The Impact of Query Refinement on Systematic Review Literature Search: A Query Log Analysis

Harrisen Scells
The University of Queensland
Australia
h.scells@uq.edu.au

Connor Forbes
Institute for Evidence-Based Healthcare, Bond University
Australia
cforbes@bond.edu.au

Justin Clark
Institute for Evidence-Based Healthcare, Bond University
Australia
jclark@bond.edu.au

Bevan Koopman
CSIRO
Australia
bevan.koopman@csiro.au

Guido Zuccon
The University of Queensland
Australia
g.zuccon@uq.edu.au

ABSTRACT

The creation of high-quality medical systematic reviews requires the development of a complex Boolean query to retrieve medical literature. An effective query in this context is critical, as it determines how many documents are to be assessed for inclusion in the resulting systematic review, as all retrieved documents must be screened. Therefore an effective query must balance a reasonable assessment workload with an estimate for how many relevant documents exist for a given topic. Getting this balance correct is naturally a difficult challenge, and there is a certain level of intuition involved in how a query should be formulated and refined. This paper reveals such intuitions and behaviours by analysing the query logs of a specialised tool developed to assist expert searchers in refining complex Boolean queries. These query logs contain unique information that permits a deeper understanding of user behaviour than previous studies. The approximately 6,000 queries collected over one year are available for further analysis at https://github.com/ielab/searchrefiner-logs-collection.

KEYWORDS

Systematic Reviews Automation, Technology Assisted Review, Query Log Analysis

ACM Reference Format:
Harrisen Scells, Connor Forbes, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. The Impact of Query Refinement on Systematic Review Literature Search: A Query Log Analysis. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539813.3545143

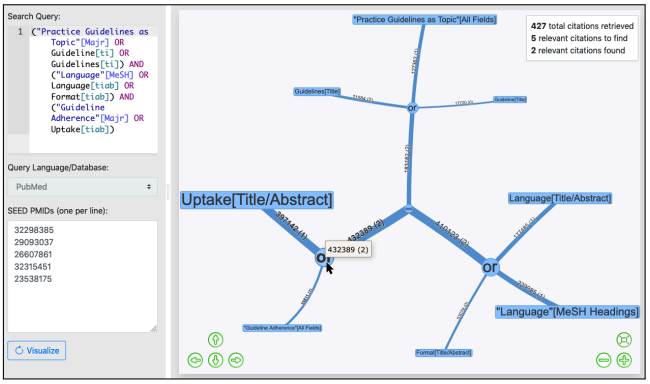


Figure 1: User interface for the searchrefiner [20] tool where the query logs in this paper are recorded. Users can enter a query and a set of document identifiers, and the tool renders a visualisation of the query as well as retrieval statistics. The visualisation is interactive, and users can click to reveal more information about parts of the query. Nodes in the visualisation correspond to the components of the query (i.e., keywords or Boolean operators), and the size of nodes corresponds to the amount of documents retrieved.

1 INTRODUCTION

The formulation of search strategies (i.e., complex Boolean queries) for systematic review literature search involves a lengthy and relatively ambiguous process; experienced librarians, also known as an information specialist, are sometimes needed construct the queries. Boolean queries are needed for a variety of reasons, including compliance with policy or regulations, reproducibility of searches, and for transparency and understandability [12]. Due to the complexity in formulating such search strategies, they are recommended to be formulated in terms of PICO [1] (i.e., population, intervention, control, and outcome). Despite such recommendations, the formulation of queries for this task, like many others, is open-ended, and there have been a variety of techniques for formulating searches [2, 8], and even methods that automatically formulate and refine queries [11, 19, 21, 22]. Although there have been several studies that have retrospectively analysed search strategies for systematic review literature search [5, 16], there are no studies that analyse query formulation and refinement prospectively (i.e., to observe and understand the changes to a query in development).

Due to the complexity of query formulation, and indeed refinement, of complex Boolean queries for systematic review literature

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545143

search, tools for assisting users in these processes are not uncommon in this space [4, 14, 17, 20]. These automation tools may be able to alleviate some of the errors commonly found in these queries, such as spelling mistakes, irrelevant terms, or incorrect logical operators [5, 18], and lowering the time to create a review [3]. However, there have been no studies investigating the query logs (especially for understanding query refinement) from such tools in this domain.

