Testing a Citation and Text-Based Framework for Retrieving Publications for Literature Reviews

M. Janina Sarol, Linxi Liu, and Jodi Schneider

University of Illinois at Urbana-Champaign
{mjsarol, lliu73, jodi}@illinois.edu

Abstract. We propose a citation- and text-based framework to conduct literature review searches. Given a small set of articles included in a literature review (i.e. seed articles), the first step of the framework retrieves articles that are connected to the seed articles in the citation network. The next step filters these retrieved articles using a hybrid citation and text-based criteria. In this paper, we evaluate a first implementation of this framework (code available at `https://github.com/janinaj/lit-review-search`) by comparing it to the conventional search methods for retrieving the included studies of 6 published systematic reviews. Using different combinations of 3 seed articles, on average we retrieved 71.2% of the total included studies in the published reviews and 82.33% of the studies available in the search database (Scopus). Our best combinations retrieved 87% of the total included studies, which comprised 100% of the studies available in Scopus. In 5 of the 6 reviews, we reduced the number of results by 34–88%, which in practice would save reviewers significant time, since the overall number of search results that need to be manually screened is substantially reduced. These results suggest that our framework is a promising approach to improving the literature review search process.

Keywords: citation relationships, text mining, literature review, systematic search

1 Introduction

Scholarly output is large and fast-growing: as of 2018, Scopus alone covers 69 million publications, comprised of journals, conference proceedings, and books[1], and this may double by 2027 as scholarly output grows about 8% each year [1]. Staying up-to-date in such an environment is difficult, especially with an increase in interdisciplinary work. This makes literature reviews important, but time-consuming to conduct.

There are multiple types of literature reviews, and each type has different specific goals [2]. For instance, a state-of-the art review may focus on current

[1] `https://www.elsevier.com/solutions/scopus/content`

literature and emerging priorities, while rapid reviews may support policymaking by assessing what is already known on a practical topic. Systematic searching is useful for all types of literature reviews, but it is fundamental for systematic reviews, which seek 100% recall. Systematic reviews try to find all available evidence pertaining to a given research question. Thus, they become increasingly time-consuming and difficult to conduct as literature grows. It is pertinent that all retrieved search results are screened, typically manually, and classified as relevant or irrelevant.

Even small improvements in the search process for literature reviews could help researchers more efficiently retrieve relevant publications. Ross-White and Godfrey [3] studied the precision of high-recall searches used for 8 systematic reviews. They calculated that an average of 142 results needed to be screened to find 1 relevant paper. The 8 reviews they described screened a total of 17,378 abstracts to find 122 relevant articles. The time required for screening can be substantial. Bannach-Brown et al. [4] suggested that 1 person can screen an estimated 1,879 results per month. Librarians reported routinely spending 40-60 hours to develop search queries that still result in thousands of results that need to be manually screened to find a handful of relevant articles [5].

Alternative or complementary approaches to conventional term- and concept-based search methods are needed, and current work in this area is promising. For instance, CitNetExplorer was originally designed to study the evolution of science, but its citation network visualizations can also help systematically retrieve publications [6]. New approaches can also take advantage of additional publication data, which is increasingly available for electronic access and efficient retrieval. Scopus, a large scientific database, provides citation information for indexed articles. A public domain corpus of citation information, OpenCitations [7], reportedly contains reference lists for 50% of CrossRef-indexed publications as of 2018.[2] Meanwhile, many publishers provide full-text access to their content, and text mining of licensed content is increasingly feasible.[3] These additional data sources allow for the development of novel techniques that leverage different kinds of information.

We propose a citation- and text-based framework for conducting literature review searches. Our approach differs from conventional search methods in that we use publications ("seed articles") as our starting point, rather than identifying search strings. We also use the citation network of seed articles as our search and retrieval space. We then filter the results by removing publications with weak citation and topical relationships with the seed articles.

We envision this framework to be useful for different types of literature reviews. In this paper, we test a first implementation of our framework on 6 systematic reviews.

