You *Can* Teach an Old Dog New Tricks: Rank Fusion applied to Coordination Level Matching for Ranking in Systematic Reviews

Harrisen Scells[1][0000−0001−9578−7157], Guido Zuccon[1][0000−0003−0271−5563], and Bevan Koopman[2][0000−0001−5577−3391]

[1] The University of Queensland, St Lucia, Australia {h.scells@uq.net.au}
[2] CSIRO, Brisbane, Australia

Abstract. Coordination level matching is a ranking method originally proposed to rank documents given Boolean queries that is now several decades old. Rank fusion is a relatively recent method for combining runs from multiple systems into a single ranking, and has been shown to significantly improve the ranking. This paper presents a novel extension to coordination level matching, by applying rank fusion to each sub-clause of a Boolean query. We show that, for the tasks of systematic review screening prioritisation and stopping estimation, our method significantly outperforms the state-of-the-art learning to rank and bag-of-words-based systems for this domain. Our fully automatic, unsupervised method has (i) the potential for significant real-world cost savings (ii) does not rely on any intervention from the user, and (iii) is significantly better at ranking documents given only a Boolean query in the context of systematic reviews when compared to other approaches.

Keywords: Coordination Level Matching · Rank Fusion · Systematic Reviews · Boolean Queries · Information Retrieval

1 Introduction

The goal of medical systematic review literature search is to retrieve all research publications relevant to a highly focused research question that satisfies an inclusion criteria. This is so that *all* literature relevant to the review's research question can be synthesised in the systematic review [23]. Search takes place using a Boolean query that is formulated by highly trained information specialists using their own intuition and domain knowledge, in order to capture the information need of the systematic review [8]. Afterwards, every study retrieved by the Boolean query is *screened* (assessed) for inclusion using the titles and abstracts of studies (abstract level assessment). Identified relevant abstracts are further processed by acquiring the full-text for additional assessment, information extraction and synthesis [23].

The process of creating a medical systematic review typically involves large monetary and temporal costs; the average Cochrane review costs $350K to create [35] and it takes up to two years to publish – thus often rendering the result

of the systematic review already out-of-date at the time of publication. The process that incurs the most cost when creating a systematic review is the screening of studies retrieved by the Boolean query; often a large set of studies is retrieved, but only a handful are relevant.

A number of solutions have arisen to address the amount of time spent screening documents, including: *screening prioritisation* (which seeks to re-rank the set of retrieved documents to show more relevant documents first, thus starting the full-text screening earlier), and *stopping estimation* (which seeks to predict at what point continuing to screen will no longer contribute gain) [25,26,40]. In this paper, we propose and evaluate a Boolean query ranking function aimed at tackling these two tasks. The proposed method incorporates intuitions from both coordination level matching of Boolean queries and search engine rank fusion.

This paper proposes an extension to coordination level matching (CLM) by exploiting the query-document relationship with rank fusion. CLM is a ranking function originally proposed for Boolean queries that scores documents using the occurrences of documents retrieved by different clauses of the query. The proposed extension, *coordination level fusion* (CLF), has many advantages over CLM that enable it to use multiple weighting schemes (rankers) and different fusion methods dependent on the Boolean clauses. We use CLF to rank studies in the screening prioritisation task of systematic reviews. We further plan to study the use of a cut-off threshold tuned on training data to control when the screening of studies should be stopped based on the CLF retrieval score. The empirical results obtained on the CLEF Technology Assisted Review datasets [25,26] show that CLF significantly outperforms existing state-of-the-art methods that consider similar settings, including the ranking method currently used in PubMed (a popular database to search for literature for systematic reviews).

2 Related Work

Systematic reviews are costly and often out-of-date by the time they are published due to the amount of time involved in their creation. A wide range of systematic review creation processes have been considered for automation or improvement using semi-automatic techniques [40], including: query formulation [27,46,48], screening prioritisation [37,7,3,54,2,47,29,28,1,56], stopping prediction [16,7,24], assessment of bias [43,33], among others. This paper proposes a technique for screening prioritisation, thus the remainder of this section focuses on this specific task.

Active learning has been explored extensively for screening prioritisation and automatic assessment [37,10,56,1]. However, the main drawbacks of active learning are that a poor initial ranking will slow down the rate of learning, and that explicit human effort is required to update the ranking. While current practice prescribes all documents must be screened (therefore explicit assessments could be used for active learning), an initially poor ranking would require many assessments before the system is able to identify relevant documents. Thus the analysis of the full-text of eligible documents may be delayed. Automatic assess-

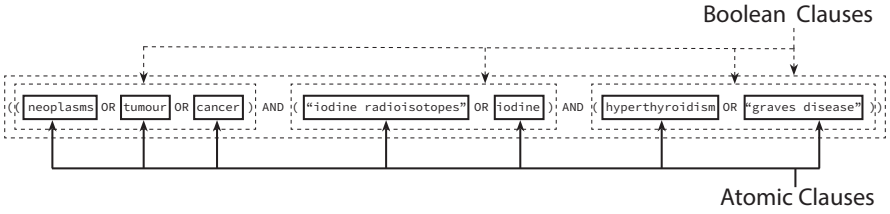


Fig. 1: Types of clauses in a Boolean query. Dashed lines surround Boolean clauses, dotted lines surround atomic clauses.

ment has been suggested to be used in place of a second researcher performing screening [40]. Fully automatic methods of screening prioritisation allow for other processes of systematic reviews to begin earlier *and* do not require the effort of humans, saving more time (and costs). In this paper, we do not consider screening prioritisation methods based on active learning. However, we note that CLF could be used as the first pass ranking in the context of an active learning method. Then, active learning could be used to augment CLF to performing re-ranking in the presence of continuous, iterative relevance feedback. We leave the study of CLF in an active learning setting for future work.

