APS: An Active PubMed Search System for Technology Assisted Reviews

Li, D.; Zafeiriadis, P.; Kanoulas, E.

DOI
10.1145/3397271.3401401

Publication date
2020

Document Version
Final published version

Published in
SIGIR '20

License
Article 25fa Dutch Copyright Act

Citation for published version (APA):
APS: An Active PubMed Search System for Technology Assisted Reviews

Dan Li
University of Amsterdam
Amsterdam, The Netherlands
d.li@uva.nl

Panagiotis Zafeiriadis
Tailo
Amsterdam, The Netherlands
zafeiriadis@hotmail.com

Evangelos Kanoulas
University of Amsterdam
Amsterdam, The Netherlands
e.kanoulas@uva.nl

ABSTRACT
Systematic reviews constitute the cornerstone of Evidence-based Medicine. They can provide guidance to medical policy-making by synthesizing all available studies regarding a certain topic. However, conducting systematic reviews has become a laborious and time-consuming task due to the large amount and rapid growth of published literature. The TAR approaches aim to accelerate the screening stage of systematic reviews by combining machine learning algorithms and human relevance feedback. In this work, we built an online active search system for systematic reviews, named APS, by applying an state-of-the-art TAR approach – Continuous Active Learning. The system is built on the top of the PubMed collection, which is a widely used database of biomedical literature. It allows users to conduct the abstract screening for systematic reviews. We demonstrate the effectiveness and robustness of the APS in detecting relevant literature and reducing workload for systematic reviews using the CLEF TAR 2017 benchmark.

CCS CONCEPTS
• Information systems → Search interfaces; Chemical and biochemical retrieval.

KEYWORDS
PubMed; TAR; Systematic Reviews; Active Search

1 INTRODUCTION
Evidence-Based Medicine (EBM) plays a significant role in health care and policy-making [7–9, 12]. The cornerstone of EBM is the synthesis of evidence presented in scientific publications through systematic reviews. Systematic reviews appraise, summarize, and synthesize all available evidence or studies regarding a certain topic (e.g., a treatment or a diagnostic test) [7–9]. However, conducting a systematic review has become a laborious task due to the large amount and rapid growth of published literature. The average time to conduct a systematic review is around 67 weeks from registration to publication, equal to more than 1000 hours of manual labor [2].

To write a systematic review, researchers have to conduct several searches that will retrieve all relevant studies and screen these studies for potential inclusion in the review. Existing databases for searching include PubMed, Medline and Embase etc. The searches typically identify thousands of potential relevant studies, which are screened on the basis of their title and abstract, and with the vast majority excluded from the final review. Studies that are not excluded during the title & abstract screening, are assessed on the basis of their full-text. The title & abstract screening phase is the most time-consuming step in the systematic review process. Hence, the need for automation in this process becomes of utmost importance.

In general, the Technology-Assisted Review (TAR) approaches retrieve a substantial number (or all) of the relevant studies by iteratively training machine learning models on the basis of human relevance feedback. We call it the TAR process. The Continuous Active Learning (CAL) approaches have been demonstrated one of the highly effective TAR approaches [3–6]. Given a document collection and a query, a ranker is trained to identify documents to be shown to reviewers for relevance assessment. Then, the assessed documents are used as training data to re-train the ranker. As more and more documents are identified by the ranker and assessed by the reviewers, the training data is further populated with more examples, which leads to more effective ranker. The TAR process continues until “enough” relevant documents have been found. The Baseline Model Implementation approach is a state-of-the-art version of CAL [3, 4]. Abualsaud et al. [1] further built a TAR system based on the CAL approach.

In this paper we describe an online active search system we developed specifically for systematic reviews in biomedical domain, named Active PubMed Search (APS). Our main contributions are (1) an new publicly available search system built on the top of the PubMed collection, for conducting systematic reviews; (2) a demonstration of the effectiveness of the system in detecting relevant literature and reducing workload for conducting systematic reviews; (3) a comparison with the Wolters Kluwer Ovid search system for conducting systematic reviews.
spent in screening irrelevant ones. The system retrieves the the top-
$k$ most relevant studies (see Initial Retrieval), with $k$ in this paper
set to 100,000, and then iteratively displays the top-10 studies from
this subset to the user and collects relevance feedback to update its
relevance prediction model (See CAL). The PubMed collection in
the initial retrieval module is indexed in advance and this is done
done offline (See PubMed Index). The TAR process continues as long
as the user clicks the Submit button to submit his/her relevance
feedback, and stops when the user clicks the Stop button. In the
end, the user gets a list of relevant studies to include in his/her
target systematic review.

The system is designed in a modular manner so that it enables
easy extension. It is logically composed of an initial retrieval mod-
ule and a CAL module. Furthermore, the initial retrieval consists of
query transformation and PubMed index; the CAL consists of query
representation, document representation and ranking model. The
default configuration of the current implementation is designed for
the convenience of biomedical experts in conducting systematic
reviews, which can be found in the later paragraphs. Possible ex-
tensions of the system could be: (1) for the query transformation
module, support Medical Subject Heading (MeSH) queries which
is designed for indexing studies in PubMed; (2) for the document
representation module, support contextual semantic representation
using BioBERT [10]; (3) for the ranking model, support ranking
model based on BioBERT [10] etc. We describe the default configu-
ration of the current implementation in the rest paragraphs.