In Section 2, we provide related work on both citation and text-based information retrieval. In Section 3, we describe our framework, a sample implementa-

[2] `https://i4oc.org/#faqs`

[3] e.g. through the Crossref Text and Data Mining APIs `http://tdmsupport.crossref.org`

tion, and an experimental evaluation. In Section 4 we report our results, which we analyze and discuss in Section 5. Finally, in Section 6, we conclude the paper.

2 Related Work

2.1 Text-based Techniques for Information Retrieval

Topic modeling is one of the text mining techniques that has been frequently used for information retrieval-based tasks. Wang, McCallum, and Wei found that the use of topical phrases can improve the performance of information retrieval systems [8]. Combining collaborative filtering and topic modeling has also been shown to be a promising approach in recommending scientific publications [9].

Text mining has also been used specifically for systematic review tasks. A 2015 systematic review by O'Mara-Eves et al. [10] provides a detailed discussion of proposed solutions for screening documents. More recent approaches include a text-mining framework for screening documents for systematic reviews introduced by Li et al. [11] and a semi-supervised approach for screening relevant documents developed by Kontonatsios et al. [12].

2.2 Citation-based Techniques for Information Retrieval

Citation-based methods have also been proposed for retrieving and ranking relevant scientific publications. In a field study using real searches in health science libraries in the early 1990's, Pao [13] found that citation searching was able to add an average of 24% recall. Recent approaches include using term frequency-inverse document frequency metrics, commonly used for text-based ranking, to rank co-cited papers [14] and citation proximity analysis to recommend scientific publications [15].

Belter [16] explored a citation-based approach for retrieving studies for inclusion in systematic reviews, which has shown promising results, in particular, substantial increases in precision. Our implementation bases its search and citation-based filtering steps on Belter's approach; we add additional text-based filtering and further automation. Belter's test set also inspired the experiment we describe below. We use 6 of the 14 systematic reviews in Belter's study [16].

2.3 Hybrid Techniques for Information Retrieval

Wolfram [17] emphasized the synergy between information retrieval, bibliometrics, and natural language processing. Adopting and integrating methods across these domains seems natural, especially with the increasing availability of citation data and full-text papers.

Glanzel [18] proposed the use of bibliometrics-aided retrieval and hybrid methods for studying scholarly disciplines. Silva et al. [19] demonstrated the utility of using a hybrid citation and text-based approach for science mapping. However, we were not able to find prior frameworks that combine citation and text-based methods to aid literature review searches. We hypothesize that such a hybrid approach would also be useful in searching for studies for literature reviews.

3 Methods

3.1 Proposed Framework

We propose a three-step framework for searching and filtering articles for literature reviews starting from one or more seed articles.

1. **Select seed article(s)**: Identify 1 or more publications relevant for inclusion in the review to use as seed articles.
2. **Search**: Collect papers connected by citation relationships to at least one seed article.
3. **Filter**:
   (a) **Citation-based**: Remove papers with weak citation relationships to the seed articles.
   (b) **Text-based**: Filter the list of papers using keywords or topics found in the set of all seed articles.

   These two filtering methods can be interchanged or combined.

3.2 A Sample Implementation

Select seed article(s) We use all possible combinations of 1-, 2-, or 3- seed articles.

Search We retrieved the references, citations, co-citing papers, and co-cited papers of all seed articles. These relationships to the seed article are shown in Figure 1. References (RP) are publications cited by a seed article (i.e. usually listed at the end of articles), while citations (CP) are publications that cited a seed article. Co-citing papers (CC) are papers that also cited the same articles that the seed article cited, while co-cited papers (CR) are papers that are also cited by the same articles that cited the seed article. For the rest of this paper, we refer to this set of articles as the citation space of the seed article. We used the Scopus APIs[4] to retrieve the citation spaces.

Filter We implemented a two-step filtering approach by first removing the articles that do not pass our citation-based criteria, then further filtering the list of papers using keywords of the seed articles. The resulting list contains the final set of retrieved papers.

[4] `https://dev.elsevier.com/sc_apis.html`

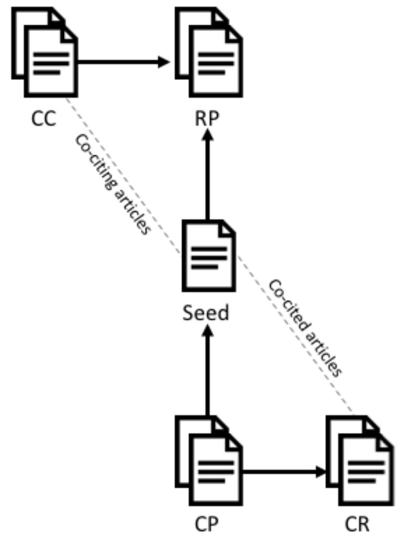

Fig. 1. Citation Space of a Seed Article

Citation-Based Filtering Our citation-based filtering removes all papers that do not meet at least one of these criteria from the retrieved set of papers:

- paper A cites paper B
- paper A is cited by paper B
- paper A shares at least 10% of its references with paper B
- paper A shares at least 10% of its citations with paper B.

We chose 10% as Belter [16] reported promising results with this threshold. Given constraints on our API usage (10,000 abstracts and 20,000 citations per week), filtering by citations enabled us to retrieve a smaller number of abstracts for text-based filtering.

Text-Based Filtering To get the final set of retrieved papers, we filtered the remaining papers based on phrases extracted from the abstracts. We deemed a paper relevant if its abstract contained at least one bigram or trigram phrase found in any of the seed articles' abstracts.

We used the Scopus Abstract Retrieval API to retrieve the abstracts. Then, phrases were extracted from abstracts using an available Python implementation[5] of the Rapid Automatic Keyword Extraction (RAKE) algorithm [20]. RAKE's graph-based approach to extract phrases has been tested on scientific abstracts, and its strength is retaining phrases that include stopwords (enabling it to find complex concepts, e.g. "curse of dimensionality"). We found that

[5] `https://pypi.python.org/pypi/rake-nltk`

the unigram output from RAKE contained uninformative words – verbs (e.g. needed), conjunctive adverbs (e.g. however), and nouns (e.g. studies), so we omitted unigrams.

4 Experiment on the Sample Implementation

4.1 Aim of Experiment

The aim of the experiment was to test our implementation against conventional search procedures used in systematic reviewing. Systematic reviews aim to find all available evidence pertaining to a given research question (i.e. get 100% recall on that question), and typically manually screen search results. Maintaining recall while increasing precision (i.e. get less results for manual screening) would save reviewers time. Therefore, for a given systematic review, our goal was two-fold: (1) retrieve all the designated major publications included in the review and (2) reduce the total number of retrieved papers.

4.2 Ground Truth from Conventional Search Methods

Table 1 shows the list of 6 systematic reviews we used as our ground truth in this experiment [21–26]. All 6 systematic reviews were conducted using Cochrane's rigorous, standardized methods for synthesizing medical evidence [27] and published in 2014 in the Cochrane Database of Systematic Reviews. Each publication indicates the month in which searches were conducted, the number of results screened, and the number of included studies with their references.

A study refers to a group of related publications (e.g. reports about the same clinical trial). For each included study in a review, one publication is designated as the the major publication, which is indicated with an asterisk in the review's list of included studies. For our purposes, the designated major publications are used as both seed articles and retrieval targets.

Searches were conducted in 2013 or 2014, and all 6 reviews reported searching the references of included studies or ongoing trials to find more relevant studies. All reviews reported several search strings in their search strategies, and most of these reviews have different search strategies for different search databases. These search strategies are often refined iteratively, and thus also take a significant amount of time to construct.

Reviewers manually screened 502 to 2762 results and resulted in 6 to 13 included studies, as reported in the published reviews.

We selected these 6 from the 14 systematic reviews used in the study by Belter [16].

4.3 Implementation Details for the Experiment

Code used in the experiment is available at GitHub.[6]

[6] https://github.com/janinaj/lit-review-search

Table 1. List of Reviews Used in our Experiment

Review	Article Title
1	Antibiotic regimens for management of intra-amniotic infection [21]
2	Interventions for preventing and ameliorating cognitive deficits in adults treated with cranial irradiation [22]
3	Co-enzyme Q10 supplementation for the primary prevention of cardio-vascular disease [23]
4	Intermittent self-dilatation for urethral stricture disease in males [24]
5	Electronic cigarettes for smoking cessation and reduction [25]
6	Long-term proton pump inhibitor (PPI) use and the development of gastric pre-malignant lesions [26]

For citation retrieval, we approximated the search date specified in each review by using the year. For example, if the search was reported as conducted in February 2013, we retrieved citations to the seed articles that were published prior to or in the entire year of 2013. It should also be noted that not all of the studies were indexed by Scopus. We return to this point in Section 5.

For each review, the major publications are used as both seed articles and retrieval targets. In each case, the goal was, given some set of major publications as seed articles, to retrieve all of the remaining major publications as retrieval targets. In the following, for simplicity, we refer to the major publications from a review as its studies or included studies.

We tested our method on all possible 1-, 2-, and 3-seed combinations. For instance, review #1 has 10 included studies indexed in Scopus: there are 10 1-seed combinations, 45 2-seed combinations, and 120 3-seed combinations. Consequently, our implementation tested a total of 175 seed combinations for review #1.

5 Results

Table 3 shows the average, maximum, and minimum number of studies (including the seed articles themselves) retrieved using the different numbers of seeds. Using the 3-seed combinations, we were able to retrieve all included studies in 2 of the 6 reviews. Overall, we were able to retrieve all the included studies indexed by Scopus (recall = 100% within Scopus), which covered 48 out of the 55 total included studies (recall = 87.27% overall). However, not all 3-seed combinations could achieve this result. The combinations that were able to achieve this result range from 1% (2/165 combinations) to 75% (3/4 combinations). On average, 40 total included studies (recall = 71.85%) were retrieved for 3-seed combinations.

For 1-seed combinations, the worst-performing seeds were not able to retrieve any of the other included studies in 5 of the 6 reviews. However, if we added a second seed that was also unable to retrieve any other included studies, combined, they could retrieve other included studies. In 3-seed combinations, the worst-performing combinations were still able to retrieve at least 50% of the included studies.

Table 2. Conventional Search Results vs. Average Number of Search Results Retrieved from Scopus in our Experiments

Review	Conventional	1-Seed	2-Seed	3-Seed
1	1,001	86	223	381
2	2,762	84	210	343
3	1,348	146	342	542
4	276	45	113	181
5	594	75	163	250
6	502	140	332	530

Table 3. Number of Included Studies Retrieved from Scopus in our Experiment

Review	Included Studies	Indexed in Scopus	Seeds	Avg	Max	Min
1	11	10	1	4	6	1
			2	6.91	9	4
			3	8.4	**10**	6
2	6	4	1	1.25	2	1
			2	2.83	**4**	2
			3	3.75	**4**	3
3	6	5	1	2	3	1
			2	3.7	**5**	2
			3	4.5	**5**	4
4	11	11	1	4.45	8	1
			2	7.13	9	4
			3	8.4	**11**	6
5	13	10	1	4	7	1
			2	6.36	9	3
			3	7.72	**10**	6
6	8	8	1	3.38	5	3
			2	5.46	**8**	3
			3	6.75	**8**	4

Table 3 shows the average number of search results retrieved for the different numbers of seeds. When the number of seeds is increased from 1 to 2, the retrieved results increased by an average of 24%, and when the number of seeds is increased from 2 to 3 they increased by an additional 16%. Overall the recall compares favorably to conventional search methods: for 3-seed combinations, we reduced the number of results by 34-88% in 5 of the 6 reviews, while the sixth review increased the number of results by 5%; in practice, reductions in results would save reviewers significant time, since the overall number of search results that need to be manually screened is substantially reduced.

6 Discussion & Future Work

Our implementation and results suggest that our proposed framework can be a useful strategy for supporting literature reviews. For 5 out of the 6 reviews, we achieved the same recall as the conventional search methods used in conducting the reviews, but increased the precision. This means that if our method were used in conducting the original reviews, the reviewers could have read and screened fewer of the abstracts while still finding all included publications indexed in Scopus. This reduction could have been substantial – over 2,000 fewer abstracts for review #2. This shows that our framework can help alleviate the time-consuming task of manually screening documents for systematic reviews. In our method, the reviewer only needs to find 1-3 articles that are definitely included in the review. Presumably, these articles are easy to identify and select as seed articles.

Papers are cited for various reasons, and a citation relationship between two documents (i.e. paper A cites paper B) means that they are related in some way. Thus, we can hope that documents are connected in the extended citation network when they have high similarity in their content. This is the case in the included studies for the 6 reviews in our experiment, where all but one included study were connected to at least one other included study in our defined citation space. The only included study that was not connected to any other study also had no citations as of the year of the review and had no reference data in Scopus. These results suggest that navigating the citation network of the seeds may be able to retrieve all the included studies, but the citation space may contain thousands of publications. The largest citation space for a single article used in our experiment contained 107,149 publications, for review #2.

While we can filter the publications in the citation space using only a text-based or only a citation-based method, we hypothesize that a combination of both approaches can filter the list of articles better. This is because the text can provide more details on the topical similarity of documents, while the citation data may capture relations that are not evident in the text. While we have not conducted a comparison with the text-based filtering, we did retrieve fewer results than Belter's citation-based filtering [16] for the 6 reviews, using the same number of seeds.

One of the advantages of this framework is that it can be largely automated. While the seeds need to be selected manually, the retrieval of the citation space and the filtering steps can be done programmatically, assuming that the data is available.

However, one of the limitations of this approach is that it relies on the completeness of the available data. In our experiment, not all of the 55 included studies were in Scopus: only 48 of the included studies were primary documents indexed in Scopus; 6 were secondary documents not indexed in Scopus (but containing title and citation data in Scopus); and 1 document (a meeting abstract published in the appendix of a journal) had no information in Scopus. In addition, of the 48 primary documents in Scopus, 1 had no abstract and 9 were missing reference data.

Further testing how the framework can be integrated into current literature review processes is warranted. While our framework cannot guarantee 100% recall all the time (although this is also the case with conventional methods), we envision that it can be easily integrated in the processes for developing and updating systematic reviews. The framework could also be used to estimate the number of included studies when developing reviews. It has particular promise for finding recent studies when updating systematic reviews, using the previously included studies as seeds. Further, Belter [16] suggested that a citation-based approach may retrieve articles that are not retrieved by the search methods used by the reviewers, so our approach could also be used to seek additional studies for inclusion in a review.

While we have shown that our framework can work well for systematic reviews, we also plan on testing our framework on different kinds of literature reviews, such as scoping reviews. Our future work will explore how different variations in citation space definitions and filtering criteria work for various kinds of literature reviews. We also want to explore how we can use the framework to rank the retrieved publications. This could be tested on the CLEF E-health 2018 Task 2; as in 2017 [28], given a Boolean query and its MEDLINE search results for 20 Cochrane Diagnostic Test Accuracy reviews, systems will rank titles and abstracts and determine a screening threshold. While the ranking of results may not be as important in systematic reviews, it may be very useful for other reviews, such as scoping reviews and state-of-the-art reviews.

7 Conclusion

In this paper, we presented a citation and text-based framework for retrieving publications for literature reviews. Our proposed framework retrieves papers from a set of seed articles through citation relationships, then filters the papers using citation and text-based methods. Our experiment on an implementation of the framework showed that we can achieve up to 100% recall within the limits of the data while improving the precision, but a careful selection of seeds is required. Further testing of the performance and utility of the framework is warranted, but our preliminary results suggest that a hybrid citation- and text-based approach can be a useful strategy in supporting literature reviews.

8 Acknowledgements

Linxi Liu's work on this project was funded by the Illinois Informatics Institute undergraduate research program.