The CLEF Technology Assisted Reviews (TAR) track [25,26] considers both screening prioritisation and stopping prediction tasks. The screening prioritisation task has gained substantial interest from CLEF participants, with submitted methods including active learning [12,13], relevance feedback [4,39,21,52,55,36,38,18], automatic supervised [17,30,51,47,9], and automatic unsupervised methods (which do not rely on any relevance feedback or human intervention) [3,2,7,54]. Meanwhile, the stopping prediction task has seen little participation and naïve techniques like static score-based cut-offs [24], as well as techniques based on continuous relevance feedback [16] are used. Many of the participants also do not use the Boolean queries directly, instead resorting only to the title of the review (a sentence), which is contrived and unrealistic in the context of systematic review literature search. This work overcomes these shortcomings by only using the Boolean query to rank documents, with no additional effort required by the information specialist.

Several approaches to ranking documents retrieved by Boolean queries were proposed in the '80s and '90s outside of the context of systematic review creation. Most of these approaches rely on users explicitly weighting terms in the query [42], probabilistic retrieval using fuzzy set theory [41,6] and term dependencies [15]. A drawback of these methods is their heavy reliance on the users to impose a ranking over retrieved documents (e.g., the requirement that users must specify individual term weightings). Users often are unable to provide such weights, or it creates an additional hindrance in using the retrieval system.

A ranking function for Boolean queries which relies solely on the structure of the Boolean query, without further user intervention, is Coordination Level Matching (CLM) [31]. The intuition behind CLM is that nested sub-clauses

of a Boolean query could be considered as separate but related queries, and therefore documents that appear in multiple clauses should be ranked higher. For example, a very common way information specialists formulate Boolean queries for systematic review literature search is to break a search down into three of four categories based on the Population, Intervention, Controls, Outcomes (PICO) framework [8]. Query terms from each category become a clause in the Boolean query, grouped together by a single `AND` operator [8]. Formally, in CLM the score of a document d is the number of Boolean clauses of the query Q that are satisfied by it. A clause can be considered as both a single atomic keyword, and the grouping of several keywords or other nested groupings by a single Boolean operator (Boolean clause). Figure 1 visualises the differences between atomic clauses and Boolean clauses.

Rankings produced by CLM typically perform poorly (as supported by our empirical findings in Section 5.1). This is because the amount of information about the query being exploited to produce a document ranking is low. CLM has been noted to be more effective when weighting occurrences of documents by, for example, IDF or TF-IDF [14]. Which weighting scheme to use for CLM is then unclear, and some documents may be ranked higher than others using different weighting schemes. Moreover, when computing scores, CLM does not account for the different Boolean operators present in the query, i.e., scores are summed in the same manner irrespective of the operator used, e.g., `AND`, `OR`.

The CLF method proposed in this paper exploits rank fusion [49], i.e., the combination of multiple document rankings, typically returned by different systems or weighting schemes for the same query (although recent work has applied fusion to different query variations [5]). There are many methods for fusion of rankings, and they can be classified into two main categories [22]: score-based [49] and rank-based [32]. Score-based methods fuse rankings using the original scores of documents in different rankings to infer the new fused ranking. As systems and weighting schemes will typically assign wildly different scores to documents, scores are often normalised before fusion (e.g., using min-max normalisation). Rank-based methods fuse rankings using only the rank positions of documents (similarly to electoral vote fusion [32]).

The novelty of our contribution is that by combining insights from decades-old research about ranking documents directly with Boolean queries with relatively more recent research about the fusion of ranked lists, significant gains in effectiveness can be obtained.

3 Coordination Level Fusion

In this paper, we propose Coordination Level Fusion (CLF), a novel method that extends the traditional Coordination Level Matching (CLM) [31] by integrating rank fusion into the Boolean retrieval model by exploiting the semantic and syntactic aspects of the Boolean query.

CLM's intuition is that documents retrieved by many clauses should be considered more likely to be relevant. We note that this intuition is supported by

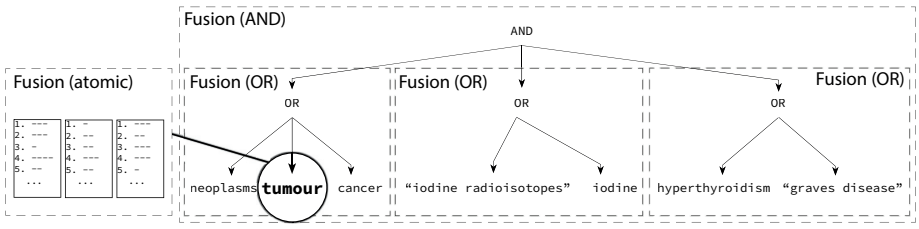


Fig. 2: Bottom-up visualisation of the fusion of ranked lists using the CLF method. First one or more ranked lists of an atomic clause are fused, then the results of each Boolean clause are fused. Each clause that has fusion applied to is encapsulated in a dashed box. The nested clauses which it encapsulates are included inside it. Each applicable fusion method is labelled within each respective box. Note that all atomic clauses use the same range of weighting schemes: in this figure only one is shown for space reasons.