This paper seeks to analyse and understand the behaviour of users refining complex Boolean queries for systematic review literature search by way of interrogating the query logs of searchrefiner [20]. A screenshot of the tool in use is presented in Figure 1. This tool is used for assisting with systematic review literature search query refinement and has approximately 1,000 users signed up from health institutions and universities across Australia, USA, and Europe. The query logs from searchrefiner are somewhat unconventional for two reasons: (1) the users upload several document identifiers (i.e., 'seed studies', documents that are known a priori to be at least somewhat relevant to the systematic review under construction) from the PubMed database (a typical database used to search for medical literature). as a way to weakly validate their search; and (2) the users never see any documents, instead they observe (and indeed are only interested in) the total number of documents and seed studies retrieved.

The contributions of this paper are: (1) an analysis of sessions collected from a live tool that information specialists use to refine their queries, including whether the tool enables users to obtain more effective queries and if there is any relationship between the number of documents retrieved and the number of seed studies retrieved; (2) a case study looking at several sessions in greater detail to identify any strategies that users make use of to refine their queries; and (3) an investigation into the broader procedures that users make use of to achieve their refinement goal by identifying several general patterns that arise across sessions.

2 RELATED WORK

This paper attempts to understand the user behaviour for vastly different kinds of queries that are often seen in typical web search engine query logs [23, 26]. Unlike web query logs, where queries are often only a handful of terms, the queries in this paper are significantly more complex, involving specialised syntax and sometimes upwards of over 100 terms. Examples of these queries are visible in Figure 3. Indeed, several studies have investigated the query logs of the PubMed search engine. These studies range from attempting to understand overall user behaviour [9, 10], to understanding search behaviours of experienced versus novice users [27], or even exploiting the query logs for suggestions [15]. However, these logs contain a mix of term-based and Boolean-based queries. Although the number of query impressions is much larger than the number available in this paper (millions of queries versus thousands), the logs analysed in this paper contain a critical piece of user-submitted information that is not present in the previously considered studies: *seed studies*. These are documents known prior to assessing studies to be included in a systematic review and are considered 'weakly relevant' (i.e., unlikely to be relevant after assessing). Seed studies are typically used to 'validate' a search [1, 2], but have also been used as a methodology for developing search strategies [7, 8, 22]. Unlike prior studies, this paper investigates user behaviour solely in the context of complex Boolean queries. Further, the novel utilisation of seed studies in the analysis permits a deeper understanding of query refinement by knowing how many documents are retrieved and how many (weakly) relevant documents are retrieved. To further understand the place of seed studies and how they fit into the systematic review creation process, Wang et al. [24] provide a suite of use cases. There are also several 'query by document' methods that have been proposed in the literature that utilise seed studies [13, 25].

Given the weak source of relevance that seed studies provide, a strong source of relevance in this context are the studies that are included in the systematic review after assessing the set of retrieved studies. In the IR nomenclature, it is common to refer to an 'effective query' as one that is measured retrospectively in the presence of strong relevance assessments. This definition contrasts with what we refer to in this paper as a 'suitable query', which represents a query where a searcher is confident that, given a set of weak relevance assessments such as seed studies, the query will be effective given the same search task, the same set of documents, but different and stronger relevance assessments. In this paper we are only able to deal with suitable queries, as logs are only collected during the query formulation stage. That is, we do not have access to the actual queries used to retrieve literature for a systematic review nor the relevance assessments for the retrieved literature.

3 LOG ANALYSIS

The query logs from searchrefiner are collected over one year, spanning the start of December 2020 to the end of November 2021. In total, there are 5,962 queries. On average, there were approximately 500 queries submitted per month. Note that the number of queries issued here is relatively small: the logs were acquired from a specialist tool, not a general search engine, which naturally means there are relatively fewer logs than other systems such as web search engines. What follows is the methods for how we split these queries into sessions (and indeed what constitutes a session), the statistics over these sessions, and the insights that can be gleaned from analysing these statistics. The sessions are also recorded in an ad-hoc manner: we did not collect logs from a lab based user study, instead the logs are sourced from a live, production system. Although this limits out ability to infer fine grain behavioural traits about users as we did not survey users of the tool, our setup did allow us to collect a (relatively) large number of sessions, given the highly specific nature of the search tasks. For privacy reasons we do not record identifiable information about users, therefore there is no way to associate sessions that belong to the same user.

