Proceedings Paper:
Alharbi, A. orcid.org/0000-0001-9327-7284 and Stevenson, R. orcid.org/0000-0002-9483-6006 (2019) Improving ranking for systematic reviews using query adaptation. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Bürki, G.H., Cappellato, L. and Ferro, N., (eds.) CLEF 2019 Proceedings : Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019: Conference and Labs of the Evaluation Forum, 09-12 Sep 2019, Lugarno, Switzerland. Lecture Notes in Computer Science (11696). Springer , pp. 141-148. ISBN 9783030285760

https://doi.org/10.1007/978-3-030-28577-7_9

Improving Ranking for Systematic Reviews Using Query Adaptation

Amal Alharbi[1,2] and Mark Stevenson[1]

[1] University of Sheffield, UK
[2] King Abdulaziz University, Saudi Arabia
{ahalharbi1,mark.stevenson}@sheffield.ac.uk

Abstract. Identifying relevant studies for inclusion in systematic reviews requires significant effort from human experts who manually screen large numbers of studies. The problem is made more difficult by the growing volume of medical literature and Information Retrieval techniques have proved to be useful to reduce workload. Reviewers are often interested in particular types of evidence such as Diagnostic Test Accuracy studies. This paper explores the use of query adaption to identify particular types of evidence and thereby reduce the workload placed on reviewers. A simple retrieval system that ranks studies using TF.IDF weighted cosine similarity was implemented. The Log-Likelihood, Chi-Squared and Odds-Ratio lexical statistics and relevance feedback were used to generate sets of terms that indicate evidence relevant to Diagnostic Test Accuracy reviews. Experiments using a set of 80 systematic reviews from the CLEF2017 and CLEF2018 eHealth tasks demonstrate that the approach improves retrieval performance.

1 Introduction

Systematic reviews aim to identify and summarise all evidence available to answer a specific research question such as 'Is systemic inflammation present in stable chronic obstructive pulmonary disease?' [7]. Conducting a systematic review is a time-consuming and expensive process which can require up to 12 months of expert effort [5, 10] and costs as much as a quarter of a million dollars [11]. The process of identifying relevant evidence for systematic reviews consists of several tasks: (1) *Boolean search* where the experts create a Boolean query that matches the criteria of a systematic review and apply the query to a medical database such as MEDLINE. (2) *Title and abstract screening* where reviewers screen the title and abstract of each retrieved study and exclude studies that are obviously not relevant to the systematic review. (3) *Content screening* where the reviewers screen the full studies identified as relevant from the previous step and decide their relevance to the systematic review [8]. A significant part of the experts' time is spent screening studies to identify those relevant to the review. Reviewers often manually screen a large number of irrelevant studies to identify the few relevant ones. Text mining techniques have been proven their

effectiveness in reducing the workload and time needed to conduct systematic reviews [12, 13, 16].

In 2017 and 2018, the CLEF eHealth forum ran a task on systematic reviews that aimed to support the screening phase by (semi)automatically ranking the studies by relevance to the review [8, 9]. Results from these exercises demonstrated that automating the screening stage of systematic review can be efficient in identifying most, if not all, relevant studies with less effort and time than manual screening.

Researchers are generally interested in a specific type of evidence (e.g. Randomised Control Trials). In this paper, we aim to exploit this fact by generating queries which aim to identify a particular type of evidence, Diagnostic Test Accuracy (DTA) studies. DTA reviews are considered to be challenging for text mining approaches [8]. This paper demonstrates that query adaptation methods can be used to identify the terms they are characteristic of studies likely to be relevant for DTA reviews. These terms are used to expand queries used for the CLEF2017 and CLEF2018 task where it is found that they lead to improved performance. These results show that incorporating information about the general type of review (e.g. DTA) improves performance when identifying relevant medical evidence. The code implementing the experiments described in this paper is available from `https://github.com/Amal-Alharbi/Improving_ranking_for_systematic_reviews.git`.

2 Query Adaptation

Query adaptation is the process of reformulating a given query with the aim of improving retrieval performance [4]. We hypothesised that there are terms which distinguish the studies that are likely to be included in DTA reviews from other literature. Expanding the Boolean query with those terms may help to find the most relevant studies. We aim to use several query adaption approaches to derive a list of key terms that indicate evidence relevant to DTA reviews. In this paper, we use lexical statistics in addition to relevance feedback.