axioms put forward in axiomatic analyses of ranking functions [19], and, more importantly for our work, it is similar to the intuition of rank fusion, namely, the *chorus effect*: the fact that "several retrieval approaches suggest that an item is relevant to a query" [53]. CLF leverages this intuition to further boost relevant documents higher up the ranking, using the agreement from multiple weighting schemes (rankers) *and* the agreement afforded by the structure of Boolean queries. Next, we describe the CLF method for ranking documents.

3.1 Producing a Ranking

We assume that a set R of rankings $r_1, r_2, \ldots, r_k$ is available for each atomic Boolean clause (i.e., a term in the Boolean query, see Figure 1) . These rankings could be produced by any weighting scheme available, e.g., IDF, BM25, etc. A ranking is an ordered list of documents: $r =< d_0, d_1, ..., d_k >$ with $s(d_i, r_j)$ representing the score of document d_i within ranking r_j. In CLF, these rankings are recursively fused, first at an atomic clause level, then at the level of (often nested) Boolean operators, until the highest level of the Boolean query is considered (typically represented by an AND operator): at this level, rankings are again fused together to produce a single, final ranking. This is achieved by applying the CLF fusion function to each document d as:

$$f_{CLF}(R, T, d) = \begin{cases} \sum_{r_j \in R} s(d, r_j) & \text{if } T = \texttt{AND} \\ |d \in R| \cdot \sum_{r_j \in R} s(d, r_j) & \text{if } T = \texttt{OR/Atomic} \end{cases} \quad (1)$$

where R is the set of rankings associated with the clauses of the Boolean query considered at the current level, and T is the type of Boolean operator applied. In this work, we consider T as being either identifying an atomic clause, or the `AND`

```
(((oesophag*[All Fields] OR endocapsule[All Fields] OR microcam[All Fields] OR esophag*[All
Fields] OR enteroscop*[All Fields] OR pillcam[All Fields] OR videocapsule*[All Fields]) AND
("Esophageal and Gastric Varices"[Mesh Terms:noexp] OR (gastroesophag*[All Fields] OR
oesophag*[All Fields] OR oesophago gastric varix[All Fields] OR paraoesophag*[All Fields] OR
oesophago gastric varic*[All Fields] OR periesophag*[All Fields] OR perioesophag*[All Fields]
 OR esophag*[All Fields]))) AND (23593613[pmid] OR 23029720[pmid] OR 22379346[pmid] OR
22346246[pmid] OR 22155754[pmid] OR 21814064[pmid] OR 21624583[pmid] OR 21429016[pmid] OR
21372764[pmid] OR 21274889[pmid] OR 20490679[pmid] OR 20684186[pmid] OR 20682230[pmid] OR
20363433[pmid] OR 20135731[pmid] OR 20054320[pmid] OR 19809355[pmid] OR 19743993[pmid]))
```

Fig. 3: Example query formatted to be issued to PubMed for re-ranking. Constructing the query like above ensures only the documents specified (e.g., document number `23593613`) are retrieved, and therefore re-ranked.

and `OR` operators. The queries we consider do not have `NOT` clauses (therefore we do not have a fusion method for this operator). According to Equation 1, CLF performs CombSUM fusion [49] if the Boolean clause is `AND` ($T = \texttt{AND}$). Likewise, CombMNZ fusion [49] is used when dealing with atomic clauses or the `OR` operator. Figure 2 visualises how fusion is performed for different Boolean clauses. When scoring exploded MeSH terms, the score provided by a weighting scheme is the summed score of each child in the subsumption (similar for phrases). Both CombSUM and CombMNZ boost the documents which multiple rankers estimate to be highly relevant (i.e., the chorus effect), however CombMNZ at the `OR` and atomic levels is used to combat less accurate estimates of relevance (i.e., the dark horse effect). That is, documents where only a single ranker estimates them as highly relevant are not boosted.

3.2 Stopping Prediction

The task of stopping prediction in systematic review literature search is that: given a ranking of the set of documents retrieved by the Boolean query, at what position should screening stop? We model this task with an equivalent description: given a set of documents retrieved by a Boolean query, what is the subset of documents which does not need to be screened? In this work, stopping prediction is performed by exploiting the scores of documents for each atomic term after fusion. Rather than setting a fixed cut-off on scores similar to participants in the CLEF TAR task [24], here a gain-based approach is used. Our approach is as follows: Given that researchers will screen documents starting at the first document and continuing to the next document for the entire list, they are accumulating gain from documents (equal to the document score) as they continue down the list of documents. Once enough gain from documents has been accumulated, they can stop screening. To model this, we use a κ parameter to control what percentage of the total gain a researcher can accumulate before stopping. The stopping point therefore becomes the position of the document in the ranked list where the cumulative gain exceeds the total allowable gain. When κ is set to 1, no documents are discarded. In the task of screening prioritisation, where documents are assessed, κ is set to 1.

4 Experimental Setup

Empirical evaluation is conducted on the CLEF TAR 2017 and 2018 collections [25,26]. For the 2018 collection, evaluation is performed on topics from Task 2. Experiments are compared with respect to two baselines: a ranking obtained by submitting queries directly to PubMed (explained in detail below), and a ranking obtained by using CLM. The results of the CLF rankings are also compared to the rankings produced by the participants of the CLEF TAR task. Note that many of these participants do not rank directly according to the terms and structure of the Boolean query (while we do), and often consider the query as a bag-of-words, and incorporate terms from the title for re-ranking. Also note that many of the participants used feedback from the relevance assessments and created active learning solutions. The comparisons between participants and our results only consider those which reported to not use relevance assessments and do not use human intervention to rank (fully automatic, thus excluding active learning settings). In other words, we experiment considering the first round of retrieval.

All experiments are run using the QueryLab domain-specific Information Retrieval framework [45]. To obtain statistics for ranking documents, the documents retrieved by each query are fetched from PubMed and indexed by QueryLab. No stopwording or stemming is applied. The particular queries in this collection contain terms which are explicitly stemmed. Therefore, we use the PubMed Entrez API [44] to identify the original terms in documents from the explicitly stemmed term (this backward approach to stemming is to allow information specialists fine-grained control over their search). The title, abstract, MeSH headings, and publication date of each PubMed document is stored in four separate fields. When a title was not available for a document, the book title field was used instead; if no book title was available, the field was left empty (this replicates how searching on the title field works in PubMed). All of the experimental code to reproduce the experiments is made available at `https://github.com/ielab/clf`.

The following weighting schemes are used in our experiments to produce document rankings for an atomic clause: IDF, TF-IDF, BM25, InL2 of Divergence from Randomness, PubMed, term position, text score, publication date, and document length. The PubMed weighting scheme uses the state-of-the-art learning to rank system of Pubmed [20]. The best match ranking system of PubMed uses a three-stage ranking system: first, documents are retrieved using the Boolean query; then, documents are ranked using BM25; finally, top-ranked documents are re-ranked using LambdaMART trained on click data, using document features such as document length, publication date, and past usage. Note that the PubMed best match ranker can *only* rank documents given a term or phrase, *not* a Boolean query. After the first stage, the Boolean query is translated into a bag-of-words type of query, similar to those seen in web search (it is often the case that the query translation results in fewer documents retrieved). Therefore, by embedding the PubMed ranker into CLF, the query translation step may be skipped entirely. The term position weighting scheme is defined as the relative position of a term in a document (0 if the term does not appear

in the document). Publication date scores documents higher if the document is newer (accounting for recency, linearly). Document length scores documents higher the longer the document is. Text score weights documents by the fields a term appears in: for example, a document is scored higher if a term appears in the title and the body than if the term appears only in the body. When queries are submitted to PubMed, they are modified to restrict them to only the PMIDs reported in the CLEF topic file (in order to account for minor discrepancies in retrieval after different time periods, see Figure 3 for an example), and set the retrieval mode in PubMed to 'relevance' in order to obtain a ranked list of documents by relevance (instead of the default ranking by publication date). Prior to fusion for any clause, ranked lists are normalised using min-max normalisation. Z-score and softmax normalisation were also considered, however through early empirical testing, min-max normalisation provided the consistently higher effectiveness compared to z-score and softmax. When there are ties in the ranking, the document which has a more recent publication date is ranked higher. The different modifications made to CLF used in this paper are taxonimised below:

CLM – The basic form of coordination level matching using the approach described in Section 1.

CLF+PuMmed – CLF, using the PubMed ranker via the PubMed Entrez API.

CLF+weighting – CLF, using the weighting schemes described in the paragraph above (excluding the PubMed weighting scheme).

CLF+weighting+PubMed – CLF, using all of the weighting schemes from **CLF+weighting** in addition to the PubMed ranker from **CLF+PubMed**.

CLF+weighting+qe – CLF, using all the the weighting schemes from **CLF+weighting**, but with a naíve query expansion method using terms from the topic titles and terms specific to DTA systematic reviews (obtained from an information specialist). Here, two additional Boolean `OR` clauses are constructed, each containing terms from the title and DTA specific terms respectively. Terms from the title have stopwording and Porter stemming applied.

CLF+weighting+PubMed+qe – CLF, using all of the weighting schemes from **CLF+weighting**, in addition to the PubMed ranker from **CLF+PubMed**, and the approach to query expansion from **CLF+weighting+qe**.

4.1 Evaluation

Evaluation is performed differently depending on the task. For the screening prioritisation task, rank-based measures are used. For comparison between the CLEF TAR participants (of which we acquired the runs), the MAP measure is included. The nDCG measure is included as a more realistic model of user behaviour. Reciprocal rank (RR) is used to demonstrate the effectiveness of systems in an active learning scenario (to show how soon the first relevant document would be shown and an update to the ranking potentially triggered). Precision after R documents (Rprec) is used to show the theoretical best possible precision obtainable in the stopping task, along with last relevant (Last Rel) that

reports at what rank position the very last relevant document was shown. Participant runs are chosen for comparison if they are a fully automatic, unsupervised method, which does not use the training data or explicit relevance feedback, and do not set a threshold (as categorised in the TAR overview papers [25,26]). Note that the tables in the CLEF TAR overview papers contain errors regarding these aspects, instead each of the participant's papers were considered to individually determine which runs to directly compare our methods to. For the stopping prediction task, several standard set-based measures are used: precision, recall, $F_{\beta=\{0.5,1,3\}}$, total cost, and reliability [11]. Reliability is a loss measure (i.e., where smaller values are better) specifically designed for the TAR task. It has two components: $loss_r = 1-(\text{recall})^2$ and $loss_e = (n/(R+100)*100/N)^2$, where n is the number of documents retrieved, N is the size of the collection, and R is the total number of relevant documents. Therefore, Reliability = $loss_r + loss_e$. Participants runs are chosen if they are fully automatic, supervised or unsupervised (thus we consider approaches that used training data), do not use explicit relevance feedback, and do set a threshold. Runs are evaluated using `trec_eval` or the evaluation scripts that are provided by the CLEF TAR organisers, where applicable.

When used for predicting when to stop screening, κ is tuned on training queries using a grid search to determine the best value. The parameter space searched in these experiments is $\{0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 0.9, 0.95\}$. Note that κ can be set at a clause-level, therefore it is possible for it to be adaptive based on the clause. We leave learning an adaptive κ for future work, and here we fix κ to a set value across all clauses.

5 Results

5.1 Screening Prioritisation

Tables 1 and 2 present the results of the screening prioritisation task for the 2017 and 2018 CLEF TAR collections. Comparing CLM to CLF (without query expansion), CLF is statistically significantly better than CLM in all of the evaluation measures presented in both 2017 and 2018 tables (using a two-tails t-test where $p < 0.05$). Comparing the CLM and CLF methods to the state-of-the-art PubMed ranking, CLM is often statistically significantly worse than the PubMed ranker, whereas some CLF-based methods are able to perform statistically significantly better than the PubMed method. Next, the best performing CLF method (**CLF+weighting+PubMed+qe**) and the best performing CLEF participant method for each year is compared. For 2017 topics, the best performing methods are Sheffield-run-2 (documents ranked with TF-IDF vector space model using terms from topic title and terms extracted from the Boolean query) and Sheffield-run-4 (same as Sheffield-run-2 except a PubMed stopword list is used) [3]. The CLF method does not perform statistically significantly better than these two methods in any evaluation measure considered (however in all measures apart from MAP and last relevant, CLF is better). For 2018 topics, the best performing method is Sheffield-general-terms (same as Sheffield-run-4

	MAP	nDCG	RR	Rprec	Last Rel
PubMed	0.1597	0.5378	0.4292	0.1786	2974.00
CLM	0.0483*†	0.3941*†	0.1344*†	0.0415*†	3763.76*
CLF+PubMed	0.1313	0.5129	0.3722	0.1387	3119.06
CLF+weighting	0.1494	0.5247	0.4213	0.1696	3307.76
CLF+weighting+PubMed	0.1643	0.5422	0.4028	0.1754	3048.10
CLF+weighting+qe	0.1960	0.5735	0.5326	0.2239	3301.73
CLF+weighting+PubMed+qe	0.2165*	**0.5939***	**0.6037***	**0.2302**	3028.03
Sheffield-run-1	0.1700	0.5404	0.3644	0.1788	2678.33
Sheffield-run-2	**0.2183**	0.5930	0.5085	0.2190	2441.70
Sheffield-run-3	0.1986	0.5770	0.4700	0.2115	2404.96
Sheffield-run-4	0.2179	0.5937	0.5099	0.2185	2382.46*
ECNU-run1	0.0905*†	0.4517*†	0.1849*†	0.0907*†	3633.16*
QUT-bool	0.1293	0.4221*	0.3465	0.1535	1972.20*†
QUT-pico	0.1197	0.4067*	0.3088	0.1565	**1873.53***†

Table 1: Results for CLEF TAR 2017. The first row of results is obtained by issuing queries to PubMed, the next set of rows is are results of the various configurations of CLF, and the last set of rows are the relevant runs from participants for that year. Two-tailed t-test between the PubMed ranker and the other methods with $p < 0.05$ is indicated by $*$ and $p < 0.01$ by $\dagger$.

	MAP	nDCG	RR	Rprec	Last Rel
PubMed	0.1918	0.5971	0.5085	0.2131	**3479.40**
CLM	0.0483*†	0.4413*†	0.1338*	0.0316*†	7194.76
CLF+PubMed	0.1734*†	0.5938	0.4942	0.2002	6363.13
CLF+weighting	0.2012	0.6186	0.5331	0.2139	6061.06
CLF+weighting+PubMed	0.2363*	0.6390	0.5289	0.2435	5937.93
CLF+weighting+qe	0.2397*	0.6501	0.5969	0.2662*†	5931.13
CLF+weighting+PubMed+qe	**0.2722***†	**0.6767***†	**0.6649**	**0.2882***†	5743.26
ECNU-TASK2-RUN1-TFIDF	0.1415*	0.5682	0.4212	0.1862	7173.00
sheffield-general-terms	0.2584*	0.6495*	0.4723	0.2779*	5519.20
sheffield-query-terms	0.2243	0.6184	0.4012	0.2425	5736.70

Table 2: Results for CLEF TAR 2018. Presentation of results and statistical significance is indicated the same was as in Table 1.

from 2017, however terms specifically designed to identify systematic reviews are added to the query) [2]. Comparing this method to CLF, the CLF method performs statistically significantly better in RR (and has gains in all evaluation measures apart from last relevant). Overall, CLF is able to obtain the highest MAP overall for 2018 topics, and the highest overall nDCG, RR, and Rprec for both 2017 topics and 2018 topics, performing statically significantly better than the state-of-the-art PubMed ranker.