PubMed Index. PubMed comprises more than 30 million studies
for biomedical literature. Once a year, a complete (baseline) set of
PubMed citation records in XML format is released for downloading
from their ftp servers. Incremental update files which include new,
revised, and deleted citations are released daily.

We use the repository last updated on 04/11/2019. The PubMed
studies we downloaded includes 30,262,491 unique records, among
which 19,798,457 records contain title and abstract, and the rest
10,464,034 records only contain title. The full text is not available
for all the records. As suggested by two biomedical experts (also the
potential users of APS) that it is less likely to include studies with
only title in a systematic review, we use the 19,798,457 studies with
both title and abstract in our PubMed collection. We ignore the other
metadata such as authors and the journal, and we only use the title
and abstract as the text of the studies. Figure 3 shows the text length
distribution in our collection. We fit a Mixture Gaussian model to
the data. As seen in the figure, the length of studies do not vary
much as the text of studies (title and abstract) are homogeneous. As
that seek to identify all studies relevant for conducting a systematic review. We use the 42 topics in the dataset to evaluate our system. For each topic, the following are provided: a topic description, a subset of the studies in the PubMed collection which are related to the topic and needs to be ranked, and the relevance assessments of the studies in this set.

**Evaluation metrics.** Following [13], we use gain curve to evaluate the effectiveness of the system. Gain curve is defined as recall as a function of effort, where cost is the number of documents reviewed by the user and recall is the percentage of relevant documents among reviewed documents. Besides, for the evaluation of the initial retrieval module, we also report recall, mean average precision (MAP) and mean R-precision (RP) metrics.

**Baseline.** We define PubMed Search (PS) system the Wolters Kluwer Ovid system². It is a widely used medical research platform to search PubMed. There are two major differences between PS and APS: (1) PS supports queries of the MeSH format which are designed for indexing studies in PubMed, while APS supports key word or natural language queries. (2) PS is static while APS is dynamic in the sense that the search system is updated with user interactions.

All the experiments are conducted on the complete 42 topics with 32 cores (Intel Xeon Gold 5118 CPU @ 2.30GHz).

### 4.2 Results

We first investigate the effectiveness of initial retrieval. In Table 1, we compare 6 ranking models in the initial retrieval module: BM25, QLDir and QLJM, as well as the corresponding extensions using the query expansion model RM3. BM25+RM3 performs the best for the initial ranking. It achieves recall of 0.8871 at cutoff 100,000. In Figure 4, we further examine how the performance of the retrieval model varies over different topics and cutoffs. The median and mean value become stable when cutoff exceeds 30,000, indicating that it is quite safe to set cutoff at 100,000 with respect to effectiveness.

In Figure 5, we investigate the effectiveness of CAL by comparing the performance of APS and PS. The ranked list of documents produced by APS is based on the aforementioned best method BM25.

---

1https://scikit-learn.org/

2http://demo.ovid.com/demo/ovidsptools/launcher.htm
Table 1: Performance of initial retrieval models at cutoff 100,000.

<table>
<thead>
<tr>
<th>Method</th>
<th>BM25 QLDir QLJM</th>
<th>BM25 QLDir QLJM +RM3</th>
<th>BM25 QLDir QLJM +RM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.8688 0.8652 0.8457</td>
<td><strong>0.8871</strong> 0.8762 0.8780</td>
<td>0.8070 0.0575 0.0690</td>
</tr>
<tr>
<td>MAP</td>
<td>0.0581 0.0508 0.0507</td>
<td><strong>0.0700</strong> 0.8688 0.8652</td>
<td>0.1976 0.2354 0.1893</td>
</tr>
<tr>
<td>RP</td>
<td>0.1976 0.2354 0.1893</td>
<td><strong>0.2440</strong> 0.2689 0.2453</td>
<td></td>
</tr>
</tbody>
</table>

* RM3 with a cutoff of 100,000. It can be observed that APS performs very well at the beginning and achieves very high recall after screening 5000 studies, while PS needs 10,000 or more studies to achieve the same recall. Note that when collecting relevance labels in the CLEF TAR dataset, the organisers used MeSH query and PS to create a study pool for each topic, which may be missing relevant studies that are retrieved by APS. Therefore, the true performance of APS could even be higher than in Figure 5.

Figure 4: Boxplot of the best model BM25+RM3 at different cutoffs. Red line is median value, green triangle is mean value over topics.

Figure 5: Gain curve of APS and PS.

Despite its complexity, the speed of APS is not a bottleneck when conducting systematic reviews. In the aforementioned experiments, the time to get the returned articles after the user clicks the search button is less than 1 second. Then every time the user clicks submit, the time to get the returned articles is much less than 1 second.

5 CONCLUSION

In this paper, we described the design of an online active search system for systematic review – the APS system. The system can assist systematic review practitioners to conduct system reviews and conduct comparative studies with the PS system. The modular design also makes it to serve as a platform to study TAR approaches for researchers. The experiment with simulated interaction demonstrated its effectiveness in detecting relevant literature and reducing workload for systematic reviews.

This work has two key limitations: (a) we did not conduct an actual systematic review with the developed system, and (b) the developed system does not allow for multiple users to use it concurrently.

6 ACKNOWLEDGMENTS

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the China Scholarship Council, and the Google Faculty Research Awards program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES