International Journal of Computer Applications (0975 - 8887)
Volume 180 - No.29, March 2018

# A Comparison Between Selective Collection Enrichment and Results Merging in Patient Centered Health Information Retrieval

Edwin Thuma
University of Botswana
Gaborone, Botswana

Onneile G. Tibi
University of Botswana
Gaborone, Botswana

Gontlafetse Mosweunyane
University of Botswana
Gaborone, Botswana

## ABSTRACT

In this article, an empirical investigation was conducted to determine whether merging search results generated by multiple query variants with the same information need can improve the retrieval performance in patient centered health information retrieval. In addition, this approach was compared with the selective collection enrichment approach, where only the results generated by a single query, which was predicted to perform better on the local collection is used. Three different results merging strategies predominantly used in distributed search environments with large overlapping databases were used in this study. The results of this investigation suggests that merging results using multiple query variants with the same information need can improve the retrieval performance. Also it was observed that the choice of an external collection used in generating these query variants can have an impact in the retrieval performance as it can sometimes lead to a degradation in the retrieval performance. When a comparison was made between results merging strategies and the selective collection enrichment approach, it was observed that the selective collection enrichment approach ranks fewer and highly relevant documents in the top 10 retrieved documents while the results merging strategies ranks more and slightly relevant documents in the top 10 retrieved documents.

### General Terms

Query Expansion, Collection Enrichment

### Keywords

Query Variants, Distributed Information Retrieval, Results Merging

## 1. INTRODUCTION

In the traditional Information Retrieval (IR) system, a query is posted to the retrieval system to satisfy an information need. To do so, searchers need to translate their information need into an explicit query that is submitted to the search system. With the ever increasing volume and complexity of health information on the web, recent research has discovered that very often in a single search session, a single user may issue multiple queries for the same information need (hence creating query variants) [1, 2, 3]. These query variants in turn may achieve different retrieval results. To address this, several investigators have explored improving retrieval effectiveness by deploying different retrieval strategies and different query representation. Extensive research exists on improving retrieval effectiveness of patient centered search systems through techniques such as Query Expansion (QE) [4, 5, 6]. However, research has shown that although QE has been proven to be an effective technique for IR, in some cases it can lead to little or poor retrieval performance [7]. Moreover, research has shown that queries expanded from different resources may yield very different top-ranked results when issued to a search engine, and these results may vary in quality and relevance to a given query [1, 8]. In addition, although extensive research exists on improving retrieval effectiveness, little concentration has been directed towards building search systems that are robust to query variations. Therefore, the aim of this article is to investigate the effect on retrieval performance when rather than selecting a single expanded query to retrieve on the local collection, query variants (expanded queries from different resources with the same information need) are used to retrieve on the local collection and their results merged into a single list of top ranked documents. Previous research in distributed information retrieval [9, 10, 11] has shown that the result merging sub-process is vital to the overall effectiveness of the retrieval process, especially in precision oriented environments where users expect a significant number of relevant documents in the top ranks of the returned document list. Even if the most appropriate information sources have been chosen in the resource selection phase, if the result merging is not effective, the overall quality of the retrieval process will deteriorate [11]. This importance is augmented particularly in the web environment where users rarely look past the top 20 results and most often do not browse after the top 5 results as observed. For that reason, three different results merging strategies that have show to perform well in large overlapping databases are deployed in this study. The rationale for using these stategies is that a single resource is searched using various query variants with the same information need and this scenario emulates a federated search environment where large overlapping databases are searched with the same query.

The remainder of this article is organized as follows: Section 2 covers brief background on results merging methods, query expansion, selective collection enrichments and literature review of relevant existing research in patient centered health information retrieval. Section 3 presents methodology. In Section 4 , a description of the experimental setup is provided. Section 5 presents the results and analysis, and finally Section 6 provides conclusions and future work.

## 2. BACKGROUND AND RELATED WORK

### 2.1 Background

In this article, three different results merging methods that have shown to perform better on large overlapping databases are used to merge search results generated by multiple query variants from a single collection. These methods includes Round Robin, Shadow Document Method and Multiple-Evidence Method, which are described in the following sections. In addition, a description of query expansion, which is used to generate the multiple query variants is provided. Moreover, a description of the selective collection enrichment approach is provided, which is used in the baseline system.

#### *2.1.1 Round Robin (RR).*

Round-robin merges the search results by taking one result from each of the input results lists [12]. Round Robin merge activity takes the search result lists as input and creates a new list to hold the results of the round-robin merge. Input search result lists are then traversed such that one result is selected from each list at a time and added to the newly created results. A Round Robin merging method is defined as follows: given n result lists $L_1, L_2, ..., L_n$, take the first result $r_1$ from each list $L_i$ as the first n results. Then take the second result $r_2$ from each list as the next n results, and so on. Round Robin merging produces a list: $L_1r_1, L_2r_1, , L_nr_1, L_1r_2, L_2r_2, L_nr_2, L_1r_3, L_2r_3, L_nr_3$,etc. The Round Robin algorithm in this article is deployed in the fashion presented in the flow-chat in figure 2.