	Precision	Recall	F_1	$F_{0.5}$	F_3	Total Cost	Reliability
No stopping	0.0415	**1.0000**	0.0752	0.0505	0.2345	3918.70	0.5441
CLF/0.4	**0.1040***†	0.7836*†	**0.1545***†	**0.1186***†	**0.3286***†	1324.63*†	**0.1259***†
ecnu-run2	0.0397	0.7075*†	0.0696	0.0478	0.2085	**1000.00***†	0.4445
ecnu-run3	0.0399	0.7164*†	0.0700	0.0480	0.2102	**1000.00***†	0.4433
sis.t1	0.0461*†	0.9868	0.0834*†	0.0561*†	0.2544*†	3435.03*†	0.4453*†
sis.t1.5	0.0482*†	0.9727*	0.0865*†	0.0585*†	0.2596*†	3165.56*†	0.3843*†
sis.2	0.0517*†	0.9531*†	0.0919*†	0.0626*†	0.2684*†	2824.6667*†	0.3309*†
sis.t2.5	0.0577*†	0.9382*†	0.1007*†	0.0695*†	0.2815*†	2536.80*†	0.2724*†

Table 3: Results of CLF for stopping prediction for CLEF TAR 2017. The first row are the results from the original queries, the second row is when CLF with $\kappa = 0.4$. Two-tailed t-test between the original results and the other methods with $p < 0.05$ is indicated by $*$ and $p < 0.01$ by $\dagger$.

	Precision	Recall	F_1	$F_{0.5}$	F_3	Total Cost	Reliability
No stopping	0.0471	**1.0000**	0.0851	0.0573	0.2622	4640.23	0.3981
CLF/0.4	**0.1225***†	0.8582*†	**0.1827***†	**0.1400***†	**0.3794***†	**1140.06***†	0.4330*†

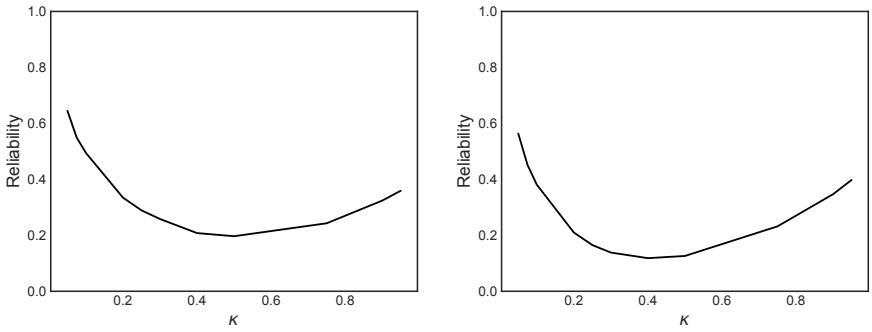
Table 4: Results of CLF for stopping prediction for CLEF TAR 2018. The first row are the results from the original queries, the second row is when CLF with $\kappa = 0.4$. Significance is indicated the same as in Table 3.

5.2 Stopping Prediction

Tables 3 and 4 present the results of the stopping prediction task using the cut-off parameter κ. A κ value of 0.4 through parameter tuning on training data was found to provide the least loss in Reliability, and was therefore chosen for the test queries for both 2017 and 2018. Results of the parameter tuning process on the training portion of the CLEF 2017 and 2018 topics are presented in Figure 4. The CLF method used in this task was **CLF+weighting+PubMed+qe** as it obtained the highest performance on the screening task.

Examining first Table 3, CLF obtains the highest precision, F_1, $F_0.5$, F_3, and lowest loss in reliability. CLF also obtains the second-lowest total cost, and maintains both a low total cost and reliability for this set of queries. Losses in recall are within a tolerable threshold [11].Table 4, reveals similar results to the 2017 topics. Significant improvements over the original queries in terms of precision, F_1, $F_{0.5}$, F_3, and total cost, with a tolerable reduction in recall can be observed. However, the Reliability on this set of queries is higher (thus worse). Given that the total cost is low, this indicates that the $loss_r$ component of Reliability does not decrease at the same rate as $loss_e$ increases for these topics. There were no participants which contributed a comparable run to the 2018 TAR task, therefore no comparisons to other systems can be made for this collection.

While there is a drop in recall, there are real monetary savings associated with the increase in precision. Across the 2017 and 2018 topics, the CLF method provides savings between approximately USD$5000 and USD$12,000, according to estimates reported by McGowan et al. [35] when considering double screening.



(a) 2017 topics. **(b)** 2018 topics.

Fig. 4: Tuning the κ parameter on the training portions of the 2017 (left) and 2018 (right) CLEF TAR topics. Lowest value for both plots is 0.4.