3.1 Session Detection

Hagen et al. [6] suggest one taxonomy of sessions from query logs: physical sessions (i.e., the time gap between queries), logical sessions (i.e., consecutive queries for the same information need within the same physical session), and search missions (i.e., disparate logical sessions that connect to the same information need). Within this hierarchy of sessions classification, there exist two kinds of search sessions within the searchrefiner logs: logical sessions and search missions. We use seed studies, not term overlap information,

for identifying a session. Only queries with seed studies and more than a single reformulation are considered when grouping queries into sessions. The median session length (i.e., number of queries in a session) for logical sessions is 5 (10.12 mean), and for search missions is 4 (14.27 mean). The method that group queries into logical sessions is described below. Our analysis does not cover search missions so this description is omitted, although they are included in the data that we release for further analysis.

Logs are grouped into logical sessions by creating a unique hash of the seed studies used, the year, and the day of the year the query was submitted. This ensures that each logical session corresponds to a specific period (i.e., one day of a given year) and a given information need (i.e., the seed studies used for weak validation of the search). The number of queries that have no associated seed studies (i.e., the user never uploaded seed studies) or those with seed studies but no reformulations (i.e., a single, stand-alone query that cannot be grouped into a session) totalled 3,188. The remaining queries are grouped into 274 logical sessions. All of the proceeding analysis is performed on logical sessions.

3.2 Session Analysis

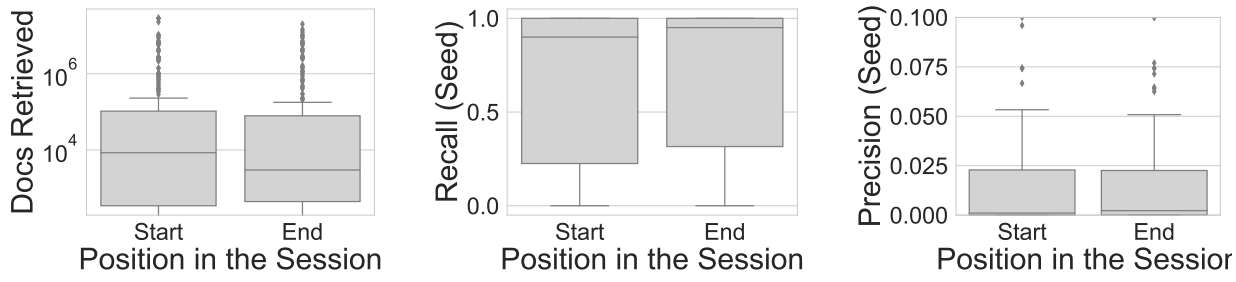
The logs from searchrefiner also permit an interesting analysis that cannot usually be performed on query logs. Since users upload the document identifiers (seed studies) that they are using to weakly validate their search, it is possible to say with some small amount of confidence whether users of the tool are successful in their search task. First, this analysis naïvely assumes that the start and end of a session represent the 'completion' of a refinement or formulation task. In other words, this analysis assumes that the goal (defined in Table 1) of users is to end their session in the tool with a more suitable query than when they started (minimising total documents while maximising seed studies). This is a natural assumption to make, and such a naïve analysis is done so in order to determine whether this assumption holds or if users have other behaviours that make the assumption no longer hold. Figure 2 presents several retrieval results for the logical sessions. Analysing these observations, the three figures suggest that users are generally successful in refining queries (under the naïve assumption above). The mean recall and precision are higher at the end of sessions than at the start. Note that the relatively small increase in precision is due to the large number of documents retrieved in this context: the median increase in precision from 0.001 to 0.002 represents a median reduction from 8,467 retrieved documents down to 2,984 (although these results are not significant under a two-tailed t-test).