#### *2.1.2 Shadow Document Method (SDM).*

In SDM for results merging, the document scores returned by multiple resource collections are normalized by a regression function that compares the scores of overlapped documents between the returned ranked [13]. The mode of operation of SDM is based on two assumptions. For a given query $q$, and resource collections A and B:

(1) If a document $d$ is contained in the results of resource collections A and B, the score values of $d$ from both databases are summed as $d$s global score.

(2) If $d$ is only retrieved from resource collections A and not from B, it is assumed that a shadow document of $d$ exists in B with a score value of $k$. Then the score of A and B are summed up.

In a nutshell, if a document was present in more than one result, then all the scores of that document were added up to get the global score. In the case that the document was present in only one result, then the documents score was added to the coefficient value. The value of $k$ has to be determined through empirical tests or can be derived from the degree of overlap between the two resource collections [13]. CombSUM is used to merge all the results. Equation 1 shows the possible ways to calculate the shadow score of a document:

$$score(d) = \sum_{i=1}^{m} s_i + k\frac{n-m}{m}\sum_{i=1}^{m} s_i \qquad (1)$$

Where, $k$ is a weighting coefficient, which lies between 0 and 1. In this case $k = 0.5$. $d$ is the document retrieved by any one of the resources and $s_i$ is the score of the document in the $ith$ result file. The number of result files available is denoted by $n$ and $m$ denotes the number of result files in which the document $d$ occurred. If the document $d$ occurred in $m$ result files out of $n$ result files with a score $s_i(d)$(where $1i \leq m \leq n$) then the global score of d was calculated using the Equation 1. In the SDM results merging experiments, the resulting merged file was sorted in descending order of their scores and re-ranked 1-1000. The documents with rank greater than 1000 were discarded from the results.

#### *2.1.3 Multiple-Evidence Method (MEM).*

A simple way of deploying the MEM is to average the score of every document and multiply the score by a factor $f(i)$ [14]. The factor $f(i)$ is the function of the number of the participating collections that return the document $d$. The value of $f(i)$ indicates the evidence about the relevancy of the document to the given query. The combined score obtained varies according to the $f(i)$ factor. The value of $f(i)$ increases with the value of $i$ . Therefore, it is important to state how $f(i)$ is determined. For example, if document $d_1$ is retrieved by collection A with a normalized score of 0.2, and document $d_1$ is also retrieved by collection B with a normalized score of 0.3, and retrieved by collection C with a normalized score of 0.1 then:

(1) if we let $f(i) = i for (i = 1, 2, 3, \ldots)$, then we have CombSUM.

(2) if we let$f(i) = i^2, for (i = 1, 2, 3, \ldots)$, then we have CombMNZ [13, 15].

Lee [15] found that when experimenting with $f(i) = avg\_score * i^{\beta}$, where $\beta$ have different values (1, 1.4, 2, 3, 6 and 11) assigned to f(i), the best scores was when $\beta = 2$. In the same manner, we experiment with $f(i) = i^2$ .

#### *2.1.4 Query Expansion (QE).*

Query Expansion is the term given when a search engine add terms to a users original search query [16]. The goal of query expansion is to improve precision and/or recall. Queries submitted by users are usually very short. Therefore, QE expands the original user query with other words that best capture the actual user intent, or that simply produce a more useful query, a query that is more likely to retrieve relevant documents.
The most common way to do query expansion is by using some form of a thesaurus. For each term t in the query, the thesaurus can be used to automatically expand the query using synonyms or other related words. There are several methods for building a thesaurus for query expansion.

(1) Maintaining a controlled vocabulary. For each concept, there is an assigned canonical term.

(2) Manually constructing a thesaurus, where human editors have assigned different names for concepts without any canonical terms.

(3) Automatically deriving a thesaurus, where word co-occurrence statistics over a collection is used to automatically build a thesaurus.

(4) In case of web search, using the query log where query reformulation from other users are utilized to make suggestion to new users. This requires a large volume of queries, which is why this approach is more suitable for web based systems. Expansion terms (i.e. the terms that are added to the original query) generally are selected from 3 types of resources: local resource, global resource or external resource.

To expand a query with a local resource, candidate expansion terms are selected from a set of documents retrieved in response to the original (unexpanded) query. Ideally, expansion terms should be drawn from some initially retrieved relevant documents. Since

these documents are relevant, terms present in these documents are expected to be related to the query, and should help to retrieve other similar documents which are also likely to be relevant. However, when Global resource is used for query expansion, expansion terms are selected from the entire collection of documents. Candidate terms are usually identified by mining term-to-term relationships from the target corpus. When an external resource is used, query expansion terms are obtained from other resources besides the target corpus. These resources may include other document corpora (including the Web), linguistic resources like Wordnet4, user-query logs etc.

*2.1.5 Selective Query Expansion (SCE).*
As mentioned in Section *2.1.4*, QE has been proven to be an effective technique for IR, however, in some cases can lead to little or poor retrieval performance [7]. The effectiveness of query expansion on retrieval is correlated with the quality of the top-ranked documents returned. Therefore, it is important to know in advance if QE will improve the retrieval performance, or on the contrary, degrade the retrieval performance. The selective query expansion approach has been studied as a possible mechanism to enable decision making for this. The basic idea of selective query expansion is to decide whether QE should be applied or not, such that QE could be disabled if the query is predicted to perform poorly. A selective QE mechanism was proposed by Amati et al. [7]. In the context of the Divergence from Randomness (DFR) framework. Their model predicts the performance of QE by assessing the query difficulty. This method looks at the divergence of the query terms distribution in the top-ranked documents from this distribution in the whole collection. The intuition is that as the query term distribution diverges further away from the distribution of the whole collection, it would be beneficial to deploy query expansion as it is likely to yield better results. In their selective QE approach, Amati et al. [7]. deployed InfoQ, an information theoretic function that combines query length and Inverse Document Frequency (IDF). Amati et al. [7]. reported an improvement in Mean Average Precision (MAP) when this selective QE method is deployed. Another selective QE approach was proposed by Cronen-Townsend et al. [17] in the context of language modelling. In their approach, they used the comparison between language models of the unexpanded and the expanded retrieval results to predict when the expanded retrieval results have strayed from the original sense of the query. In these cases, the unexpanded retrieved results are used, while the expanded results are used in the remaining cases (where such straying is not detected). Clarity score was used to classify queries in two groups: the ones which would benefit from QE and the ones that would not. Kwok et al. [18] studied the idea of using an external resource for QE. They suggested that the failure of QE is caused by the lack of relevant documents in the local collection. Therefore, the performance of QE can be improved by using a large external collection, which possibly contains more relevant documents and has better collection statistics for the query term reweighting. He et al. [19] suggested a method to select amongst several term-weighting models depending on the query. Queries are characterized by various features which are used to cluster them using agglomerative hierarchical clustering. Training associates the best term-weighting schema with each query cluster. After training, a new query is first clustered into the existing clusters and the pretrained system is used to process it. This is a pre-retrieval method since the query characteristics are calculated from the query itself and from the general characteristics of the document collection (query length, term idf, and clarity/ambiguity of the query). When evaluated on Robust TREC, the method slightly improves MAP when enough training queries are used. Another approach of SQE was proposed by Tibi et al. [3] where three different external resources are used to enrich a user query, thus generating three different expanded queries. Pre-retrieval query performance predictors are then deployed to select an expanded query that is most likely to perform better when retrieving on a local collection being searched. In their investigations, the effects of combining several pre-retrieval query performance predictors scores using data fusion techniques for Selective Collection Enrichment(SCE) were evaluated. Their empirical evaluation shows marked improvement in the retrieval performance in terms of nDCG@10 when the SCE approach was deployed.

## 2.2 Related Work

*2.2.1 Patient Centered Health Information Retrieval.*
Research in patient centered health information has been driven by the Conference and Labs Evaluation Forum (CLEF) since 2013. In particular, the CLEF 2013 eHealth Evaluation Lab shared task was focused on natural language processing (NLP) and IR for clinical care. The goal of the task was to investigate the effect of using additional information such as the discharge summaries and external resources such as medical ontologies on the IR effectiveness. The Lab aimed at evaluating systems that support laypeople in searching for and understanding their health information. Goeuriot, et al. [20] analyzed the approaches used by the participating teams. In their findings, CLEF 2013 participating teams results showed that combining BM25 with relevance feedback provided the strongest baseline. The combination of query expansion methods and external resources seemed to be efficient and produced better results [20]. This approach was used by Zhu et al. [4], who produced the overall best results. In their approach, Zhu et al. [4] used external resources for query expansion, as well as re-ranking based on concepts from the query and the discharge summary reports. The use of concept re-ranking provided better results for the top three best teams runs as well [4, 21, 22]. However, the overall results showed that research still needs to be conducted to make the best out of the external resources.
The CLEF 2014 eHealth Evaluation lab focused on facilitating understanding of information in narrative clinical reports, such as discharge summaries, by visualizing and interactively searching previous eHealth data. The goal of the lab was to also evaluate systems that support laypeople in searching for and understanding their health information. The results from CLEF 2014 showed a different approach by the participating teams. An analysis of the approaches deployed by participants at the CLEF 2014 task showed slight similarity with the approaches deployed in the previous year. The results showed that effective systems can be created using statistical language modelling techniques, along with query expansion mechanisms based on structured domain knowledge, and the exploitation of information from the discharge summaries [23]. Several state-of-the-art baselines were implemented for the CLEF 2014 eHealth task. The highest performance was achieved using language models with Dirichlet smoothing [24].
The main aim of the CLEF 2015 eHealth evaluation lab task 2 was to investigate the effectiveness of IR systems when searching for health-related information on the Web to answer queries posted by ordinary users who want to self-diagnose certain medical ailments. Unlike in the previous CLEF eHealth tasks (CLEF 2013 and 2014), in the CLEF 2015 eHealth task there was a shift in the query formulation focus. The queries depicted information needs that often arise before attending a medical professional appointment. Such queries often fail to deliver effective search results