6 Conclusion & Future Work

In this paper, a novel approach to ranking documents for systematic review literature search using rank fusion applied to coordination level matching was presented. The method, dubbed Coordination Level Fusion (CLF), outperformed the current state of the art for two different tasks. For the screening prioritisation task, CLF significantly outperformed the existing PubMed ranking system, as well as participants that submitted comparable runs to the CLEF TAR tasks. The results of the screening prioritisation task demonstrate the applicability of CLF to systematic review literature search when prioritisation is considered, and suggest it may also be applied to obtain an effective early ranking in settings that consider active learning. For the stopping prediction task, CLF could significantly reduce the cost of screening with tolerable losses in recall. The results of the stopping prediction task demonstrate the applicability of CLF to specific systematic reviews where total recall is not essential, such as in rapid reviews [34].

There are many aspects about CLF that require further investigation. First, we propose to study the effectiveness of CLF within an active learning setting. In this context, CLF can be used as the first ranker, before relevance feedback is collected. Then, feedback could be further weaved into CLF by devising and integrating weighting schemes that account for this. We also plan to investigate the use of CLF as a method for query performance prediction (e.g., as a post-retrieval predictor using reference lists [50], or as a candidate selection function in query transformation chain frameworks [48]). In terms of extending CLF, the weighting schemes themselves can be weighted (i.e., one weighting scheme may have more importance over others); e.g., using the linear combination fusion method [53] which assigns weights to each ranker being fused. The problem then is learning the weight to assign to each weighting scheme (ranker) used for rank fusion. Rather than using fusion methods like CombMNZ, it is foreseeable to use a different combination of weights for each Boolean clause considered.

Acknowledgements. Harrisen is the recipient of a CSIRO PhD Top Up Scholarship. Dr Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award. This research is supported by the National Health and Medical Research Council Centre of Research Excellence in Informatics and E-Health (1032664).

References

1. Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M.D., Cormack, G.V., Grossman, M.R.: A system for efficient high-recall retrieval. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1317–1320 (2018)
2. Alharbi, A., Briggs, W., Stevenson, M.: Retrieving and ranking studies for systematic reviews: University of sheffield's approach to clef ehealth 2018 task 2. In: CEUR Workshop Proceedings. vol. 2125. CEUR Workshop Proceedings (2018)
3. Alharbi, A., Stevenson, M.: Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to clef ehealth 2017 task 2. In: CLEF (Working Notes) (2017)
4. Anagnostou, A., Lagopoulos, A., Tsoumakas, G., Vlahavas, I.P.: Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In: CLEF (Working Notes) (2017)
5. Benham, R., Culpepper, J.S., Gallagher, L., Lu, X., Mackenzie, J.: Towards efficient and effective query variant generation. In: DESIRES. pp. 62–67 (2018)
6. Buell, D.A.: A general model of query processing in information retrieval systems. Information Processing & Management **17**(5), 249–262 (1981)
7. Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., He, L., Yang, Y.: Ecnu at 2017 ehealth task 2: Technologically assisted reviews in empirical medicine. In: CLEF (Working Notes) (2017)
8. Clark, J.: Systematic reviewing. In: Suhail A. R. Doi, G.M.W. (ed.) Methods of Clinical Epidemiology. Springer (2013)
9. Cohen, A.M., Smalheiser, N.R.: Uic/ohsu clef 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In: CEUR Workshop Proceedings. vol. 2125 (2018)
10. Cohen, A., Hersh, W., Peterson, K., Yen, P.: Reducing workload in systematic review preparation using automated citation classification. JAMIA **13**(2), 206–219 (2006)
11. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 75–84 (2016)
12. Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. In: CLEF (Working Notes) (2017)
13. Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2018. In: CLEF (Working Notes) (2018)
14. Crestani, F.: Exploiting the similarity of non-matching terms at retrieval time. Information Retrieval **2**(1), 27–47 (2000)
15. Croft, W.B.: Boolean queries and term dependencies in probabilistic retrieval models. Journal of the American society for Information Science **37**(2), 71–77 (1986)