2.1 Lexical Statistics

We use three lexical statistics: Log-Likelihood, Chi-Squared and Odds-Ratio. These statistics are widely applied in corpus linguistics where they are used to identify the key terms that are characteristic of a sub-corpus [6, 14, 15]. We treat the relevant documents as a sub-corpus and aim to identify the terms that characterise it so that they can be used to adapt the query.

Log-Likelihood is computed as

$$\textit{Log-Likelihood} = 2 \times \left(O_{rel} \times \log \frac{O_{rel}}{E_{rel}} + O_{irrel} \times \log \frac{O_{irrel}}{E_{irrel}} \right) \tag{1}$$

where O_{rel} and O_{irrel} are the observed frequency of the term in different subsets of the collection (e.g. relevant and irrelevant documents). E_{rel} and E_{irrel} are the term's expected frequency calculated as

$$E_{rel} = N_{rel} \times \frac{O_{rel} + O_{irrel}}{N_{rel} + N_{irrel}} \quad , \quad E_{irrel} = N_{irrel} \times \frac{O_{rel} + O_{irrel}}{N_{rel} + N_{irrel}} \tag{2}$$

where N_{rel} and N_{irrel} represent sub-corpus size (e.g. relevant and irrelevant documents). Terms are assigned high Log-Likelihood scores for a particular corpus when their observed frequency is (much) higher than the expected frequency.

Chi-Squared is calculated as

$$\textit{Chi-Squared} = \frac{(O_{rel} - E_{rel})^2}{E_{rel}} + \frac{(O_{irrel} - E_{irrel})^2}{E_{irrel}} \tag{3}$$

where O_{rel} and O_{irrel} are the observed values and E_{rel} and E_{irrel} are expected values calculated using equation 2.

Odds-Ratio is most commonly applied for keyword analysis and terms identification [14]. The Odds-Ratio for each term calculated as

$$\textit{Odds-Ratio} = \frac{O_{rel} \times (N_{irrel} - O_{irrel})}{O_{irrel} \times (N_{rel} - O_{rel})} \tag{4}$$

where O_{rel} and O_{irrel} are the frequency counts of the term in the relevant and irrelevant sub-corpus and N_{rel} and N_{irrel} are the total number of terms in each of those sub-corpus.

2.2 Relevance Feedback

This approach aims improve a query by taking account of feedback about the results it returned. Rocchio's algorithm [3] (equation 5) was used to reformulate the query by enriching it with additional terms weighted using information about the relevance of the documents it returned.

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{N_{rel}} \sum_{\forall \vec{d_j} \in D_{rel}} \vec{d}_j - \frac{\gamma}{N_{irrel}} \sum_{\forall \vec{d_j} \in D_{irrel}} \vec{d}_j \tag{5}$$

where $\vec{q}$ is the original query vector, $\vec{d}_j$ is a weighted term vector associated with abstract j. D_{rel} is the set of relevant abstracts among the abstracts retrieved and N_{rel} is the number of abstracts in D_{rel}. D_{irrel} is the set of irrelevant abstracts among the abstracts retrieved and N_{irrel} is the number of abstracts in D_{irrel}. α, β and γ are weighting parameters.

3 Experiments

The experiments use reviews titles, Boolean queries, list of PubMed identifiers (PMIDs) and relevance judgements provided for the CLEF2017 and CLEF2018 tasks on Technology Assisted Reviews in Empirical Medicine [8,9]. The Boolean queries were manually constructed by experts and designed to match the criteria of each systematic review. The PMIDs refer to the documents retrieved from the Boolean queries. The relevance judgements files indicate which of the documents returned by the Boolean query were indicated as being relevant after the *Title and Abstract Screening* and *Content Screening* stages. All reviews are related to DTA studies. i.e. reviews which report the accuracy of a specific test for identifying a disease.

The CLEF2017 dataset contained 266,967 abstracts divided into training and test datasets containing 20 and 30 reviews, respectively. The CLEF2018 dataset contained 460,165 abstracts and divided into a training dataset consisting of 42 reviews and test dataset of 30 reviews. (Note that the training split of the CLEF2018 dataset is a subset of CLEF2017 dataset.)