References

1. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11) (2015) 2215–2222
2. Grant, M.J., Booth, A.: A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal **26**(2) (2009) 91–108
3. Ross-White, A., Godfrey, C.: Is there an optimum number needed to retrieve to justify inclusion of a database in a systematic review search? Health Information and Libraries Journal **34**(3) (2017) 217–224
4. Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A.S., Ananiadou, S., Liao, J., Macleod, M.R.: The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error [biorxiv:255760]. (2018)
5. Hoang, L.K., Schneider, J.: Opportunities for computer support for systematic reviewing-a gap analysis. In: iConference 2018 Proceedings, iSchools (2018)
6. Van Eck, N.J., Waltman, L.: Systematic retrieval of scientific literature based on citation relations: Introducing the CitNetExplorer tool. In: International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2014) at the European Conference on Information Retrieval. (2014) 13–20
7. Peroni, S., Shotton, D., Vitali, F.: One year of the OpenCitations corpus. In: International Semantic Web Conference, Springer (2017) 184–192
8. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA. (2007) 697–702
9. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2011) 448–456
10. 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) (2015) 5
11. Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M.H., Liu, H.: A text-mining framework for supporting systematic reviews. American Journal of Information Management **1**(1) (2016) 1–9
12. Kontonatsios, G., Brockmeier, A.J., Przybyła, P., McNaught, J., Mu, T., Goulermas, J.Y., Ananiadou, S.: A semi-supervised approach using label propagation to support citation screening. Journal of Biomedical Informatics **72** (2017) 67–76

13. Pao, M.L.: Term and citation retrieval: A field study. Information Processing and Management **29**(1) (1993) 95–112
14. White, H.D.: Bag of works retrieval: TF*IDF weighting of co-cited works. In: International Workshop on Bibliometric-enhanced Information Retrieval at the European Conference on Information Retrieval. (2016) 63–72
15. Knoth, P., Khadka, A.: Can we do better than co-citations? bringing citation proximity analysis from idea to practice in research article recommendation. In: Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). (2017) 14–25
16. Belter, C.W.: Citation analysis as a literature search method for systematic reviews. Journal of the Association for Information Science and Technology **67**(11) (2016) 2766–2777
17. Wolfram, D.: Bibliometrics, information retrieval and natural language processing: Natural synergies to support digital library research. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). (2016) 6–13
18. Glänzel, W.: Bibliometrics-aided retrieval: Where information retrieval meets scientometrics. Scientometrics **102**(3) (2015) 2215–2222
19. Silva, F.N., Amancio, D.R., Bardosova, M., Costa, L.d.F., Oliveira, O.N.: Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics **10**(2) (2016) 487–502
20. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Mining: Applications and Theory (2010) 1–20
21. Chapman, E., Reveiz, L., Illanes, E., Bonfill Cosp, X.: Antibiotic regimens for management of intra-amniotic infection. Cochrane Database of Systematic Reviews **12** (2014)
22. Day, J., Zienius, K., Gehring, K., Grosshans, D., Taphoorn, M., Grant, R., Li, J., Brown, P.D.: Interventions for preventing and ameliorating cognitive deficits in adults treated with cranial irradiation. Cochrane Database of Systematic Reviews **12** (2014)
23. Flowers, N., Hartley, L., Rees, K.: Co-enzyme Q10 supplementation for the primary prevention of cardiovascular disease. Cochrane Database of Systematic Reviews **2** (2014)
24. Jackson, M.J., Veeratterapillay, R., Harding, C., Dorkin, T.J.: Intermittent self-dilatation for urethral stricture disease in men. Cochrane Database of Systematic Reviews (12) (2014)
25. McRobbie, H., Bullen, C., Hartmann-Boyce, J., Hajek, P.: Electronic cigarettes for smoking cessation and reduction. Cochrane Database of Systematic Reviews **12** (2014)
26. Song, H., Zhu, J., Lu, D.: Long-term proton pump inhibitor (PPI) use and the development of gastric pre-malignant lesions. Cochrane Database of Systematic Reviews **12** (2014)
27. Higgins, J.P., Green, S., eds.: Cochrane handbook for systematic reviews of interventions. Volume 5.1.0. John Wiley & Sons (2011)
28. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Volume CEUR 1866. (2017) 1–29