because search engine users who try to self-diagnose often construct circumlocutory queries, using colloquial language instead of medical terms [25]. For example, using white part of eye turned green as a query instead of using the medical term jaundice. In the CLEF 2015 eHealth task 2, participating teams explored different methods of query expansion. The results show that query expansion had an important effect in improving search effectiveness [26]. This was evident in the approach used by team ECNU [27], which obtained the best results among all the participating teams. In particular, they obtained the highest retrieval performance when they used a query expansion method that mined expansion terms from Google top search results returned for the original queries. However, other participating teams explored learning to rank and other term weighting models for QE. The CLEF 2016 eHealth evaluation lab was a continuation from previous CLEF eHealth IR tasks that ran in 2013, 2014, and 2015. The CLEF 2016 task 3 investigated the effectiveness of IR systems when health consumers search the Web for health information. The 2016 task 3 as well aimed at fostering research and development of search engines to support health information seeking [28]. CLEF 2016 task 3 is similar to the CLEF 2015 task 2 with additions of considering also health information needs related to the treatment and management of health conditions. However, unlike in the previous years (2013, 2014 and 2015), the dataset used in CLEF 2016 was ClueWeb12 B13, which consist of over 50 million articles and is more representative of the current state of the content on the Web. It differs with the datasets used in previous CLEF eHealth tasks, which had about 1 million health-related Web pages provided by the Khremoi project [26, 29]. The participating teams in the CLEF 2016 task 3 deployed different query expansion strategies such as testing different term weighting models, for example TF-IDF, Dirichlet-smoothed language model, PL2 etc. [30, 31, 32], Web-based QE, document re-ranking, query re-formulation with medical terms and collection enrichment [29].

## 3. EVALUATION METHODOLOGY

To measure the effectiveness of the proposed retrieval approach, an empirical evaluation is carried out based on the Cranfield paradigm [33]. The prime motivation to use the Cranfield paradigm was the system-centric evaluation mode that is embodied in the Cranfield paradigm evaluation model. The Cranfield paradigm uses a test collection. The test collection supports the automated evaluation of relevance-based effectiveness, through its three components: a document collection from which the system attempts to satisfy the user information need; statements of information needs called queries each describing a different information need; and relevance assessments or judgments as to which retrieved documents are relevant to which topics.

### 3.1 Document Collection

Experiments were carried out using four corpuses; one as the local resource being searched and three as external resources for collection enrichment (expanding the original query to create three different query variants). The local collection is made of the 2016 CLEF eHealth Evaluation Lab corpus ClueWeb12 B13[1] (a collection of more than 50 million Web pages). The document collection was provided by the organizers of the CLEF eHealth 2016 Lab in the form of a standard index of this corpus built with the Terrier-4.2[2] [34] IR platform. The details of how the ClueWeb12 B13 document collection was generated can be found in Lidah et al. [28] and in Zuccon et al. [29]. This standard index had query expansion disabled. To perform collection enrichment, three different indexes (external resources) were used, which were also built using the Terrier-4.2 IR platform. These indexes had query expansion enabled and were built using the 2015 CLEF eHealth Evaluation Lab corpus, TREC Clinical Decision Support(CDS) track 2015 corpus and the English Wikipedia dump of 2008 corpus. The 2015 CLEF eHealth Evaluation Lab corpus was provided by the Khremoi project [26] for the CLEF 2015 eHealth Evaluation Labs conference. This corpus consists of about a million of crawled web pages obtained from predominantly popular health and medicine resources. These web pages consist of a broad range of health information that is likely to contain health topics in both laypeople and professionals vocabulary. The TREC 2015 track is a snapshot of the Open Access Subset of the PubMed Central (PMC). PMC is an online digital database that consists of free full-text biomedical articles [35]. The TREC 2015 corpus consists of 733 138 articles in the biomedical domain. The Wikipedia dump of the 2008 corpus is a snapshot of the English dump of October 2008 and consist of close to 818 741 general articles. Wikipedia is an enormous freely available resource of information. This resource is an encyclopedia that is collaboratively written and edited by its users. Therefore, Wikipedia is more likely to contain medical terms expressed in laypeople vocabulary.