Next, analysing the most effective queries in logical sessions, there are 118 out of 274 (approximately 43%) where a query is more effective in retrieving fewer documents but maintaining or increasing the number of seed studies retrieved. Of these, there are 20 sessions where one of the queries in the session is more effective both in retrieving more seed studies and retrieving fewer studies overall (i.e., increasing the precision and recall of the search, given seed studies as a validation set). This observation suggests that in the context of these logs, the query submitted last in a session is not necessarily the most effective. We found this observation interesting because our intuition was that users progressively improve their query in increments (i.e., the start of a session begins with a query with effectiveness, and the measured effectiveness continues to increase at some rate as session length increases). In reality (as we will show in Section 3.3), the effectiveness of some sessions fluctuates dramatically as the session length increases, often to the point where some sessions look as if random modifications are being made to the query. When retrospectively identifying the most effective query in a session and comparing it to the first query submitted in a session, the results are naturally better than those seen in Figure 2, and the improvements in the number of documents retrieved, recall, and precision (with respect to seed studies) are all statistically significant (two-tailed paired t-test, $p < 0.05$). For this comparison, when a more effective query could not be identified, the first query of a session is used.

One possible reason why sessions continue past the point of the 'most effective query' might be because the user refining the query has total recall in mind. In essence, as we assumed above, the general goal of users may not simply be to minimise the number of retrieved documents while maximising the number of retrieved seed studies. That is, just because the query retrieves all seed studies does not necessarily mean that the query retrieves all of the relevant documents for a systematic review (as knowledge of relevant documents requires assessing the entire retrieved set of documents). A strong relationship between these two variables suggests that users continue refining until they meet some threshold of total documents retrieved. Figure 4 presents this relationship for individual queries in sessions that could potentially be the query selected from a session to be used for a systematic review literature search. Although the particular query in a session that will be used for retrieval for a systematic review is unknown, it is possible to make assumptions. Using the two kinds of queries previously mentioned (the last query in a session and the 'most effective' query in a session) for retrieval, in both cases, there is almost no correlation between the number of seed studies retrieved and the number of documents retrieved. Indeed, as shown in Figure 4b, the weak correlation is statistically significant (indicated by *). These observations suggest that the number of seed studies retrieved bears very little on the decision of the user to stop refining their query. Knowing when to stop refining a query does have important, if not immediately apparent, implications for query formulation and refinement. Predicting when to stop refining is a crucial area of research in automatic query refinement [19, 21]. It could also be integrated into tools such as the one used to collect these logs to provide an estimate for query effectiveness.

3.3 Session Case Study

Of the logical sessions where one of the queries is more effective than the first query in a session both in terms of reducing the total number of documents and in terms of increasing the number of seed studies retrieved, two are short enough to be included and visualised here as case studies. The objective of these case studies are to identify any strategies (as defined in Table 1) that users may use to achieve immediate goals within a session. These sessions are presented in Figure 3. The sessions are identified using the first eight characters of their logical hash (Section 3.1). Each query in a session can be identified by the number in a black square the precedes it. The session is essentially presented as a 'diff', where characters in



(a) Docs retrieved. **(b) Recall (seed).** **(c) Precision (seed).**

Figure 2: Retrieval statistics for logical sessions. Note recall and precision are measured against the seed studies that the users uploaded themselves and not documents relevant to a systematic review.

Label	Description
Goal	High-level task that dictates what should be achieved in a session (i.e., retrieve half the total documents while maintaining the number of seed studies). Goals can be immediate (corresponding to strategies) or distant (corresponding to procedures).
Strategy	Actions taken between each query in a session to achieve immediate goals (i.e., the next query should retrieve fewer documents)
Procedure	Overall plan used to achieve a distant goal (i.e., iteratively remove the term that retrieves the most documents but contributes nothing to retrieving all the seed studies).

