Browsing citation clusters for academic literature search: A simulation study with systematic reviews

Juan Pablo Bascur[1, 2], Suzan Verberne[2], Nees Jan van Eck[1], Ludo Waltman[1]

[1]Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands
[2]Leiden Institute for Advanced Computer Science, Leiden University, The Netherlands
j.p.bascur.cifuentes@cwts.leidenuniv.nl

Abstract

Our aim is to test if citation clusters can be useful in academic literature search for systematic reviews. We performed an initial offline evaluation using simulated user behaviour on a browsing tool for academic literature search over a set of 17 systematic reviews. To perform the evaluation, we clustered papers in a citation network obtained from the Web of Science database. The clustering solution was a system of seven hierarchical levels of clusters that allowed the simulated user to navigate from larger to smaller clusters. We simulated five user models with different emphasis on precision and recall. We found that citation clusters are more helpful for tasks focused on recall than for precision-oriented tasks. Our future research includes evaluation on a larger set, and a comparison to query search, followed by a study with real users.

Keywords: Academic literature search, Evaluation, Simulation of interaction, Macroscope, Scatter/Gather, Citation network, Document clustering

1 Introduction

Knowledgeworkers need special information retrieval (IR) tools because their IR tasks and practices differ from the general public and from each other [1]. Several special IR tools for academic knowledge workers have been proposed, some of which visualize the search space of literature [2, 3]. These tools are sometimes called *macroscopes*, that is, tools for visualizing big or complex data [4]. Macroscopes facilitate document search through browsing because the visual content and context provides additional information. This information is particularly helpful when Boolean queries are inadequate for an IR task, e.g. if the user does not know the relevant terms to search for. It is difficult to perform offline evaluations of IR macroscopes because there are no standards for the simulation of the stopping point in a browsing task. An analysis of this difficulty of stopping point simulation for document retrieval can be found in the work of Maxwell et al. [5].

To assist knowledge workers, we have prototyped an IR macroscope for academic search literature based on citation clusters. We refer to this tool as SciMacro (Science

Macroscope). SciMacro clusters papers based on their citations and summarizes the content of each cluster. This is different from previous works that cluster papers based on their textual content, like Iris [6], or use citations to find related papers, like PaperPoles [7]. The user can obtain smaller clusters from the papers of a given cluster for a more detailed visualization. As an example, let us consider a fictive user that has a set of documents from a multidisciplinary journal and that wants to know which ones are related to the visualization of big data. The user provides his initial set of documents to SciMacro, which are then clustered into 3 clusters. The descriptors of the clusters are *Mathematics*, *Statistics* and *Physics*. Then, the user selects the *Statistics* cluster and gets 3 smaller clusters from the next clustering level labelled *Visual*, *Modeling* and *Bayesian*. Because these new clusters were created for documents from the *Statistics* cluster, the user knows that they are also related to statistics. Finally, the user selects the *Visual* cluster, and obtains the smaller clusters *Analytic*, *GIS,* and *Big*. Now the user sees the *Big* cluster related to visualization and statistics, enabling him to screen the documents in that cluster.

In this paper, we evaluate the performance of SciMacro at retrieving relevant scientific literature for systematic reviews. Our work is of particular relevance because the potential of IR macroscopes that allow for cluster-based browsing of the complete document set of all scientific literature has not been studied before.

The task goals in our study were to find the relevant literature for 17 systematic reviews (SRs). SRs are review articles that report the relevant literature found by the authors [8, 9]. We used a public test set of these reports to simulate the IR tasks [10]. We obtained the results of the SciMacro IR tasks from a simulation of five user models with different emphasis on precision and recall.

We address the following research questions:

1. What is the potential of SciMacro for finding the relevant literature for SRs?
2. How do the user's preferences affect the performance of SciMacro?

The main contribution of this paper is that it presents the first evaluation of SR search through citation clusters.

2 Related work

The idea of using citations for academic IR browsing is not new: A number of prior works [11, 3] have proposed IR tools that visualize papers based on their citation networks, while others [6, 7] have proposed IR tools that visualize clusters of papers. The tool Citation Gecko [2] visualizes a citation network that expands from given papers, while the work by Haunschild and Marx [12] uses a citation network to find the seminal paper on a topic. On the other hand, text processing can also be used for IR browsing: A number of prior works [13, 14] have proposed IR tools that cluster the semantics of the papers, while others [15, 16] have proposed IR tools that suggest terms for complex Boolean queries.

