from the documents of the same cluster, which are clustered based on their database similarity. In the former case, the databases receive different queries, while in the latter case, databases in the same cluster receive the same query. The local approach in which different queries are issued to the databases performed worse than the unexpanded query.

Palakodety and Callan [31] departs from using sampled documents as a source of the expansion terms. Rather, they selected the expansion terms from a trained vocabulary corpus. More specifically, they trained the Google News corpus, which contains almost 100 billion tokens with 3 million words, and then selected the expansion terms from the trained corpus. They reported marginal result improvement for the expanded query compared to the unexpanded ones due to topic drift. Nowadays, learning object repositories are very popular on the web because they enable the sharing and re-use of educational materials [35]. Koutsomitropoulos et al. [26] proposed a federated search mechanism that integrates learning objects repositories and optimizes user experience using QE.

To summarize, some conclusions drawn include: (i) In most cases, the expansion terms selected from feedback documents are unable to improve the retrieval effectiveness of the expanded query results. (ii) The selection of expansion terms from large external vocabulary corpora leads to topic drift. To address these issues, we propose selecting expansion terms from the URLs of documents. In line with this proposal, the next section presents the methodology of this research work.

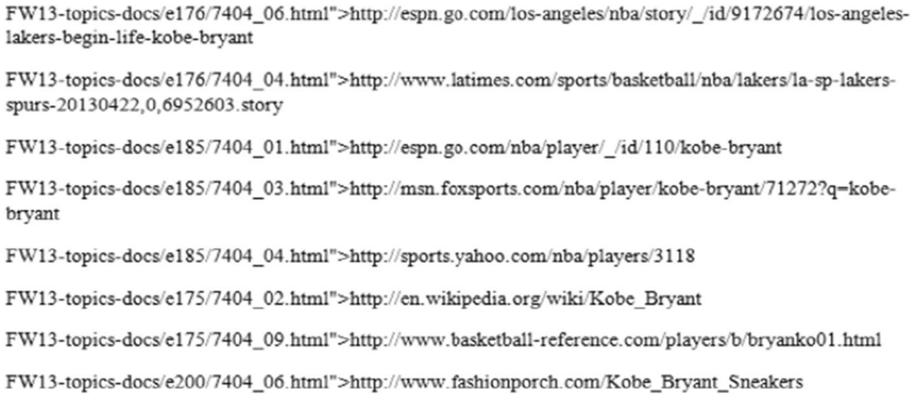
3 Materials and methods

In a federated setting, the query can be expanded either pre-retrieval (i.e., before being issued to the databases) or post-retrieval (i.e., based on the PRF). We consider the latter approach since the expansion terms are selected from the returned documents' URLs.

3.1 Proposed method

As previously mentioned, studies that selected expansion terms from external vocabulary corpora and pseudo-relevance documents ended up with poor results on most of the expanded queries compared to the unexpanded ones. For this reason, this work proposes using the documents' URLs as the source to select expansion terms. The benefit of our approach is that the size of candidate expansion terms and the possibility of topic drift have been drastically curtailed. Furthermore, selecting expansion terms from the documents' URLs minimizes the chances of the terms being biased towards a particular database. As in most cases, there is an overlap of similar terms in the documents' URLs irrespective of the database that returns the documents. The overlap terms are therefore more likely to reflect the content of the documents, so we consider them discriminatory enough to be candidate expansion terms.

Figure 1 shows the URLs of some documents for query no. 7404 "kobe bryant" as provided in the TREC 2013 FedWeb dataset. It can be observed that terms such as "nba," "kobe," and "players" appear in almost every document's URL returned by the different databases. When these terms are used as expansion terms to augment the query, the expanded query is more likely to return high-relevant documents while avoiding topic drift.



FW13-topics-docs/e176/7404_06.html">http://espn.go.com/los-angeles/nba/story/_/id/9172674/los-angeles-lakers-begin-life-kobe-bryant

FW13-topics-docs/e176/7404_04.html">http://www.latimes.com/sports/basketball/nba/lakers/la-sp-lakers-spurs-20130422,0,6952603.story

FW13-topics-docs/e185/7404_01.html">http://espn.go.com/nba/player/_/id/110/kobe-bryant

FW13-topics-docs/e185/7404_03.html">http://msn.foxsports.com/nba/player/kobe-bryant/71272?q=kobe-bryant

FW13-topics-docs/e185/7404_04.html">http://sports.yahoo.com/nba/players/3118

FW13-topics-docs/e175/7404_02.html">http://en.wikipedia.org/wiki/Kobe_Bryant

FW13-topics-docs/e175/7404_09.html">http://www.basketball-reference.com/players/b/bryanko01.html

FW13-topics-docs/e200/7404_06.html">http://www.fashionporch.com/Kobe_Bryant_Sneakers

Fig. 1 Some of the URLs for query No 7404 as provided in the TREC 2013 FedWeb dataset

Figure 2 and Algorithm 1 illustrate the steps involved in selecting the expansion terms from the documents' URLs. Suppose that a user issues a query to the broker, and the broker selects N number of relevant databases and routes the query to them. Each database processes the query and returns an initial ranked result list. It is assumed that the result list holds snippets of documents since that is what users find in the result list of real-world search engines. These returned result lists are combined into a single ranking list. Next the URLs of the top n documents are identified and extracted from this ranking list. For each URL, we first remove "https" and the domain extensions (i.e., ".com," ".org," ".uk," etc.). Then we use the URL conventional separators (i.e., a hyphen, slash, underscore, etc.) to split the remaining characters into tokens. Further, we remove all special characters, stop words, alphanumeric, and numbers. Finally, we sort the remaining terms and rank them based on their frequency. Although term frequency (TF) has been recognized as one of the best and most straightforward ways to quantify the importance of a term in a document, its raw use in federated settings might not be discriminative enough because multiple databases return documents in federated settings. Therefore, the importance of each term is quantified based on its score obtained using the proposed term weight functions, as given in Eq. (1), where $S(t)$ is the rank score of term t, n is the total number of documents' URLs that term t occurs in, $tf(t, u_i)$ is the frequency of t in those URLs, N is the total number of the databases that t occurs in.

