Reducing Workload of Systematic Review Searching and Screening Processes

Harrisen Scells
Queensland University of Technology
harrisen.scells@hdr.qut.edu.au

Systematic reviews, in particular medical systematic reviews, are time consuming and costly to produce. The largest contributing factors to the time and monetary costs are the searching (including the formulation of queries) and screening processes. These initial processes involve researchers reading the abstracts of thousands and sometimes hundreds of thousands of research articles to determine if the retrieved articles should be included or excluded from the systematic review. This research explores automatic methodologies to reduce the workload relating to the searching and screening processes.

Keywords: Information Retrieval, Systematic Reviews, Workload Reduction

1. INTRODUCTION

A systematic review is an exhaustive literature review with a single focused research question. Systematic reviews are used primarily, but not exclusively, in evidence based medicine to guide clinical decisions and inform policy. Medical systematic reviews are costly and time consuming and can take years to complete. In the worst case scenario, a systematic review is out of date by the time it is published. The two components of systematic reviews that comprise the majority of the workload are the *searching* (including query formulation) and *screening* processes, whereby researchers manually analyse thousands and sometimes millions of medical studies depending on the type of the review. Reducing the workload required for medical researchers to compile a systematic review can save tens of thousands of dollars and hundreds of hours of medical researchers' time. Each systematic review defines a search strategy like that of Figure 1 comprising of an often large and complex boolean query. This boolean query is significant to a systematic review as it is the method for retrieving studies to be analysed for inclusion in the review. A search strategy is devised in the searching phase of a systematic review. Information experts such as librarians work closely with medical researchers to formulate a search strategy that captures the information need set out by the research question of the systematic review. Once a search strategy is established, the reviewers begin the screening process, whereby the abstracts of medical studies retrieved by the search strategy are screened to determine if the studies should be included in the review. The manual appraisal of large quantities of documents in the screening process reveals a problem that technology can solve: *how well can the searching and screening processes be automated*? The aim of this research is to reduce the workload required by reviewers for both processes while maintaining the search quality. In doing so, this research will identify automatic techniques to support the creation of search strategies and investigate retrieval systems that are more effective at retrieving fewer non-relevant studies, and will

```
1. Diabetic Ketoacidosis/
2. Diabetic Coma/
3. (diabet* and (keto* or coma)).tw.
4. DKA.tw.
5. or/1-5
6. Insulin Aspart/
7. Insulin, Short-Acting/
8. (glulisine or apidra).tw.
9. (humulin or novolin).tw.
10. (novolog or novorapid).tw.
11. acting insulin*.tw.
12. or/6-11
13. 5 and 12
```

***Figure 1:** A typical search strategy found in a systematic review. Note that the line numbers are part of the search strategy. The search strategy is nested by referring to the line numbers, for instance `or/1-5` on line 6 means that lines 1 - 5 should be combined with a Boolean **OR** operator. Medical subject headings (/) and the fields to search on (**tw** - title and abstract) are also encoded in the search strategy.*

Proceedings of the 7th Symposium on Future Directions in Information Access 2017

define measurements to identify what makes a good search strategy. Due to systematic reviews requiring stringent reproducibility requirements, this research aims to improve upon existing systems while closely matching the reproducibility requirements of existing search strategies. To perform these experiments, this research requires a high quality collection of search strategies and annotated studies; which has been constructed and published in the first stage of this research (Scells et al. 2017c). Additionally, the CLEF 2017 Technology Assisted Reviews track has developed a similar collection; although limited to one type of review (Kanoulas et al. 2017; Goeuriot et al. 2017). The technologies that will be investigated with the aim of decreasing workload are ranking algorithms, rank cutoff estimation, query performance prediction, faceted search, retrievability analysis, and retrieval bias. Reducing the workload required for reviewers to search and appraise studies for systematic reviews has many benefits. Firstly, it reduces associated time and cost factors, and secondly, reviews can be published in a more timely manner leading to more confident clinical decisions. Finally, updating a review, which can also be an expensive process, may become more convenient. The outcomes of this research has significant benefits: more cost effective and timely reviews leading to more confident clinical decisions, ultimately improving the quality of healthcare delivery. As such, the overarching research question is *what can technology do to decrease the workload for compiling medical systematic reviews?*

2. OBJECTIVES

Systematic reviews are a cornerstone of evidence based medicine, guiding clinical decisions and inform policy outside of academia. Evidence based medicine relies on both clinical expertise and the best available external clinical evidence in the form of systematic reviews (Sackett et al. 1996). As such, this research will focus on medical systematic reviews. Compiling a systematic review is time consuming and costly (Wallace et al. 2016). For example, Allen and Olkin (1999) found that many systematic reviews consume more than 1,000 man hours with the majority of that time spent searching for and screening medical studies. Meanwhile, McGowan and Sampson (2005) found that the total cost of systematic reviews can often be upwards of a quarter of a million dollars and that librarians, who assist with the searching and screening processes, contribute to a significant portion of the total cost of a systematic review. Reducing the workload associated with searching and screening medical studies will enable potentially significant reductions to the time and cost factors of a systematic review. As such, the objectives of this research are to:

1. Create a test collection in order to empirically evaluate the effectiveness of different aspects of systematic review retrieval and screening. The test collection is a crucial component of this research. Preliminary research has investigated existing collections and found them to be insufficient for the needs of this research. An initial test collection using PubMed and 94 queries has been created to facilitate evaluation (Scells et al. 2017c) as the majority of research in this area has been evaluated on an outdated and significantly smaller collection (Cohen et al. 2006). In parallel, the CLEF 2017 Technology Assisted Reviews track has created a similar collection, with slightly queries Kanoulas et al. (2017); Goeuriot et al. (2017). The TREC *Total Recall* track focuses on similar problems in a different domain, however domain specific approaches such as medical concept extraction would not be applicable.

2. Produce models and algorithms that reduce the workload surrounding the searching and screening processes of systematic reviews. This research will investigate information retrieval techniques. Preliminary research using learning to rank models to re-rank systematic review literature searches (Liu 2009) has been recognised in the CLEF2017 eHealth Lab working notes (Task 2) (Scells et al. 2017a).

3. Understand the structure of a search strategy. This problem is addressed in this objective by better understanding what makes an effective search strategy, and to offer suggestions of, more effective search strategies. The outcome of this objective is to support medical researchers through query suggestion; given a search strategy, automatically identify alternative search strategies that are more effective.

4. Detect bias in search strategies. A systematic review must ensure that all documents that are relevant to the research question have been retrieved. Omitting a study from a systematic review can have serious consequences when informing clinical decisions. For example, when a study on the effects on smoking fails to include significant publications relating to the negative effects smoking has on the lungs, the systematic review is said to be bias. This research aims to formulate techniques for detecting search strategies that omit highly relevant studies from the search results.

3. METHODOLOGIES

My research is composed of five tasks (visualised in Figure 2) that relate to reducing the workload associated with compiling a systematic review. These tasks mostly involve no user-in-the-loop and when a user is required they will be simulated with pseudo relevance feedback, with the exception of

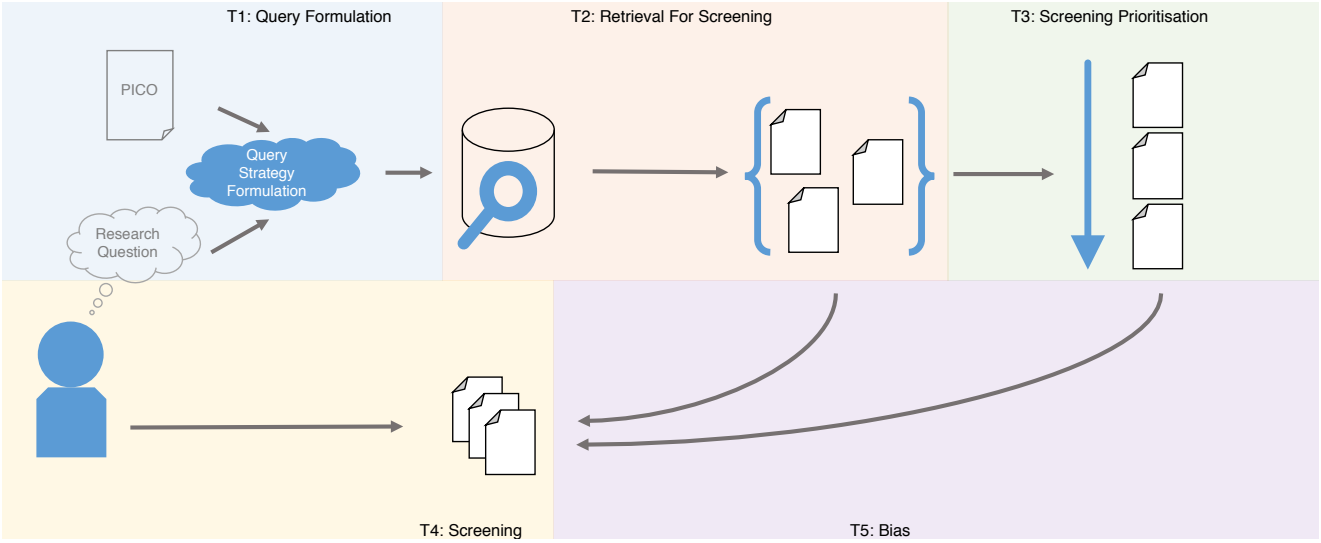

Figure 2: *The five tasks involved in the searching and screening processes that this research is undertaking.*

Task 1, in which users may be needed to understand how well performance estimations help them.

Task 1: Query Formulation A search strategy (Boolean query) is used to retrieve medical studies for screening and inclusion. This task investigates the usefulness of domain specific features, such as PICO (population, intervention, control, outcome) elements, in query formulation. PICO is a framework extensively used in systematic reviews as a means of formulating clinical questions. This task also aims to evaluate and measure the effectiveness of search strategies, as this is currently done manually with possibly little or no rigour by the information specialist. The outcome of measuring and evaluating the effectiveness of search strategies is to support the users of search systems by providing feedback on the quality of the search strategy and to suggest higher quality search strategies. This task involves a study to explore different query performance predictors (QPPs) to determine which ones are the most applicable for systematic review search strategies. There are many QPPs and each measures different aspects of a query.

Task 2: Retrieval for Screening Medical databases that contain clinical studies are often searched using boolean queries. This technology is often sub-optimal and other information retrieval domains such as patent retrieval have benefited from transitioning away from boolean queries. This task investigates systems for medical citation retrieval, including better exploiting the document structure of the indexed documents. The use of information related to the PICO element for retrieval is largely unexplored. The few examples that have attempted to exploit PICO in this context include Demner-Fushman and Lin (2007) and Boudin et al. (2010), but these were not thoroughly evaluated. Even though the majority of abstracts indexed in medical databases adhere to the PICO structure – and PICO is often used in the formulation of research questions and search strategies Boudin et al. (2010). This task involves building an annotated collection of search strategies from existing systematic reviews and searching on an annotated collection of PubMed citations. The keywords in the boolean query of the search strategies are annotated with relevant PICO elements manually. Studies are also annotated automatically to extract the PICO elements. Preliminary experiments I have conducted reveal an increase in precision when matching terms in a boolean query to sentences in abstracts that relate to the same PICO element (Scells et al. 2017b).

Task 3: Screening Prioritisation Prioritising the citations for screening (in the form of a ranked results list) enables reviewers to screen and include the most relevant studies faster. Theory of ranking and associated methods are central to the development of information retrieval research (Robertson 1977; Fuhr 2008; Zuccon and Azzopardi 2010), and document ranking is common practice in many search engine applications. One overlooked method for increasing the precision of a ranked list in systematic review information retrieval is exploiting document fields such as the journal category/type or domain specific features such as medical ontologies. This task will investigate the effectiveness of domain specific field statistics and domain specific ontologies in a learning to rank framework as features for a model to learn. Learning to rank has been used extensively outside of medical information retrieval (Liu 2009), as well as medical applications outside of systematic reviews (Limsopatham et al. 2013; Palotti et al. 2016). This task will further use active learning feedback to build a learning to rank model. Learning to rank models will be compared to active learning baselines.

Task 4: Screening Determining whether a citation should be reviewed for inclusion is a manual process that is performed by information experts. There is room for improvement, as automatic and semi-automatic processes such as active learning and text mining have been identified as inadequate or insufficient for reducing the workload associated with the screening process (Olorisade et al. 2016). Specifically, this task investigates the use of a stopping point to determine when a systematic reviewer should stop screening citations. In a ranked list of citations there may exist a point whereby continuing down the list does not provide any more citations that are relevant for inclusion in the systematic review. Predicting this point in the list supports researchers by reducing the number of citations the need to screen.