Our tool, SciMacro, belongs to the academic IR browsing tools that visualize clusters of papers. One possibility is to cluster papers based on textual similarity (see for instance the tool Iris [6]). SciMacro does not use textual similarity, but instead uses citations. Citations represent primarily the intellectual relation between papers, while text may represent (broader) topical clustering.

Following citations is a common strategy for authors of SRs [17]. Some IR tools have been proposed for SRs that classify the relevance of papers combining their citations and text [18]. Other tools reduce the workload of the authors by ranking their search results [19, 20]. These tools are different from SciMacro because they are not based on browsing. For a more complete overview of IR techniques for SRs, we refer to [21].

For non-academic IR browsing, a prominent model is the Scatter/Gather browsing model [22, 23], which has inspired the development of SciMacro. In this browsing model, a cluster of documents is split into smaller clusters, each with their own label. The clusters can also be combined to give the user control over the clustering solution. Scatter/Gather has been previously used for clustering web services [24] and web search results [25]. SciMacro is the first to use the Scatter/Gather model for academic literature search.

During the SIGIR 2010 Workshop on the Simulation of Interaction it was argued that the simulation of different search types (browsing, directed and drifting) requires different user models [26]. We follow up on this work, but instead of simulating query search we simulate the browsing behaviour in citation clusters.

Beyond simulations, Mahdi et al. [27] proposed a framework for evaluating browsing tasks with real users. We also want to highlight the work of Leuski [28], who worked with real users to evaluate a web search tool in which clusters of search results are presented. His work emphasized that clustering the search results gives the user a sense of control over the feedback process, which is a highly valued feature in professional search [1]. In line with this, we designed SciMacro in such a way that the user also has control over the feedback process.

3 Methods

3.1 Search model using queries

We model the search process of the authors of SRs as an IR task. We start from the following idea: When an author of a SR decides to read the full text of a document, based on the abstract and/or title, we consider this document to be relevant. Therefore, we argue that an IR tool should find all documents that the user considers relevant enough to read for a SR. With this consideration in mind, we decided to use the SRs published by the Cochrane Library database [29], which requires authors to report the documents of which they read the full text, regardless of whether they included these documents in the SR or not. We will refer to this set of documents as the *relevant documents* of a particular SR.

3.2 Search model using SciMacro

This search model retrieves the documents of a given cluster (see Section 3.4 for an explanation of the clustering). To select a cluster, we defined the following simulation protocol, similar to a greedy algorithm (see Figure 1 for an example):

1. Select the cluster from level 1 with the highest relevance score (the score is explained below).

2. If the selected cluster has subclusters, obtain the highest score of the subclusters. Otherwise, retrieve the selected cluster.
3. If the highest score of the subclusters is higher than the score of the selected cluster, select the subcluster with the highest score and go back to step 2. Otherwise, retrieve the selected cluster.

The goal of this protocol is to simulate the behaviour of a real user. In our simulation the user has perfect knowledge of the relevance scores of the clusters; this is a common simplification in simulation for evaluation [10]. In a real situation the user has to deduce this knowledge from the cluster labels.[1]

To evaluate this IR task, we need to know the number of relevant retrieved documents, non-relevant retrieved documents and relevant non-retrieved documents from the simulation:

- The total number of retrieved documents equals the number of documents in the retrieved cluster.
- The number of relevant retrieved documents equals the number of relevant documents that are in the retrieved cluster.
- The number of non-relevant retrieved documents equals the total number of retrieved documents minus the number of relevant retrieved documents.
- The relevant non-retrieved documents are the documents of the SR that are not in the retrieved cluster.

With these values we can also calculate the weighted F-score of each cluster:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (1)$$

We created five user models that differ in which F-score they prefer: $F_{0.25}$, $F_{0.5}$, F_1, F_2 or F_4. The different F-scores reflect the different needs of real users: for example, a real user that wants a short overview of a topic will emphasize precision over recall. Lower subscript F-scores emphasize precision, and higher subscript F-scores emphasize recall. F_1 gives an equal weight to precision and recall. For each user model, the relevance score in the cluster selection protocol is given by the weighted F-score. The goal is to maximize this F-score.

[1] Generating informative cluster labels is outside the scope of this paper.

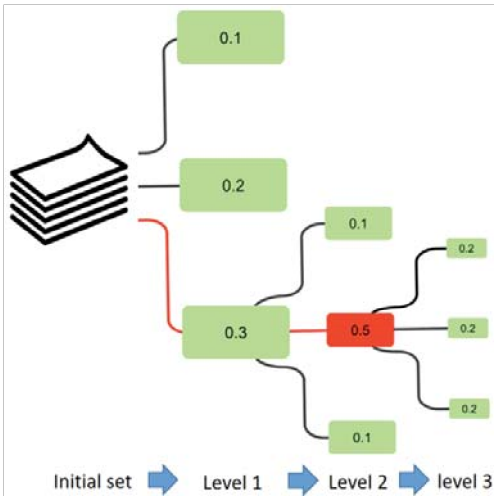

Fig. 1. Example of the cluster selection simulation protocol. Rectangles are clusters. The labels in the rectangles are the scores of the clusters for a given SR. The lines connect clusters with their parent and their children. The red lines indicate the steps that the user follows. The red rectangle is the cluster where the user stops. The steps that the user follows are: **1**- Select the cluster with the highest score at level 1 for the current SR. **2**- Select the cluster with the highest score at level 2 among the subclusters of the current cluster selected at level 1. **3**- Stop because all level 3 subclusters of the current cluster selected at level 2 have a lower score than the current cluster.

3.3 Dataset

We obtained the SRs and their *relevant documents* from the dataset published by Scells et al. [10]. It contains the PubMed ids of 177 randomly selected SRs published by the Cochrane Library plus their *relevant documents*, excluding *relevant documents* that lack PubMed ids. We selected the SRs from this collection that had 10 or more *relevant documents*.

The citation network was created based on the in-house Web of Science database at the Centre for Science and Technology Studies (CWTS) at Leiden University, the Netherlands, which includes papers published since 1980 (as well as a small number of papers published in earlier years). We excluded the documents in this database without PubMed id. We transformed citation links into undirected links (to comply with the requirements of our clustering algorithm). The SRs in the Cochrane Library database include the *relevant documents* in their reference list. Therefore our database also contains these citation links. These links are advantageous for the clustering algorithm, but they would not exist in a real IR task, so we excluded from the citation network all documents published in the same year as the SR and in later years. We selected as our focal year the year with the largest number of SRs. This was the year 2014 with 17 SRs. Therefore our citation network only included documents published before 2014. In the end, the citation network contained over 13.2 million documents and 280.4 million citation links. The SR set contained the 17 SRs published in 2014 and for each SR it contained the *relevant documents* present in the citation network.

3.4 Clustering

We clustered the citation network with the Leiden algorithm [30] based on the methodology developed by Waltman and van Eck [31]. However, they built the clustering hierarchy in a bottom-up manner while we took a top-down approach. Also, they merged small clusters, which we did not do. The use of citation links to cluster documents is common practice in the field of bibliometrics. Textual information, for instance from the titles and abstracts of documents, is also often used to cluster documents. In this paper, we choose to use citation links. The use of textbased co-occurrence links could be explored in future research. We refer to Waltman et al. [32] for a comparison of different approaches for clustering documents. The clustering algorithm maximizes the following quality function:

$$V(x_1, \dots, x_n) = \sum_i \sum_j \delta(x_i, x_j)(a_{ij} - r) \quad (2)$$

In this quality function, i and j are documents, x_i is the cluster of document i, and r is the resolution parameter. a_{ij} equals 1 if there is a citation link between documents i and j, otherwise a_{ij} equals 0. δ equals 1 if i and j are in the same cluster, otherwise δ equals 0. The value of r is given for each level of the clustering hierarchy (see below). The quality function ensures that related documents tend to be assigned to the same cluster. The higher the value of the resolution parameter, the larger the number of clusters and the smaller the number of documents per cluster.

We created a hierarchical clustering consisting of 7 levels. Starting from the highest level, at each level we applied our clustering algorithm to the citation network of the documents of each cluster obtained at the prior level, except at the highest level, where we applied the clustering algorithm to all documents in our citation network. At each lower level we multiplied the value of the resolution parameter by 10 to obtain smaller clusters.

At the highest level, we used the value 10^{-7} for the resolution parameter. This is similar to the value of $8*10^{-8}$ used by Waltman and van Eck [31] in their clustering solution with the lowest granularity. Using the value 10^{-7}, we obtained a clustering solution in which 40% of the documents belonged to the largest cluster and 88% of the documents belonged to the 10 largest clusters. The size of the 15 largest clusters is shown in Figure 2.

In the end, we obtained a nested system of clusters and subclusters that was used for the evaluation. We removed clusters with fewer than 5 documents to avoid precision artefacts.

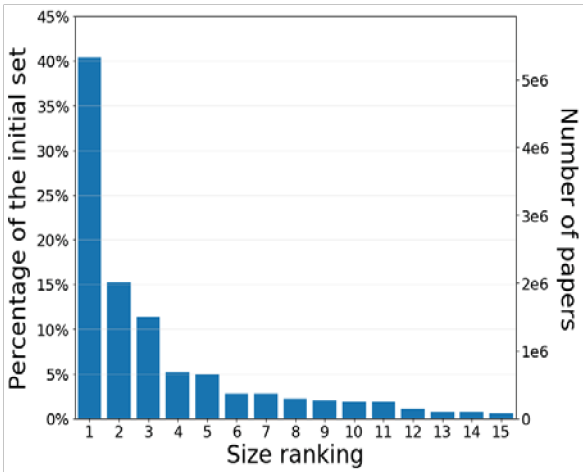

Fig. 2. Size of the 15 largest clusters at level 1 of the clustering hierarchy. X: Size ranking. Y left: Number of documents in a cluster relative to the total number of documents. Y right: Absolute number of documents in a cluster.

4 Results

Figures 3 and 4 show the results of the tasks. Figure 3 shows the F-scores of each SR for the different user models. We observe that the F-score of most SRs increases monotonically from $F_{0.25}$ to F_4. Figure 4 shows the recall and precision values of each SR in the user models $F_{0.25}$ and F_4. Most SRs have higher recall and lower precision in $F_{0.25}$ than in F_4, 5 SRs have the same recall and precision, and 1 has both lower recall and lower precision.

The monotonic increase of the F-score from the $F_{0.25}$ to the F_4 user model observed in Figure 3 suggests that SciMacro is better at tasks that require high recall than tasks that require high precision. This inference is supported by the results observed in Figure 4. Here we observe that most tasks had higher recall than precision and that the user model F_4 greatly increased recall at the expense of relatively little precision, especially when the recall of $F_{0.25}$ was low.

The SR in Figure 4 for which the user model $F_{0.25}$ has both lower recall and lower precision than the user model F_4 is an instance where optimizing for precision was objectively worse than optimizing for recall. We will explore this phenomenon in more detail in the future. The fact that many SRs had the same precision and recall for both user models suggests that the selected cluster was unambiguously the best for these SRs.

In order to draw conclusions from the obtained F-scores, we need a comparison to other (query-based) search methods on the same testset. This is part of our currently ongoing work. We are also extending our evaluation set to include more SR tasks. However, we can make an informed guess of the effectiveness of SciMacro by comparing our results with the results reported by Scells and Zuccon [33]. They replicated the self-reported Boolean query searches of 51 SRs from the same test set as we used, and reported the average F-scores (Table 1), average precision and recall (Table 2) and average number of retrieved and relevant documents (Table 3). The

Fscores and precision are 10 to 20 times higher for SciMacro than for the Boolean query searches, while the recall is 20% to 60% lower, depending on the user model. Therefore, we expect that SciMacro will perform well in precision-focussed tasks when compared with query-based search methods.

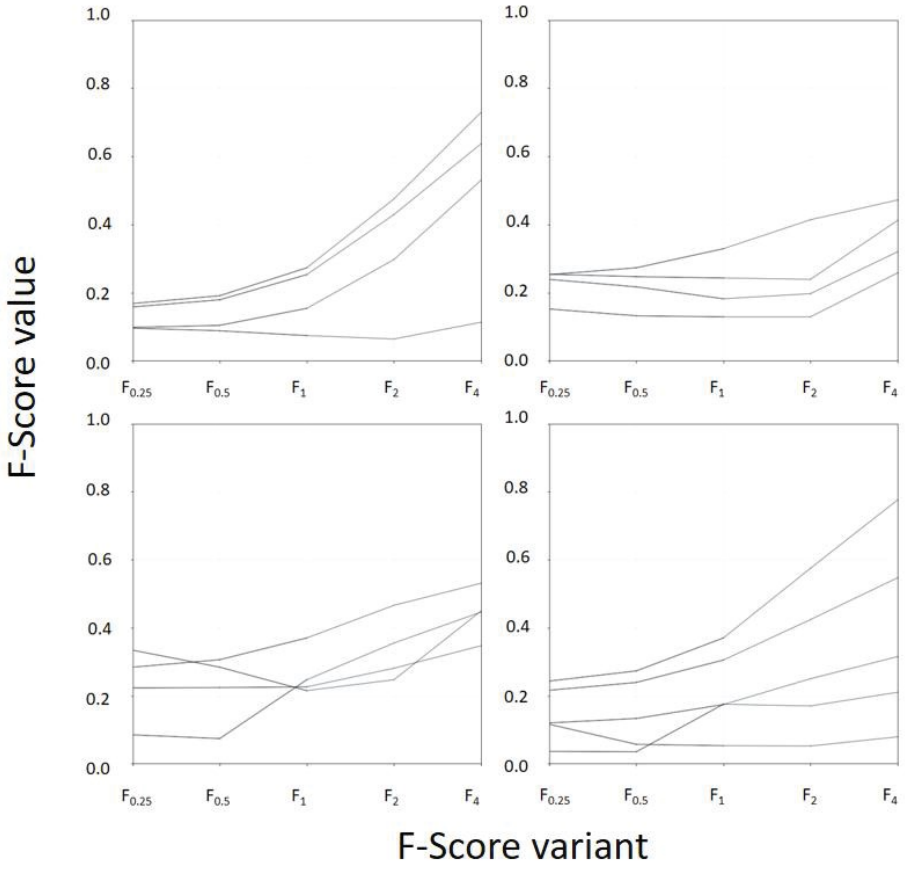

Fig. 3. Performance of SciMacro by F-score. X-axis: F-score variant. Y-axis: F-score value. Quadrants: the quadrants have no specific meaning, they serve to better visualize the 17 SRs. Lines: SRs.

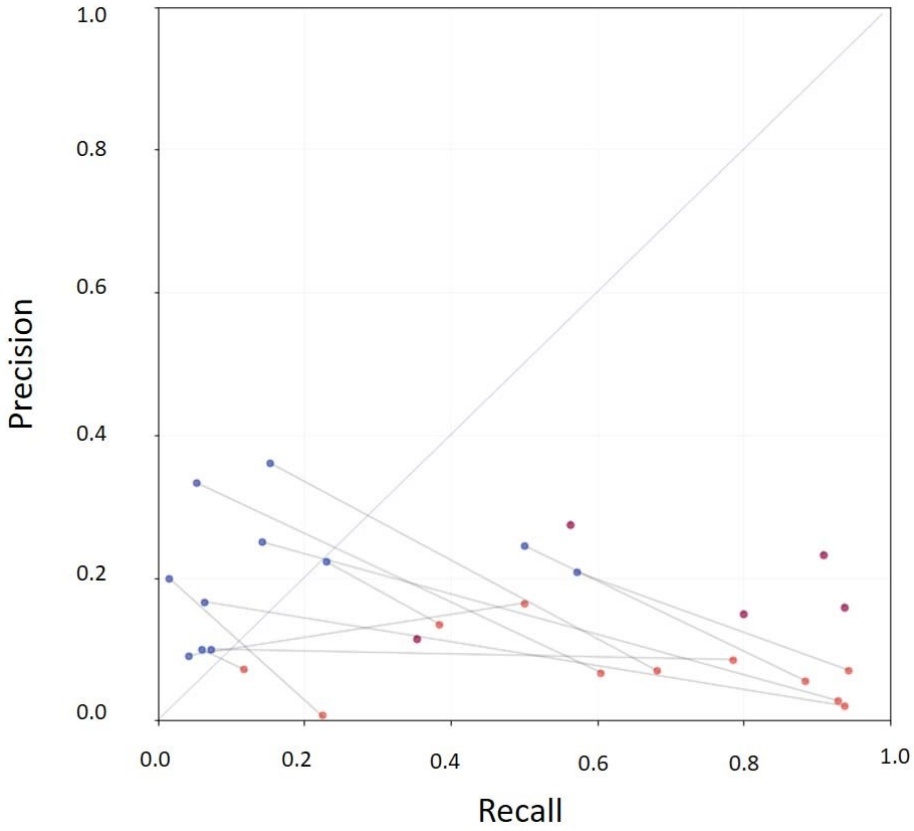

Fig. 4. Performance of SciMacro by recall and precision. X-axis: Recall. Y-axis: Precision. Lines connecting dots: SRs. Diagonal: Points where recall and precision are equal. Colors are user models. Blue: $F_{0.25}$. Red: F_4. Purple: $F_{0.25}$ and F_4 have the same precision and recall.

	$F_{0.25}$	$F_{0.5}$	F_1	F_2	F_3	F_4
Boolean query	-	0.012	0.018	-	0.053	-
SciMacro	0.181	0.180	0.222	0.298	-	0.422

Table 1: Average values of F-score. For SciMacro, the F-score variant is both the user model and the evaluation metric. The Boolean query search values were reported by Scells and Zuccon [33] for 51 SRs of the test set.

	Precision	Recall
Boolean query	0.010	0.815
SciMacro $F_{0.25}$	0.191 ± 0.086	0.323 ± 0.316
SciMacro F_4	0.102 ± 0.073	0.643 ± 0.269

Table 2: Average and standard deviation of precision and recall. The Boolean query search values were reported by Scells and Zuccon [33] for 51 SRs of the test set.

	Relevant documents	Retrieved documents	Relevant retrieved documents
SciMacro $F_{0.25}$	28.8 ± 20.1	37.9 ± 32.6	7.5 ± 7.2
SciMacro F_4		379.6 ± 518.8	17.1 ± 12.9

Table 3: Average and standard deviation of the number of relevant and retrieved documents.

5 Conclusions

The answers to our research questions can be summarized as follows:

What is the potential of SciMacro for finding the relevant literature for SRs? The preliminary results presented in this paper do not allow us to draw strong conclusions. We will need further simulations on a larger benchmark set, a more direct comparison to query search, and ultimately a follow-up user study to answer this question. However, our informed guess, based on the results reported in prior work, is that SciMacro will perform well in precision-focussed tasks when compared with querybased search methods.

How do the user's preferences affect the performance of SciMacro? Our simulation has shown that SciMacro performs better at tasks focused on recall than precision.

At the moment, we are working on evaluating SciMacro using our current simulation setup and comparing it to other academic IR methods. Also, we are currently expanding our evaluation set with more SR tasks to present more rigorous results on the potential of SciMacro.

References

1. Russell-Rose, T., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: A comparison of professional search practices. Inf. Process. Manag. 54, 1042–1057 (2018). https://doi.org/10.1016/j.ipm.2018.07.003
2. Citation Gecko, https://github.com/CitationGecko/gecko-react

3. Choo, J., Lee, C., Kim, H., Lee, H., Liu, Z., Kannan, R., Stolper, C.D., Stasko, J., Drake, B.L., Park, H.: VisIRR: Visual analytics for information retrieval and recommendation with large-scale document data. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 243–244. IEEE, Paris, France (2014)
4. Börner, K.: Plug-and-play macroscopes. Commun. ACM. 54, 60 (2011). https://doi.org/10.1145/1897852.1897871
5. Maxwell, D.M.: Modelling Search and Stopping in Interactive Information Retrieval, (2019)
6. iris.ai, https://iris.ai/
7. He, J., Ping, Q., Lou, W., Chen, C.: PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. J. Assoc. Inf. Sci. Technol. 70, 843–857 (2019). https://doi.org/10.1002/asi.24171
8. Giang, H.T.N., Ahmed, A.M., Fala, R.Y., Khattab, M.M., Othman, M.H.A., Abdelrahman, S.A.M., Thao, L.P., Gabl, A.E.A.E., Elrashedy, S.A., Lee, P.N., Hirayama, K., Salem, H., Huy, N.T.: Methodological steps used by authors of systematic reviews and meta-analyses of clinical trials: a cross-sectional study. BMC Med. Res. Methodol. 19, (2019). https://doi.org/10.1186/s12874-0190780-2
9. Peterson, J., Pearce, P.F., Ferguson, L.A., Langford, C.A.: Understanding scoping reviews: Definition, purpose, and process. J. Am. Assoc. Nurse Pract. 29, 12–16 (2017). https://doi.org/10.1002/2327-6924.12380
10. Scells, H., Zuccon, G., Koopman, B., Deacon, A., Azzopardi, L., Geva, S.: A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17. pp. 1237–1240. ACM Press, Shinjuku, Tokyo, Japan (2017)
11. Nakazawa, R., Itoh, T., Saito, T.: Analytics and visualization of citation network applying topic-based clustering. J. Vis. 21, 681–693 (2018). https://doi.org/10.1007/s12650-018-0483-5
12. Haunschild, R., Marx, W.: Discovering seminal works with marker papers. (2019)
13. Mirylenka, D., Passerini, A.: ScienScan – An Efficient Visualization and Browsing Tool for Academic Search. In: Salinesi, C., Norrie, M.C., and Pastor, Ó. (eds.) Advanced Information Systems Engineering. pp. 667–671. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
14. Open Knowledge Maps: A Visual Interface to the World's Scientific Knowledge., https://openknowledgemaps.org
15. Smith, P.J., Krawczak, D., Shute, S.J., Chignell, M.H.: Bibliographic information retrieval systems: increasing cognitive compatibility. Inf. Serv. Use. 7, 95–102 (1987). https://doi.org/10.3233/ISU-1987-74-502
16. Zhu, Y.: Graph-based Interactive Bibliographic Information Retrieval Systems, (2017)
17. Horsley, T., Dingwall, O., Sampson, M.: Checking reference lists to find additional studies for systematic reviews. Cochrane Database Syst. Rev. (2011). https://doi.org/10.1002/14651858.MR000026.pub2