### 3.2 Queries

The queries used in this study were provided in the 2016 CLEF eHealth Evaluation Lab test collection. Queries in this test collection explore real health consumer information needs posted on health Web forums. Query posts were extracted from the health forum askDocs of Reddit [29]. The details of how these queries were generated can be found in Zuccon et al. [29]. Table 1 provides a sample of queries used in this study. Queries 1-6 were created for post 101, i.e. queries 101001 101006, and query 102001 was created for post 102. Query identifier 1, 2 and 3 shows queries generated by expert query creators e.g. queries 101001-101002, while identifier 4, 5, and 6 represent queries created by laypeople query creators [29], e.g. queries 101004-101006.

Table 1. A SAMPLE OF QUERIES

| Query Identifier | Queries |
|---|---|
| 101001 | inguinal hernia repair laparoscopic mesh benefits risks |
| 101002 | inguinal hernia laparoscopic mesh surgery |
| 101003 | inguinal hernia success rate |
| 101004 | inguinal hernia surgery or surgical “complications” |
| 101005 | inguinal hernia laparoscopic with mesh surgery reviews |
| 101006 | inguinal hernia surgery story, is it safe? |
| 102001 | anal Tag removal options |

### 3.3 Relevance Assessment

The relevance assessments also referred as the relevance judgments are used to quantify the system effectiveness [33]. The document collection and the query set are used to run retrieval experiments,

[1] https://lemurproject.org/clueweb12/specs.php
[2] http://terrier.org

and produce systems outputs in a standard format called runs, which are then evaluated against the set relevance assessments for relevance, using a suitable evaluation measure. In this study, the query relevance judgments provided by the 2016 CLEF eHealth Evaluation Lab Conference organizers in conjunction with precision at rank K (P@K) and nDCG@K were used to evaluate the performance of the proposed retrieval strategy. Details of how the query relevance judgments were created can be found in Zuccon et al. [29, 36, 37]. The topical relevance judgments provided were graded with respect to the grades Highly relevant, Somewhat relevant and Not Relevant [29] on a 3-point scale: 0, "Not Relevant; 1, "Somewhat Relevant; 2, "Highly Relevant. The qrel file has the form,
query-number 0 document-id relevance
Each element in the qrel is delimited by the spaces where query-number is the number of the query, document-id is the external ID for the judged documents, 0 is a constant and relevance is the relevance grade assigned to the document for the particular query. Table 2 presents a sample of a few lines from the qrel file provided for the 2016 CLEF eHealth Evaluation Lab.

Table 2. A SAMPLE OF QREL FILE

| QRELS |
|---|
| 101001 0 clueweb12-0000tw-08-16795 0 |
| 101001 0 clueweb12-0000wb-06-29427 0 |
| 101001 0 clueweb12-0000wb-15-21867 0 |
| 101001 0 clueweb12-0000wb-20-07932 0 |
| 101001 0 clueweb12-0000wb-48-08896 0 |
| 101001 0 clueweb12-0000wb-54-11411 1 |
| 101001 0 clueweb12-0000wb-70-15174 0 |
| 101001 0 clueweb12-0000wb-84-05434 1 |
| 101001 0 clueweb12-0000wb-85-17525 0 |
| 101001 0 clueweb12-0000wb-97-30058 1 |

## 4. EXPERIMENTAL SETUP

### 4.1 Indexing and Retrieval Platform

Terrier-4.2[3] [34], an open source IR platform [34] was used in this study for indexing and retrieval. The ClueWeb12 B13 collection was first pre-processed before indexing, and this involved tokenizing the text and stemming each token using the full Porter stemming algorithm. Stopword removal was enabled during indexing using the Terrier-4.2 stopword list.

### 4.2 Query Variants Generation

Three different query variants with the same information need were generated for each of the queries described in Section 3.2 using the collection enrichment approach. In particular, for each external resource described in Section 3.1 (2015 CLEF eHealth Evaluation Lab corpus, TREC Clinical Decision Support(CDS) corpus and the Wikepedia dump of 2008 corpus), the Terrier-4.2 Divergence from Randomness (DFR) Bose-Einstein 1 (Bo1) model was used to select the 10 most informative terms from the top 3 returned documents as expansion terms. These 10 new terms together with the original query terms were combined together to form new query variants to be used for retrieval on the local collection.