$$S(t) = N \sum_{i=1}^{n} tf(t, u_i) \tag{1}$$

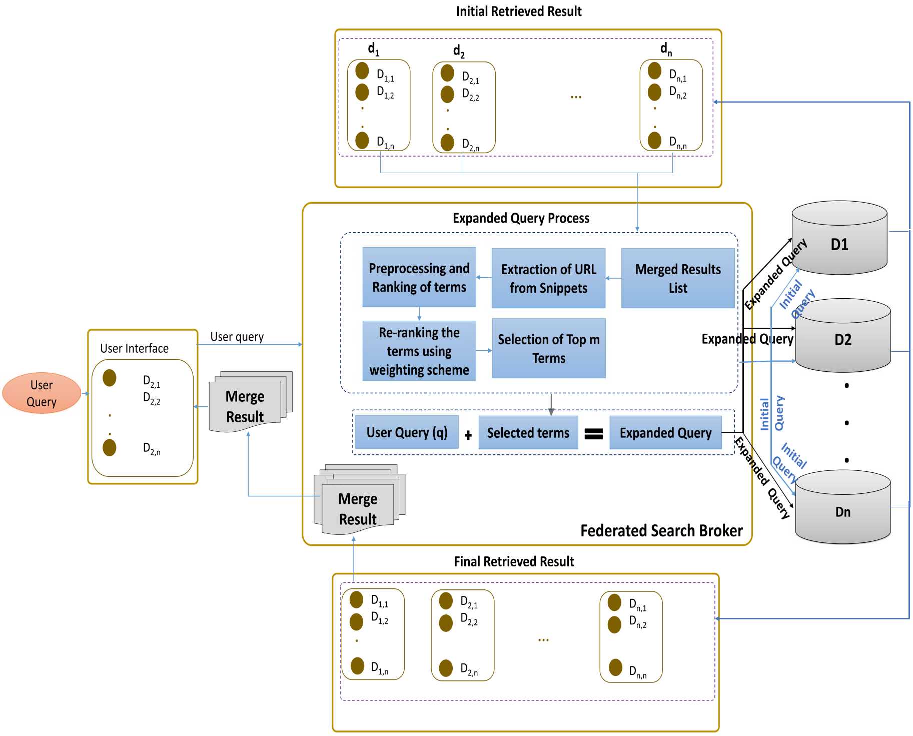


Fig. 2 A schematic view of the working of the proposed method

The multiplication by N boosts the score of those terms that occur in the document URLs of many databases. Based on this new score, the terms are re-ranked, and the top m terms are selected and added to the query. This expanded query is routed to the databases again. The databases process the expanded query and return a final ranked result lists. This final returned result lists are merged into a single ranking list and presented to the user.

Algorithm 1 QE Based on Documents' URLs for Federated Search

```
Input: User query, sets of databases
Output: Merged result list
User enters query to the search user interface.
procedure FederatedSearch($query$, $databases$)
    The broker routes the received query to the most relevant databases.
    Each database processes the query and returns a ranked results list.
    The returned results are merged using Equation (1).
    Equation (1) weights the processed terms extracted from the URLs.
    The user query is expanded with the top-weighted terms.
    The expanded query is routed to the relevant databases again.
    The returned result is merged using Equation (2).
    The merged result is returned to the user.
end procedure
```

3.2 The results merging score

Previous studies on federated search reported that the process of merging multiple result lists impacted QE performance. As such, we estimate the merging score based on the information the databases provide in their returned results. Therefore, Eq. (2) is used to compute the merging score of the returned list documents. Here, r is the rank of the documents in the database result list, n is the total number of documents in the database ranked list, and $s_i(s)$ is the snippet score of the document returned by a retrieval model. We use BM25 as the retrieval model.

$$score(s) = \exp - (r/n) \times s_i(s) \tag{2}$$

3.3 Advantages of the proposed method

As mentioned in Section 2.3, most of the results of the previous studies that investigated QE in the federated environment showed little or no benefit of QE, largely due to challenges in selecting appropriate expansion terms. In contrast, the proposed method uses documents' URLs as a source of expansion terms, which neither overwhelms the broker nor causes topic drift. This is due to the fact that the proposed method selects the expansion terms from the URLs, which all the documents have irrespective of their content (i.e., text, video, etc.). Selecting expansion terms from URLs reduces the chances of topic drift and the number of candidate expansion terms. In contrast, most existing approaches select expansion terms from feedback documents, which prevents non-textual documents from contributing toward QE. In addition, our experimental results are promising, especially for a single database mode (see Section 4).

The proposed method uses the sampled snippets instead of the full-text documents in the experiments, as using the full-text documents rules out many relevant databases beforehand. For example, consider the contents of the two databases in the dataset (i.e., e022 and e122)[3]; e022 has video contents while e122 has video and images. These databases have no full-text documents. Therefore, a strategy that uses full-text documents, will rule out those databases that have no full text, regardless of how relevant they are to a given query. Moreover, the study in [49] experimented with sampled snippets and full-text documents and reported that sampled snippets performed better than full-text documents.

3.4 Experimental setup

For the experiment, we used the sampled search engines' snippets provided in the TREC 2013 FedWeb dataset [8] as the databases. This dataset is the first standard corpus created to promote the research on federated search. It also aims to discourage the artificial creation of databases using TREC web track datasets. It holds the results downloaded from 157 real-world search engines in 24 vertical categories (i.e., academics, blogs, entertainment, jobs, kids, etc.). Each search engine uses its retrieval model to retrieve results. We use the 50 queries released with the dataset as search queries. The queries are judged using the five-level graded relevance judgments. These include not relevant (NRel), relevant (Rel), high relevant (HRel), top relevant (Key), and navigational (Nav) by the team of experts.