3.1 Experiment 1: Lexical Statistics

The Log-Likelihood, Chi-Squared and Odds-Ratio statistics were used to derive lists of terms that indicate evidence relevant to DTA reviews as described in Section 2.1. The training sets from CLEF 2017 and CLEF 2018 were partitioned into relevant and irrelevant documents depending upon whether the study was included in the systematic review. Terms that occurred fewer than ten times were excluded since it is difficult to generate reliable statistics for these rare terms and, also, they are unlikely to be useful for identifying relevant studies. After computing the lexical statistics for each term in every review, the average for each statistic for each term across all the reviews in the training dataset was computed as

$$Avg\_statistic = \frac{\sum_{i=1}^{T} statistic_i}{T} \tag{6}$$

where $statistic_i$ represent the statistic (Log-Likelihood, Chi-Squared or Odds-Ratio) for the term in review i and T is the total number of reviews in the training portion of the dataset (20 for the CLEF2017 dataset and 42 for the CLEF2018 dataset). For each lexical statistic, the terms with the highest scores are identified and added to the query for each review in the test portion of the dataset. The abstracts in the test dataset are ranked by matching terms from the expanded queries against those in the abstracts using a simple TF.IDF weighted cosine similarity measure.

The evaluation was carried out using software similar to trec_eval provided by CLEF2017 organisers[3]. The performance metrics reported here are average precision (AP) and work saved over sampling at 95% and 100% recall (WSS@95 and WSS@100). Table 1 shows the results of experiment 1. The baseline results

[3] `https://github.com/leifos/tar`

Table 1. Lexical Statistic results for CLEF2017 and CLEF2018 test datasets.

		(a) CLEF2017 Dataset			(b) CLEF2018 Dataset		
Lexical Statistic	**Terms**	**AP**	**WSS@100**	**WSS@95**	**AP**	**WSS@100**	**WSS@95**
Baseline	-	0.218	0.385	0.493	0.224	0.377	0.506
Log-Likelihood	5	0.232	**0.389**	**0.507**	0.244	0.389	0.525
	10	0.227	0.380	0.497	0.251	0.407	0.535
	20	**0.233**	0.384	**0.507**	**0.259**	**0.414**	**0.545**
Chi-Squared	5	0.214	**0.389**	0.490	0.232	0.380	0.515
	10	**0.230**	**0.389**	0.507	0.242	0.396	0.530
	20	**0.230**	**0.389**	**0.508**	**0.253**	**0.409**	**0.547**
Odds-Ratio	5	0.214	**0.389**	0.490	0.221	0.377	0.505
	10	0.214	0.388	0.489	0.231	0.380	0.515
	20	**0.233**	**0.389**	**0.506**	**0.252**	**0.398**	**0.541**

were obtained without adding any additional terms to the query. The lower part of the table shows the results that were obtained when different numbers of terms with the highest scores were added to each query using different statistics (Log-Likelihood, Chi-Squared and Odds-Ratio). Retrieval performance improved when the additional terms are added to the queries, and this improvement is consistent across evaluation metrics for both dataset (CLEF2017 and CLEF2018). Enriching the query with more key terms generally improved performance. These results demonstrate that the additional terms, generated from an independent set of reviews, provide information about the types of studies that are likely to be relevant for DTA reviews, independently of their specific review.

Table 2. Ten terms with highest Log-Likelihood scores derived from CLEF2017 training dataset.

Term	**Score**	**Term**	**Score**
sensitivity	58.249	mtb rif	31.054
predictive	41.675	positive	30.308
gonadotropin	38.557	vulva	29.346
hcg	32.743	protein	28.686
false	31.090	fetoprotein	28.053

Table 2 shows the ten terms with the highest Log-Likelihood scores derived from the CLEF 2017 training dataset. We noticed that the top terms identified by the lexical statistic include ones that are highly indicative of the discussions that are found in DTA reviews, for example "sensitivity", "predictive" and "positive" are terms which describe the accuracy of a medical test. The presence of these terms is likely to indicate that the study will be relevant to a DTA review so adding them to the query improves performance. It is also interesting to note that several of the terms that appear in this list are also used in standard filters for DTA reviews that have been developed to support information professionals searching for relevant literature [1]. However, we also note that the list also includes terms that appear to be specific to particular DTA reviews (e.g. "gonadotropin"). The CLEF 2017 training dataset contains only 20 reviews, and if

a particular term proves to be very important for a small set of reviews, then its overall score can be high enough for it to be included in this list.