[3]http://terrier.org

### 4.3 Baseline System

For comparison, baseline experiments were conducted on the local collection using the query variants generated in Section 4.2. In these baseline and subsequent experiments, the Terrier-4.2 PL2 term weighting model from the Divergence from Randomness (DFR) framework was used to rank the retrieved documents.

### 4.4 Query Variants Results Merging (QVRM)

Single queries expanded on multiple external resources generate query variants. When these query variants are used to retrieve on the local collection, three sets of results are generated, each with 1000 ranked documents. Three different result merging strategies which were described in Section 2 ( Round Robin, Shadow Document Method and Multi-Evidence Method ) are then deployed to investigate whether combining results generated by multiple query variants can improve the retrieval effectiveness. In addition, a comparison of the proposed approach was made to the selective collection enrichment approach deployed by Tibi et al. [3], where they deployed query difficulty predictors to selectively expand their queries rather than generating and merging results from multiple query variants.

## 5. RESULTS AND ANALYSIS

### 5.1 Baseline Result

As stated earlier in Section 4.3, a baseline system was created by deploying the PL2 term weighting model from the DFR framework to retrieve and rank the documents on the local collection (ClueWeb13 B13) using the various query variants. Table 3 and Table 4 presents the evaluation results at different cut-off levels. The highest values are emphasized with bold and the symbol ↑ emphasize retrieval increase over other query variants.

Table 3. BASELINE EVALUATION RESULTS, P@5 and nDCG@5

| External Collection | P@5 | nDCG@5 |
|---|---|---|
| **PL2 - No Expansion** | 0.3213 | 0.1832 |
| **CLEF 2015** | ↑**0.3380** | ↑ **0.1960** |
| **TREC 2015** | 0.3080 | 0.1781 |
| **WIKIPEDIA 2008** | 0.3207 | 0.1822 |

Table 4. BASELINE EVALUATION RESULTS, P@10 and nDCG@10

| External Collection | P@10 | nDCG@10 |
|---|---|---|
| **PL2 - No Expansion** | 0.3067 | 0.1824 |
| **CLEF 2015** | ↑**0.3380** | ↑ **0.1960** |
| **TREC 2015** | 0.2817 | 0.1664 |
| **WIKIPEDIA 2008** | 0.933 | 0.1713 |

The results in Table 3 and Table 4 shows that query variants generated after expanding the original queries with the CLEF 2015 external corpus improves the retrieval performance than when the original queries without expanding with new terms. This is evidenced by an increase in nDCG@5 from 0.1832 (no expansion) to 0.1960 (CLEF 2015 expansion). Similar results were observed when P@5, P@10 and nDCG@10 evaluation measures were used. However, when other external collections are used, there was a degradation in the retrieval performance for all evaluation measures. The question

raised by these baseline systems was whether there could be an improvement in the retrieval performance if the results generated by these multiple query variants are merged into a single result list.

### 5.2 Query Variants Results Merging (QVRM)

Table 5 and Table 6 present the evaluation results for the results merging methods; Round Robin, Shadow Document Method and Multi-Evidence Method after merging results generated by CLEF 2015, TREC2015 and WIKIPEDIA 2018 query variants. The results are presented for precision and nDCG evaluation measures both at cut-off levels 5 and 10 respectively. The evaluation results presented in these tables show retrieval improvement over the PL2 - No Expansion baseline, TREC 2015 baseline and WIKIPEDIA 2008 baseline. However, merging retrieval results from these multiple query variants resulted in the degradation in the retrieval performance when compared to the CLEF 2015 baseline across the different results merging methods. These results suggests that not all external collection are suitable for selecting expansion terms to generate new queries for retrieval.

Table 5. RESULTS MERGING EVALUATION RESULTS, P@5 and nDCG@5

| Results Merging Method | P@5 | nDCG@5 |
|---|---|---|
| **Round Robin** | 0.3280 | 0.1910 |
| **Shadow Document Method** | 0.3300 | 0.1909 |
| **Multi-Evidence Method** | 0.3300 | 0.1909 |

Table 6. RESULTS MERGING EVALUATION RESULTS, P@10 and nDCG@10

| Results Merging Method | P@10 | nDCG@10 |
|---|---|---|
| **Round Robin** | 0.3063 | 0.1814 |
| **Shadow Document Method** | 0.3027 | 0.1797 |
| **Multi-Evidence Method** | 0.3027 | 0.1797 |