Two sets of experiments were conducted. In the first set, the goal is to determine whether URLs of documents are a suitable source to select expansion terms. Here, all snippets provided in the dataset are indexed in a single database repository. The second set of experiments aims to discover how QE enhances federated search retrieval performance. Here, we select the top three and five most relevant databases for the given queries using the modified resource selection algorithm proposed in [43]. In both sets of experiments, the snippets are indexed and retrieved using Apache Solr version 8.2 with default values for parameters and keeping BM25 [37] as the retrieval model. Since the snippets are indexed, we queried the title and description fields and aggregated the two scores as the relevant score of the documents. For pre-processing the URLS, we used the natural language processing toolkit (NLTK)[4] with Python as the programming language. All the experiments are carried out on an Intel Core i7 processor and 8 GB memory. As this is the first work that uses the documents' URLs as a source to select expansion terms in federated search, its performance based on the number of terms selected to augment the query is examined thoroughly. As such, from the initial merged result list, top 5, 10, and 15, documents were selected for each given query, while the expansion terms are set to 2, 4, and 6.

Since the aim of this paper is to investigate the effect of QE on federated search by exploring the documents' URLs as a source to select the expansion terms from. Consequently, we only compared the performance of the results with the unexpanded query results. Previous studies [29, 31, 41] in the literature also used this comparison method. We

[3] The identities of the databases are concealed in generating their sampled snippets' results with e000 code in the dataset. Although information about small excerpts of the dataset is allowed to be displayed in an academic article, revealing the identity of the search engines is prohibited to avoid infringement of their rights.

[4] http://www.nltk.org/

evaluate the results with the official evaluation metric for the TREC 2013 FedWeb result merging task, which is NDCG@k [5]. The NDCG metric measures the goodness of the retrieved result list compared to the best/ideal ordering of the result list. We report the result of top k positions, and the value of k is set to 5, 10, 15, and 20.

4 Results and discussion

Table 1 shows the performance of the expanded queries against the unexpanded ones. The results show that documents' URLs are good candidates for QE. From the results, the highest performance is observed at NDCG@5, which shows a 43.3% improvement over the unexpanded query. While the lowest performance was observed at NDCG@20. Furthermore, from Table 1, it is evident that increasing the number of feedback documents and expansion terms has no effect on the performance of the already expanded query. It may be because a term could only be selected as an expansion term if it appears in the URLs of multiple documents from different databases. This makes the selected terms mostly generalize across most of the databases. For example, consider the sampled URLs provided for query No 7404 in Fig. 1. When we checked the top two terms for this query on a single database repository when the feedback documents are set to 5, the terms are *kobe* and *players*. When the feedback documents are increased to ten and that of expansion terms to four, the top four terms are "kobe," "bryant," "players," and "nba". And since only the title and description are queried in the experiments, increasing the number of feedback documents and expansion terms showed no effect on the performance of the expanded query. Consequently, in the remaining experimental results of this section, we report adding two expansion terms only when the feedback documents are set to five.

Now that we have observed the performance of the expanded query on a single database repository, let us turn our attention to see if it would maintain the same performance when searching the subset of the databases considered relevant for the given queries. Table 2 shows the performance of the expanded query when the top three and five databases are selected. The results showed mixed performance of the expanded query upon variations in the number of selected databases to search. Based on the results, it can be observed that the expanded queries performed poorly compared to the unexpanded ones while selecting the top three databases. At the same time, a marginal improvement in performance can be observed for expanded queries compared to the unexpanded ones when the top five databases are selected. Although we do not expect the expanded query to maintain the same level of performance as in Table 1, we expected it to maintain its positive performance.

To understand this sudden drop in the performance of the expanded query, we critically analyzed the content of the selected databases and the expanded terms. We discovered that one of the top three databases selected by the given search queries contains video

Table 1 The QE on a single database repository. The expansion terms are selected from documents' URLs

NDCG@	Unexpanded query	Top 5 documents, 2 expansion term	Top 10 documents, 6 expansion terms	Top 15 documents, 6 expansion terms
5	0.1226	0.1763 (+43.80%)	0.1763 (+43.80%)	0.1763 (+43.80%)
10	0.1358	0.1852 (+36.38%)	0.1852 (+36.38%)	0.1852 (+36.38%)
15	0.1434	0.1852 (+36.38%)	0.1852 (+36.38%)	0.1852 (+36.38%)
20	0.1515	0.1900 (+25.41%)	0.1900 (+25.41%)	0.1900 (+25.41%)