3.2 Experiment 2: Relevance Feedback

In this experiment, abstracts in the test dataset were ranked using a simple TF.IDF weighted cosine similarity measure comparing each abstract with terms extracted from the Boolean query. Relevance judgement from the 10% top-ranked abstracts (up to a maximum of 1,000) were used to reformulate the query using Rocchio's algorithm and the remaining abstracts re-ranked using the updated query vector. A range of values for the weighting parameters were previously explored [2] and it was found that the best results were achieved by setting $\alpha = \beta = 1$ and $\gamma = 1.5$. In this experiment, we applied two approaches: (1) use all the terms of the modified query $\vec{q}_m$ and (2) exclude terms with weight less than or equal zero (i.e. terms with negative weight).

Results are shown in Table 3. Retrieval performance improves for most metrics when using relevance feedback (compared with the baseline and lexical statistics in Table 1). On the other hand, a higher AP score for the CLEF 2018 dataset is obtained using lexical statistics. Including only terms which Rocchio's algorithm weighted positively improves the AP score and saves more effort than using all terms.

Table 3. Relevance Feedback results for CLEF2017 and CLEF2018 test datasets.

	(a) CLEF2017 Dataset			(b) CLEF2018 Dataset		
$\vec{q}_m$ **terms**	**AP**	**WSS@100**	**WSS@95**	**AP**	**WSS@100**	**WSS@95**
all	0.236	0.342	0.485	0.222	0.345	0.496
+ve	**0.243**	**0.432**	**0.557**	**0.238**	**0.420**	**0.608**

4 Conclusion and Future Work

Query adaption methods can be used to identify terms characteristic of studies likely to be relevant for DTA reviews. The experiments reported in this paper demonstrate that including general information about the type of publication that is likely to be of relevance for a systematic review can improve retrieval performance. The best performance was achieved using relevance feedback.

In the future, we would like to apply those methods to other types of systematic review (e.g. prognosis reviews and intervention reviews). In addition, we would like to explore alternative methods for identifying useful terms such as synonym-based query expansion that may overcome some of the limitations of approaches used in this study.

References

1. Search Filters for MEDLINE in Ovid Syntax and the PubMed translation, https://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx. Last accessed 18 January 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: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, France (2018)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley Publishing Company, USA, 2nd edn. (2011)
4. Carpineto, C., Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys **44**(1), 1–50 (2012)
5. Cohen, A.M., Ambert, K., McDonagh, M.: A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In: AMIA Annual Symposium proceedings. AMIA Symposium. vol. 2010, pp. 121–5 (2010), `http://www.ncbi.nlm.nih.gov/pubmed/21346953`
6. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational linguistics **19**(1), 61–74 (1993)
7. Gan, W.Q., Man, S., Senthilselvan, A., Sin, D.: Association between chronic obstructive pulmonary disease and systemic inflammation: a systematic review and a meta-analysis. Thorax **59**(7), 574–580 (2004)
8. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF technologically assisted reviews in empirical medicine overview. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum. CEUR Workshop Proceedings, CEUR-WS.org, Dublin, Ireland (2017)
9. 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. CEUR-WS, France (2018)
10. Karimi, S., Pohl, S., Scholer, F., Cavedon, L., Zobel, J.: Boolean versus ranked querying for biomedical systematic reviews. BMC medical informatics and decision making **10**(1), 1–20 (2010)
11. McGowan, J., Sampson, M.: Systematic reviews need systematic searchers. Journal of the Medical Library Association **93**(1), 74–80 (2005)
12. 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)
13. Paisley, S., Sevra, J., Stevenson, M., Archer, R., Preston, L., Chilcott, J.: Identifying potential early biomarkers of acute myocaridal infarction in the biomedical literature: a comparison of text mining and manual sifting techniques. In: Proceedings of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 19th Annual European Congress. Vienna, Austria (2016)
14. Pojanapunya, P., Todd, R.W.: Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Ling. Theory **14**(1), 133–167 (2018)
15. Rayson, P.: From key words to key semantic domains. International Journal of Corpus Linguistics **13**(4), 519–549 (2008)
16. Shemilt, I., Khan, N., Park, S., Thomas, J.: Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Systematic reviews **5**(1), 140 (2016)