### 5.3 Query Variants Results Merging (QVRM) vs Selective Collection Enrichment (SCE)

In Table 7 Table 8, a comparison was made between the retrieval results obtained from QVRM over the results obtained for the SCE approach [3]. The results in Table 7 shows that the SCE has improved the retrieval performance over the QVRM in terms of P@5 and nDCG@5. In particular, improved retrieval performance is indicated for SCE approach when pre-retrieval query performance pre-predictors AvICTF and SCS are used to select an enriched query for retrieval on the local collection. Table 8 presents retrieval evaluation results of the QVRM algorithms vs the retrieval evaluation results obtained when deploying the SCE approach In terms of P@10 and nDCG@10. The results indicate that the SCE approach shows significant improved retrieval performance over the QVRM in terms of nDCG@10. In particular, these results show improved performance for all the three pre-retrieval query performance predictors when used to select an enriched query. However, QVRM shows improved retrieval performance over the SCE approach in terms of P@10. This retrieval improvement is evident for all the three results merging algorithms or methods. It is interesting to note that the different methods (QVRM vs SCE) outperforms each other depending on which evaluation measure is being used. For example, when P@10 is used, the QVRM outperforms the SCE approach. However, when the nDCG@10 evaluation measure is used, the SCE approach outperforms the QVRM. Since nDCG@K uses graded relevance and P@K uses binary relevance, these results suggests that the SCE approach ranks less and highly relevant documents in the top 10 retrieved documents compared to the QVRM which ranks more and slightly relevant documents in the top 10 retrieved documents.

Table 7. EVALUATION RESULTS OF THE QVRM vs SCE BASED ON QUERY PERFORMANCE PREDICTORS' SCORES: P@5 and nDCG@5

| Method | | P@5 | nDCG@5 |
|---|---|---|---|
| **Result Merging Algorithm** | RR | 0.3280 | 0.1910 |
| | SDM | 0.3300 | 0.1909 |
| | MEM | 0.3300 | 0.1909 |
| **Selective Collection Enrichment** | SCS | ↑**0.3320** | ↑**0.2779** |
| | AvIDF | ↑0.3233 | ↑**0.2726** |
| | AvICTF | ↑**0.3427** | ↑**0.2900** |

Table 8. EVALUATION RESULTS OF THE QVRM vs SCE BASED ON QUERY PERFORMANCE PREDICTORS' SCORES: P@10 and nDCG@10

| Method | | P@10 | nDCG@10 |
|---|---|---|---|
| **Result Merging Algorithm** | RR | ↑**3063** | 0.1814 |
| | SDM | ↑**0.3027** | 0.1797 |
| | MEM | ↑**0.3027** | 0.1797 |
| **Selective Collection Enrichment** | SCS | 0.3020 | ↑**0.2589** |
| | AvIDF | 0.2960 | ↑**0.2549** |
| | AvICTF | 0.2960 | ↑**0.2553** |

## 6. CONCLUSIONS AND FUTURE WORK

The main aim of this research was to investigate whether merging the retrieval results generated by multiple query variants with the same information need can yield improved retrieval performance. In addition, an investigation was conducted to compare the results generated by the query variants results merging with the selective collection enrichment approach, where only results generated by a single query which was predicted to perform better on the local collection was used for retrieval. Three different result merging strategies which are normally used in distributed search environments with large overlapping databases were deployed in this study. In particular, Round Robin (RR), Shadow Document Method (SDM) and Multi-Evidence Method (MEM) were used. Moreover, the results of this investigation were compared with the results generated by the individual query variants when not merged. The results of this investigation suggests that merging results generated by multiple query variants with the same information need can improve the retrieval performance. Also, it was observed that the choice of the external collection used in generating these query variants with the same information need can have an impact in the retrieval performance as it can sometimes lead to a degradation in the retrieval performance. When a comparison between the query variants results merging and the selective collection enrichment approach is made, it was observed that the different methods (QVRM vs SCE) outperforms each other depending on which evaluation measure is

being used. For example, when P@10 is used, the QVRM outperforms the SCE approach. However, when the nDCG@10 evaluation measure is used, the SCE approach outperforms the QVRM. Since nDCG@K uses graded relevance and P@K uses binary relevance, these results suggests that the SCE approach ranks fewer and highly relevant documents in the top 10 retrieved documents compared to the QVRM which ranks more and slightly relevant documents in the top 10 retrieved documents. This research has thrown up many questions in need of further investigation. Further research might explore deploying a machine learning approach to merge these search results from multiple query variants with the same information need.

## 7. REFERENCES

[1] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and ir system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 625–634, New York, NY, USA, 2015. ACM.

[2] G. Zuccon, J. Palotti, and A. Hanbury. Query variations and their effect on comparing information retrieval systems. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 691–700, New York, NY, USA, 2016. ACM.