Task 5: Bias A systematic review must report the studies included and excluded from the review. This enables external reviewers and clinical experts to ensure that the systematic review is a reliable source. However, it is unclear if a search strategy can be formulated to purposefully omit studies. Consider a well formulated and properly documented search strategy. The only way to tell if relevant studies have been omitted from the review is to formulate new, similar search strategies that may capture these omitted citations. This task will investigate the difficulty in constructing search strategies that purposefully omit relevant citations from the result set, and methods for detecting these types of search strategies. This task is used as a constraint to ensure the results of the previous tasks maintain the reproducibility of existing systematic reviews. The likelihood of a document to be retrieved for a query for each document in a collection, or retrievability (Azzopardi and Vinay 2008), identifies the level of bias document fields and queries exude on a collection. The retrievability score identifies the documents that are most likely to be retrieved by a query; assisting users in the query formulation stage. Retrievability can also by used in Task 1 to support users formulating effective search strategies.

4. CONCLUSIONS

This research investigates many open questions and unconsidered solutions to information retrieval problems within the context of systematic reviews. Systematic reviews that employ the techniques and technologies recommended by this research will be faster to compile due to a significant reduction in workload in the searching and screening processes. Potentially tens or hundreds of thousands of dollars and years of time are needed to compile a single systematic review. This research investigates suitable technologies and techniques to reduce the workload in the searching and screening processes, two highly costly components of a systematic review. This research will accelerate medical systematic review screening and inclusion processes, ensuring faster time-to-publication of systematic reviews, leading to more timely and up-to-date systematic reviews, resulting in more accurate clinical decisions and ultimately higher quality patient outcomes.

REFERENCES

I. E Allen and I Olkin. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA*, 282(7):634–635, 1999.

L Azzopardi and V Vinay. Retrievability: an evaluation measure for higher order information access tasks. In *CIKM*, pages 561–570, 2008.

F Boudin, J. Y Nie, and M Dawes. Clinical information retrieval using document and pico structure. In *NAACL HLT*, pages 822–830, 2010.

A M Cohen, W R Hersh, K Peterson, and P Yen. Reducing workload in systematic review preparation using automated citation classification. *JAMIA*, 13(2):206–219, 2006.

D Demner-Fushman and J Lin. Answering clinical questions with knowledge-based and statistical techniques. *COLI*, 33(1):63–103, 2007.

N Fuhr. A probability ranking principle for interactive information retrieval. *IR*, 2008.

L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon. CLEF 2017 ehealth evaluation lab overview. In *CLEF*, 2017.

E Kanoulas, D Li, L Azzopardi, and R Spijker. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CLEF*, 2017.

N Limsopatham, C Macdonald, and I Ounis. Learning to selectively rank patients' medical history. In *CIKM*, pages 1833–1836. ACM, 2013.

TY Liu. Learning to rank for information retrieval. *FnTIR*, 3(3): 225–331, 2009.

J McGowan and M Sampson. Systematic reviews need systematic searchers (IRP). *JMLA*, 93(1):74, 2005.

B K Olorisade, E de Quincey, P Brereton, and P Andras. A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In *EASE*, page 14. ACM, 2016.

J Palotti, L Goeuriot, G Zuccon, and A Hanbury. Ranking health web pages with relevance and understandability. In *SIGIR*, pages 965–968, 2016.

S E Robertson. The probability ranking principle in IR. *J.Doc*, 33 (4):294–304, 1977.

D. L Sackett, W. MC Rosenberg, JA. M Gray, R. B Haynes, and W. S Richardson. Evidence based medicine: what it is and what it isn't, 1996.

H Scells, G Zuccon, A Deacon, and B Koopman. Qut ielab at clef 2017 technology assisted reviews track: Initial experiments with learning to rank. In *CLEF*, 2017a.

H Scells, G Zuccon, B Koopman, A Deacon, and S Geva. Integrating the framing of clinical questions via pico into the retrieval of medical literature for systematic reviews. In *CIKM*. ACM, 2017b.

H Scells, G Zuccon, B Koopman, A Deacon, S Geva, and L Azzopardi. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In *SIGIR*. ACM, 2017c.

B. C Wallace, J Kuiper, A Sharma, M. B Zhu, and I. J Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *JMLR*, 2016.

G Zuccon and L Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *ECIR*, pages 357–369. Springer, 2010.