16. Di Nunzio, G.M.: A study of an automatic stopping strategy for technologically assisted medical reviews. In: European Conference on Information Retrieval. pp. 672–677. Springer (2018)
17. Di Nunzio, G.M., Beghini, F., Vezzani, F., Henrot, G.: An interactive two-dimensional approach to query aspects rewriting in systematic reviews. ims unipd at clef ehealth task 2. In: CLEF (Working Notes) (2017)
18. Di Nunzio, G.M., Ciuffreda, G., Vezzani, F.: Interactive sampling for systematic reviews. ims unipd at clef 2018 ehealth task 2. In: CLEF (Working Notes) (2018)
19. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 480–487. ACM (2005)
20. Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., Osipov, M., Kholodov, M., Ismagilov, R., Mohan, S., et al.: Best match: new relevance search for pubmed. PLoS biology **16**(8), e2005343 (2018)
21. Hollmann, N., Eickhoff, C.: Ranking and feedback-based stopping for recall-centric document retrieval. In: CLEF (Working Notes) (2017)
22. Hsu, D.F., Taksa, I.: Comparing rank and score combination methods for data fusion in information retrieval. Information retrieval **8**(3), 449–480 (2005)
23. JPT CHH, G.S.: Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated march 2011]. The Cochrane Collaboration (2011)
24. Kalphov, V., Georgiadis, G., Azzopardi, L.: Sis at clef 2017 ehealth tar task. In: CEUR Workshop Proceedings. vol. 1866, pp. 1–5 (2017)
25. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: CLEF'17 (2017)
26. Kanoulas, E., Spijker, R., Li, D., Azzopardi, L.: Clef 2018 technology assisted reviews in empirical medicine overview. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes (2018)
27. Karimi, S., Pohl, S., Scholer, F., Cavedon, L., Zobel, J.: Boolean versus ranked querying for biomedical systematic reviews. BMC MIDM **10**(1), 1 (2010)
28. Lagopoulos, A., Anagnostou, A., Minas, A., Tsoumakas, G.: Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 52–63. Springer (2018)
29. Lee, G.E., Sun, A.: Seed-driven document ranking for systematic reviews in evidence-based medicine. In: Proceedings of the 41st annual International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 455–464. SIGIR '18 (2018)
30. Lee, G.E.: A study of convolutional neural networks for clinical document classification in systematic reviews: sysreview at clef ehealth 2017 (2017)
31. Losee, R.: Probabilistic retrieval and coordination level matching. Journal of the American Society for Information Science **38**(4), 239–244 (1987)
32. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp. 387–396. ACM (2006)
33. Marshall, I.J., Kuiper, J., Wallace, B.C.: RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association (2015)
34. Marshall, I.J., Marshall, R., Wallace, B.C., Brassey, J., Thomas, J.: Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. Journal of clinical epidemiology **109**, 30–41 (2019)

35. McGowan, J., Sampson, M.: Systematic reviews need systematic searchers (irp). Journal of the Medical Library Association **93**(1), 74 (2005)
36. Minas, A., Lagopoulos, A., Tsoumakas, G.: Aristotle university's approach to the technologically assisted reviews in empirical medicine task of the 2018 clef ehealth lab. In: CLEF (Working Notes) (2018)
37. Miwa, M., Thomas, J., O'Mara-Eves, A., Ananiadou, S.: Reducing systematic review workload through certainty-based screening. JBI **51**, 242–253 (2014)
38. Norman, C., Leeflang, M., Névéol, A.: Limsi@ clef ehealth 2018 task 2: Technology assisted reviews by stacking active and static learning. In: CLEF (Working Notes) (2018)
39. Norman12, C., Leeflang, M., Névéol, A.: Limsi@ clef ehealth 2017 task 2: Logistic regression for automatic article ranking (2017)
40. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews **4**(1), 5 (2015)
41. Radecki, T.: A probabilistic approach to information retrieval in systems with boolean search request formulations. Journal of the American Society for Information Science **33**(6), 365–370 (1982)
42. Salton, G., Fox, E.A., Wu, H.: Extended boolean information retrieval. Tech. rep., Cornell University (1982)
43. Savoie, I., Helmer, D., Green, C.J., Kazanjian, A.: Beyond medline: reducing bias through extended systematic review search. International journal of technology assessment in health care **19**(1), 168–178 (2003)
44. Sayers, E.: A general introduction to the e-utilities. Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information (2010)
45. Scells, H., Locke, D., Zuccon, G.: An information retrieval experiment framework for domain specific applications. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (2018)
46. Scells, H., Zuccon, G.: Generating better queries for systematic reviews. In: Proceedings of the 41st annual International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18 (2018)
47. Scells, H., Zuccon, G., Deacon, A., Koopman, B.: Qut ielab at clef ehealth 2017 technology assisted reviews track: Initial experiments with learning to rank. In: CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum. vol. 1866, pp. Paper–98. CEUR Workshop Proceedings (2017)
48. Scells, H., Zuccon, G., Koopman, B.: Automatic boolean query refinement for systematic review literature search. In: The World Wide Web Conference. pp. 1646–1656 (2019)
49. Shaw, J.A., Fox, E.A.: Combination of multiple searches. NIST SPECIAL PUBLICATION SP pp. 105–105 (1995)
50. Shtok, A., Kurland, O., Carmel, D.: Query performance prediction using reference lists. ACM Transactions on Information Systems (TOIS) **34**(4), 19 (2016)
51. Singh, G., Marshall, I., Thomas, J., Wallace, B.: Identifying diagnostic test accuracy publications using a deep model. In: CEUR Workshop Proceedings. vol. 1866. CEUR Workshop Proceedings (2017)
52. Singh, J., Thomas, L.: Iiit-h at clef ehealth 2017 task 2: Technologically assisted reviews in empirical medicine. In: CLEF (Working Notes) (2017)
53. Vogt, C.C., Cottrell, G.W.: Fusion via a linear combination of scores. Information retrieval **1**(3), 151–173 (1999)

54. Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., He, L.: Ecnu at 2018 ehealth task 2: Technologically assisted reviews in empirical medicine. Methods **4**(5), 7 (2018)
55. Yu, Z., Menzies, T.: Data balancing for technologically assisted reviews: Undersampling or reweighting. In: CLEF (Working Notes) (2017)
56. Zou, J., Li, D., Kanoulas, E.: Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 949–952 (2018)