[3] O. Tibi, E. Thuma, and G. Mosweunyane. Selective collection enrichment in user-centred health information retrieval. In *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, pages 175–181, July 2017.

[4] Dongqing Zhu, Stephen T. Wu, James J. Masanz, Ben Carterette, and Hongfang Liu. Using discharge summaries to improve information retrieval in clinical domain. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

[5] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, and J. Palotti. Overview of the share/clef ehealth evaluation lab 2014. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 172–191, Cham, 2014. Springer International Publishing.

[6] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the clef ehealth evaluation lab 2015. In Josanne Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Capellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 429–443, Cham, 2015. Springer International Publishing.

[7] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval*, volume 2997, pages 127–137, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[8] L. Azzopardi. Query side evaluation: An empirical analysis of effectiveness and effort. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 556–563, New York, NY, USA, 2009. ACM.

[9] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 21–28, New York, NY, USA, 1995. ACM.

[10] J. Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in information retrieval*, pages 127–150. Kluwer, 2000.

[11] D. Hawking, N. Craswell, P. B. Thistlewaite, and D. Harman. Results and challenges in web search evaluation. *Computer Networks*, 31(11-16):1321–1330, 1999.

[12] E.M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 172–179, New York, NY, USA, 1995. ACM.

[13] S. Wu and F. Crestani. Shadow document methods of resutls merging. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, pages 1067–1072, New York, NY, USA, 2004. ACM.

[14] S. Wu and S. I. McClean. Result merging methods in distributed information retrieval with overlapping databases. *Inf. Retr.*, 10(3):297–319, 2007.

[15] J.H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 267–276, New York, NY, USA, 1997. ACM.

[16] J. Arguello, J.L. Elsas, J. Callan, and J.G. Carbonell. Document representation and query expansion models for blog recommendation. In *Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008, Seattle, Washington, USA, March 30 - April 2, 2008*, 2008.

[17] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A framework for selective query expansion. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, pages 236–237, 2004.

[18] K. L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 250–256, New York, NY, USA, 1998. ACM.

[19] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings*, pages 43–54, 2004.

[20] L. Goeuriot, G.J.F. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, H. Suominen, and G. Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF 2013 Evaluation Labs and Workshop*, Valencia, Spain, 2013.

[21] T. Chappell and S. Geva. Working notes for topsig at share/clef ehealth 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

[22] X. Zhong, Y. Xia, Z. Xie, S. Na, Q. Hu, and Y. Huang. Concept-based medical document retrieval: THCIB at CLEF ehealth lab 2013 task 3. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

[23] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G.J. F. Jones, and H. Müller. Share/clef

ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 43–61, 2014.

[24] W. Shenwei, J.-Y. Nie, X. Liu, and X. Liu. An investigation of the effectiveness of concept-based approach in medical information retrieval. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 236–247, 2014.

[25] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst.*, 27(4):23:1–23:37, November 2009.

[26] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanburyn, G. J.F. Jones, M. Lupu, and P. Pecina. CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, September 2015.

[27] Y. Song, Y. He, Q. Hu, L. He, and E. M. Haacke. ECNU at 2015 ehealth task 2: User-centred health information retrieval. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.

[28] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, and G. Zuccon. Overview of the CLEF ehealth evaluation lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 255–266, 2016.

[29] G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Müller, J. Budaher, and A. Deacon. The IR task at the CLEF ehealth evaluation lab 2016: User-centred health information retrieval. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 15–27, 2016.

[30] S. Saleh and P. Pecina. Task3 patient-centred information retrieval: Team CUNI. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 123–129, 2016.

[31] Y. Song, Y. He, H. Liu, Y. Wang, Q. Hu, and L. He. ECNU at 2016 ehealth task 3: Patient-centred information retrieval. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 157–161, 2016.

[32] E. Thuma, N. P. Motlogelwa, and T. Leburu-Dingalo. Task 3: Patient-centered information retrieval, irtask 1: ad-hoc search - TEAM ub-botswana. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 162–166, 2016.

[33] C.W. Cleverdon, J. Mills, and M.E. Keen. Aslib cranfield research project - factors determining the performance of indexing systems; volume 2, test results. Technical report, Cranfield University, England, UK, Technical Report 1966, 1966.

[34] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[35] K. Roberts, M. S. Simpson, E. M. Voorhees, and W. R. Hersh. Overview of the TREC 2015 clinical decision support track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.

[36] B. Koopman and G. Zuccon. Relevation!: An open source system for information retrieval relevance assessment. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1243–1244, New York, NY, USA, 2014. ACM.

[37] G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Proceedings of the Medical Information Retrieval Workshop at SIGIR co-located with the 37th annual international ACM SIGIR conference (ACM SIGIR 2014), Gold Coast, Australia, July 11, 2014.*, pages 32–35, 2014.