18. Portenoy, J., West, J.D.: Supervised Learning for Automated Literature Review. In: Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019). p. 9 (2019)
19. Martinez, D., Karimi, S., Cavedon, L., Baldwin, T.: Facilitating Biomedical Systematic Reviews Using Ranked Text Retrieval and Classification. Australasian Document Computing Symposium (ADCS) (2008)
20. Karimi, S., Pohl, S., Scholer, F., Cavedon, L., Zobel, J.: Boolean versus ranked querying for biomedical systematic reviews. BMC Med. Inform. Decis. Mak. 10, (2010). https://doi.org/10.1186/1472-6947-10-58
21. 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. Syst. Rev. 4, (2015). https://doi.org/10.1186/2046-40534-5
22. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. (1992)
23. Pirolli, P., Schank, P., Hearst, M., Diehl, C.: Scatter/gather browsing communicates the topic structure of a very large text collection. In: Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96. pp. 213–220. ACM Press, Vancouver, British Columbia, Canada (1996)
24. Farsandaj, K., Ding, C.: Scatter/Gather browsing of web service QoS data. Future Gener. Comput. Syst. 28, 1145–1154 (2012). https://doi.org/10.1016/j.future.2011.08.020
25. Gong, X., Ke, W., Khare, R.: Studying scatter/gather browsing for web search. Proc. Am. Soc. Inf. Sci. Technol. 49, 1–4 (2012). https://doi.org/10.1002/meet.14504901328
26. Azzopardi, L., Järvelin, K., Kamps, J., Smucker, M.D.: Report on the SIGIR 2010 workshop on the simulation of interaction. ACM SIGIR Forum. 44, 35 (2011). https://doi.org/10.1145/1924475.1924484
27. Mahdi, M.N., Ahmad, A.R., Ismail, R.: Evaluating Search Results in Exploratory Search. Int. J. Eng. Technol. 7, 276 (2018). https://doi.org/10.14419/ijet.v7i4.35.22746
28. Leuski, A.: Evaluating Document Clustering for Interactive Information Retrieval. . K. 13 (2001)
29. Chalmers, I.: The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. Annals of the New York Academy of Sciences 703(1) 156-165 (1993)
30. Traag, V.A., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, (2019). https://doi.org/10.1038/s41598-019-41695-z
31. Waltman, L., van Eck, N.J.: A New Methodology for Constructing a Publication-Level Classification System of Science. J. Am. Soc. Inf. Sci. Technol. 63, 2378–2392 (2012). https://doi.org/10.1002/asi.22748

32. Waltman, L., Boyack, K.W., Colavizza, G., & Van Eck, N.J. (in press): A principled methodology for comparing relatedness measures for clustering publications. Quantitative Science Studies. arXiv:1901.06815 (2019)
33. Scells, H., & Zuccon, G.: Generating better queries for systematic reviews. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 475-484). (2018)