Table 2 The QE performance with three and five most relevant databases. The documents' URLs are used as a source for QE

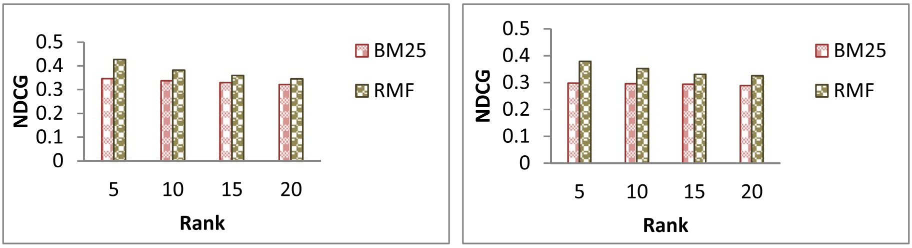
NDCG@	Top 3 Databases Selected		Top 5 Databases Selected	
	Unexpanded query	Expanded query	Unexpanded query	Expanded query
5	0.4268	0.4157 (-2.60%)	0.3787	0.3862 (+1.98%)
10	0.3819	0.3794 (-0.65%)	0.3522	0.3527 (+0.14%)
15	0.3599	0.3576 (-0.63%)	0.3308	0.3310 (+0.06%)
20	0.3448	0.3433 (-0.43%)	0.3252	0.3240 (-0.37%)

content. During the pre-processing phase, we discarded most of the content from document URLs returned by this database. The limited number of terms that remained were not among the top m. As a result, the expansion terms were selected from the remaining two databases. Unfortunately, the first database contained many relevant documents that had a higher initial rank with the unexpanded query, but a lower rank with the expanded query. These results show how challenging it is to select the expansion terms in a federated environment containing diverse databases. On the other hand, we can observe a partial marginal improvement of the expanded query results when selected from the top five relevant databases. Looking at the results in Table 2, we can observe that the performance of the expanded query increases with an increase in the number of databases selected to search.

Based on the results in Tables 1 and 2, we can answer our aims with these points: (i) Query expansion improves the retrieval effectiveness of federated search result merging when executed on a single database repository. (ii) With some exceptions, QE also improves when executed on the subset of the databases. (iii) In all the cases, the top documents' URLs are used as the source to select the expansion terms. The following subsections discuss the impact of the result merging method and database selection on federated search QE.

4.1 Impact of the results merging method

Although evaluating the performance of the proposed results merging method is not among our objectives, yet, we assess it to see its effectiveness against merging based on ranking scores produced by the retrieval model, i.e., BM25. Figure 3 shows the performance of the



(a)

(b)

Fig. 3 The performance of the proposed result merging method compared to the BM25 when (**a**) top three and (**b**) top five most relevant databases are selected

proposed results merging formula of the unexpanded query when the top three and five databases are selected. The results demonstrate that the proposed merging method is effective as using only the retrieval model score to merge the results drops the merged results' effectiveness on NDCG@5 by over 18.9% when the top three databases are selected. This drop in retrieval effectiveness is observed across the other ranks cut-off, as shown in Fig. 3(a) and (b).

4.2 Impact of database selection

The consensus in the federated search literature is to search a few databases considered relevant to the given queries. Searching all the databases increases latency and decreases performance as some may not be relevant to the given query. Although the expanded queries produce higher retrieval effectiveness on a single database repository than the unexpanded ones, as shown in Table 1, the overall effectiveness is shallow compared to the results in Table 2. Even in Table 2, the higher effectiveness is observed when the top three databases are selected than the top five. From these results, it can be concluded that the optimum result performance in federated search is obtained by selecting fewer relevant databases.

In summary, the experimental results show that the documents' URLs can be an excellent source to select expansion terms from, especially if we search short text like the documents' snippets. Increasing the number of feedback documents or the expansion terms neither aids nor hurts the performance of the first expanded query. However, when searching the subset of the databases considered relevant to the query, the expanded query shows some level of bias toward the documents of the databases that recommend the selected expansion terms. These findings show that exploring other sources for QE is desirable, even though the retrieval effectiveness of our expansion source is promising. One prominent reason is the dependence of the search effectiveness on the selection of sources for QE [32, 40].

5 Conclusions and future work

In this study, we investigated the effect of query expansion on the performance of the federated search. The main objective was to find how QE affects the retrieval effectiveness of the federated search. To achieve this objective, we designed a research question and attempted to answer it by proposing the use of the documents' URLs as a source to select the expansion terms. A series of experiments were conducted. The main findings can be summarized as follows:

- Query expansion significantly improves the effectiveness of the federated search on a single database repository. However, when subsets of the databases are searched, the expanded query produces mixed retrieval performance.
- The expanded query performed poorly when the top three databases were selected and showed marginal performance improvement with the top five databases in all the cases compared to unexpanded queries.
- The number of feedback documents or expansion terms has a negligible effect on the performance of the expanded query.
- The documents' URLs make a good source that can be leveraged in selecting the expansion terms.

In a nutshell, some of our findings are consistent with the previous studies that the performance of the federated search can be improved by expanding the query with good expansion terms. However, selecting the expansion terms from a source that can generalize across all the participating databases remains a challenge. Since there are multiple data sources to select the expansion terms from, we explored the use of documents' URLs as our data source in this paper.

Even though some of the experimental results are promising, the data source used to select the expansion terms has not eliminated the bias occasionally experienced regarding query expansion in the federated search. Nevertheless, the study has provided an avenue that can be further extended by combining it with other sources of expansion terms. For example, using any word embedding techniques to train a data source like a query log and selecting the top terms similar to the ones extracted from the URLs as expansion terms is worth exploring as a future direction. Another direction is to select the expansion terms from both the documents' URLs and any lexical dictionary.

Declarations

Competing interests The authors explicitly declare that "No Competing Interests are at stake and there is No Conflict of Interest" with other people or organizations that could inappropriately influence or bias the content of the article.

References

1. Azad HK, Deepak A (2019) A new approach for query expansion using wikipedia and wordnet. Inf Sci 492:147–163. https://doi.org/10.1016/j.ins.2019.04.019
2. Baillie M, Azzopardi L, Crestani F (2006) Adaptive query-based sampling of distributed collections. In Proceedings of the 13th International Conference on String Processing and Information Retrieval, SPIRE'06, page 316-328, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/11880561_26
3. Callan J, Connell M (2001) Query-based sampling of text databases. ACM Trans Inf Syst 19(2):97–130. https://doi.org/10.1145/290941.290974
4. Callan J (2002) Distributed information retrieval. In Advances in information retrieval, Springer. 127–150. https://doi.org/10.1007/0-306-47019-5_5
5. Clarke CLA, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 659–666. https://doi.org/10.1145/1390334.1390446
6. Cui H, Wen J-R, Nie J-Y, Ma W-Y (2002) Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web. 325–332. https://doi.org/10.1145/511446.511489
7. Damas J, Devezas J, Nunes S (2022) Federated search using query log evidence. In Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings, pages 794–805. Springer. https://doi.org/10.1007/978-3-031-16474-3_64
8. Demeester T, Trieschnigg D, Nguyen D, Zhou K, Hiemstra D (2014) Overview of the trec 2014 federated web search track. Technical report, GHENT UNIV (BELGIUM)
9. Diaz F, Mitra B, Craswell N (2016) Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891
10. Dragoni M, Rexha A, Ziak H, Kern R (2017) A semantic federated search engine for domain-specific document retrieval. In Proceedings of the Symposium on Applied Computing, pp 303–308. https://doi.org/10.1145/3019612.3019833
11. Fernández-Reyes FC, Hermosillo-Valadez J, Montes-y-Gómez M (2018) A prospect-guided global query expansion strategy using word embeddings. Inf Process Manag 54(1):1–13. https://doi.org/10.1016/j.ipm.2017.09.001

12. Furnas GW, Landauer TK, Gomez LM, Dumais ST (1987) The vocabulary problem in human-system communication. Commun ACM 30(11):964–971. https://doi.org/10.1145/32206.32212
13. Gallant M, Isah H, Zulkernine F, Khan S (2019) Xu: an automated query expansion and optimization tool. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), vol 1. IEEE, Milwaukee, WI, pp 443–452. https://ieeexplore.ieee.org/document/8754179/
14. Garba A, Khalid S, Ullah I, Khusro S, Mumin D (2020) Embedding based learning for collection selection in federated search. Data Technol Appl 54(5). https://doi.org/10.1108/DTA-01-2019-0005
15. Garba A, Wu S (2023) Snippet-based result merging in federated search. J Inf Sci. 01655515221144864. https://doi.org/10.1177/01655515221144864
16. Ghansah B, Wu S, Ghansah N (2015) Rankboost-Based Result Merging. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. IEEE, Liverpool, UK, pp 907–914. https://ieeexplore.ieee.org/document/7363176/
17. Gong Z, Cheang CW, Hou UL (2005) Web query expansion by wordnet. In International Conference on Database and Expert Systems Applications, pp 166–175. Springer. https://doi.org/10.1007/11546924_17
18. Gravano L, Chang C-CK, Garcia-Molina H, Paepcke A (1997) Starts: Stanford proposal for internet meta-searching. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data. 207–218. https://doi.org/10.1145/253260.253299
19. Han B, Chen L, Tian X (2018) Knowledge based collection selection for distributed information retrieval. Inf Process Manage 54(1):116–128. https://doi.org/10.1016/j.ipm.2017.10.002
20. Hong D, Si L (2012) Mixture model with multiple centralized retrieval algorithms for result merging in federated search. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp 821–830. https://doi.org/10.1145/2348283.2348393
21. Keikha A, Ensan F, Bagheri E (2018) Query expansion using pseudo relevance feedback on wikipedia. J Intell 50(3):455–478. https://doi.org/10.1007/s10844-017-0466-3
22. Khalid S, Khusro S, Alam A, Wahid A (2023) BERT-embedding and citation network analysis based query expansion technique for scholarly search. arXiv preprint arXiv:2301.11069. https://doi.org/10.48550/arXiv.2301.11069
23. Khalid S, Khusro S, Ullah I (2018) Crawling ajax-based web applications: Evolution and state-of-the-art. Malays J Comput Sci 31(1):35–47. https://doi.org/10.22452/mjcs.vol31no1.3
24. Khalid S, Shengli Wu, Alam A, Ullah I (2021) Real-time feedback query expansion technique for supporting scholarly search using citation network analysis. J Inf Sci 47(1):3–15. https://doi.org/10.1177/0165551519863346
25. Khalid S, Shengli Wu (2020) Supporting scholarly search by query expansion and citation analysis. Eng Technol Appl Sci Res 10(4):6102–6108. https://doi.org/10.48084/etasr.3655
26. Koutsomitropoulos D, Solomou G, Kalou K (2017) Federated semantic search using terminological thesauri for learning object discovery. J Enterp Inf Manag 30(5):795–808. https://doi.org/10.1108/JEIM-06-2016-0116
27 Li L, Zhang Z, Wu S (2018) Lda-based resource selection for results diversification in federated search. In: Meng Xiaofeng, Li Ruixuan, Wang Kanliang, Niu Baoning, Wang Xin, Zhao Gansen (eds) Web Information Systems and Applications. Springer, Cham, pp 147–156. https://doi.org/10.1007/978-3-030-02934-0_14
28. Mikolov T, Chen K, Greg Corrado, and Jeffrey Dean (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
29. Ogilvie P, Callan J (2001) The effectiveness of query expansion for distributed information retrieval. In Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01, pp 183-190, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/502585.502617
30. Paepcke A, Brandriff R, Janee G, Larson R, Ludaescher B, Melnik S, Raghavan S (2000) Search middleware and the simple digital library interoperability protocol. D-Lib Magazine 6(3):5–8
31. Palakodety S, Callan J (2014) Query transformations for result merging. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science. https://apps.dtic.mil/sti/pdfs/ADA618630.pdf. Accessed 20 Nov 2021
32. Pal D, Mitra M, Datta K (2014) Improving query expansion using wordnet. J Am Soc Inf Sci 65(12):2469–2478. https://doi.org/10.1002/asi.23143
33 Parapar J, Presedo-Quindimil MA, Barreiro A (2014) Score distributions for pseudo relevance feedback. Inf Sci 273:171–181. https://doi.org/10.1016/j.ins.2014.03.034

34. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
35. Piedra N, Chicaiza J, Lpez J, Tovar E (2014) An architecture based on linked data technologies for the integration and reuse of oer in moocs context. Open Praxis 6(2):171–187
36. Rattinger A, Le Goff J, Guetl C (2018) Local word embeddings for query expansion based on co-authorship and citations. CEUR Workshop Proc 2080:46–53
37. Robertson SE, Walker S, Beaulieu M (2000) Experimentation as a way of life: Okapi at trec. Inf Process Manage 36(1):95–108. https://doi.org/10.1016/S0306-4573(99)00046-1
38. Roy D, Paul D, Mitra M, Garain U (2016) Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608
39. Sellami S, Zarour NE (2022) Keyword-based faceted search interface for knowledge graph construction and exploration. Int J Web Inf Syst 18(5/6):453–486. https://doi.org/10.1108/IJWIS-02-2022-0037
40. Sharma DK, Pamula R, Chauhan DS (2018) A comparative analysis of fuzzy logic based query expansion approaches for document retrieval. In International Conference on Advances in Computing and Data Sciences, pp 336–345. Springer. https://doi.org/10.1007/978-981-13-1813-9_34
41. Shokouhi M, Azzopardi L, Thomas P (2009) Effective query expansion for federated search. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, p 427-434. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1571941.1572015
42. Shokouhi M, Si L (2011) Federated search. Found. Trends Inf Retr 5(1):1–102. https://doi.org/10.1561/1500000010
43. Shokouhi M (2007) Central-rank-based collection selection in uncooperative distributed information retrieval. In European Conference on Information Retrieval, pp 160–172. Springer. https://doi.org/10.1007/978-3-540-71496-5_17
44. Singh J, Sharan A (2015) Context window based co-occurrence approach for improving feedback based query expansion in information retrieval. Int J Inf Retr Res (IJIRR) 5(4):31–45. https://doi.org/10.4018/IJIRR.2015100103
45. Singh J, Sharan A (2017) A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Comput Appl 28(9):2557–2580. https://doi.org/10.1007/s00521-016-2207-x
46. Ullah I, Khusro S (2020) Social book search: the impact of the social web on book retrieval and recommendation. Multimed Tools Appl 79(11–12):8011–8060. https://doi.org/10.1007/s11042-019-08591-0
47. Ullah I, Khusro S (2023) On the analysis and evaluation of information retrieval models for social book search. Multimed Tools Appl 82(5):6431–6478. https://doi.org/10.1007/s11042-022-13417-7
48. Urak G, Ziak H, Kern R (2018) Source selection of long tail sources for federated search in an uncooperative setting. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC '18, p 720-727. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3167132.3167212
49. Wang Q, Shi S, Cao W (2014) Ruc at trec 2014: Select resources using topic models. Technical report, RENMIN UNIV BEIJING (CHINA). http://trec.nist.gov/pubs/trec23/papers/pro-info ruc federated.pdf
50. Wu T, X Liu, Dong S (2019) Ltrrs: a learning to rank based algorithm for resource selection in distributed information retrieval. In Information Retrieval: 25th China Conference, CCIR 2019, Fuzhou, China, September 20–22, 2019, Proceedings 25, pp 52–63. Springer. https://doi.org/10.1007/978-3-030-31624-2-5
51. Xu J, Callan J (1998) Effective retrieval with distributed collections. In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pp 112–120. https://doi.org/10.1145/290941.290974

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