Table 1: Terms used throughout this paper that describe the fine-grain and course-grain behaviours seen across sessions.

grey indicate no changes from the previous query, characters in black indicate that they have been added to the previous query, and characters that are crossed out indicate that they have been removed from the previous query. In addition to these visualisations of how the queries are modified over time, Figures 3b and 3c also visualise the effectiveness of these two sessions over time. These figures should be utilised by the reader to understand the relationship between the number of documents retrieved and the number of seed studies being retrieved; and are presented in this way rather than using precision and recall because this relationship is what the users of the tool use to verify their queries. Using these two figures, observations can be made about the behaviour of users of the tool, starting with session `bd24f1c7`:

(1) All of the seed studies were identified in query 3, the most effective (in terms of precision and recall) query appeared in position 7 of the session, and the remaining queries were worse, retrieving many documents and fewer seed studies.
(2) The user appears to use an unsupported logical operator (NEAR) in query 9, but is removed in 10. This suggests that further advanced logical operators are a feature that users want to use but are unsupported, or may be accustomed to in other medical literature databases.
(3) Software bugs in the tool may account for some poor performing queries in the sessions. Note the differences between query 15 and 16. Only a space is removed in the final clause, however the space in this position was an edge case in query parsing, resulting in a spike of retrieved documents.

There are also observations that can be made in session `4529ed03`:

(1) As observed above, the last query in a session is often times not the most effective. Indeed in the case of this session, the last query is identical to the first. This suggests that the last query was used to compare it to the second last query in the session.
(2) Users of the tool are editing the complex query directly, not using a structured query editor. As such, mistakes as seen between queries 4 and 5, can appear. A missing) caused the query to retrieve an order of magnitude more documents, and in query 6, this mistake is corrected. Although this particular user found the error, one addition to tools such as the one used here could be to automatically identify such mistakes and notify the user.

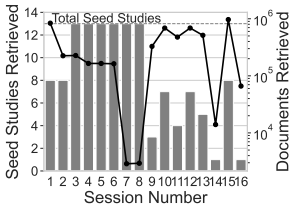
Observing the similarities between the two sessions, and particularly evident in session `bd24f1c7`, it appears that the strategy used to refine these queries was to *add terms* to retrieve seed studies, and *remove terms* to reduce the total number of studies retrieved. This strategy was also the most effective in automatic refinement [21]. Although the sessions cannot fit into this paper due to both the length of the sessions and the size of the queries, Figures 3d and 3e demonstrate these strategies over much longer sessions and much more complex queries. The refinement strategy for session `0ccfc52d` appears to have been to first dramatically reduce the total number of studies retrieved, and then identify new terms that retrieve the remaining seed studies, then continue to reduce the total number of documents from that point. Logical session `504f8ed7` follows a similar story: the user appears to enlarge the total number of studies retrieved until they find all seed studies at which point they continuously reduce the number of studies retrieved.

3.4 Refinement Procedures

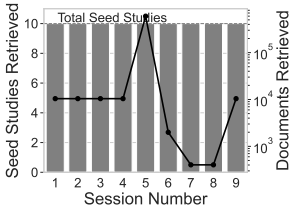
Finally, we look more broadly at the procedures used to achieve certain distant goals (as defined in Table 1). We manually identify common procedures that users take by analysing the session plots from Figure 3. In total, we identify five procedures that users broadly apply to the query refinement task. Examples of sessions that represent each of these procedures are presented in Figure 5. We also identified several sessions where neither the number of seed studies nor the number of retrieved documents changed throughout the

```
2 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode*/OR/underuse*/OR/wasteful* OR overus*
↪ OR/misus* OR unnecess* OR wrong OR unacceptable/OR/poor OR disinvest*) AND (treatment OR therapy OR intervention OR care
↪ OR diagnosis OR test OR screening OR procedure OR surgery OR operation OR referral OR 'health technolog*' OR practice*)
3 ( low-value OR 'low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR wasteful* OR overus* OR
↪ unnecess* OR wrong OR unacceptable OR disinvest*) AND (treatment OR therapy OR intervention OR care OR diagnosis OR test
↪ OR screening OR procedure OR surgery OR operation OR referral OR 'health technolog*' OR practice*)
4 ( low-value OR 'low value' OR 'low added value' OR harmful/OR/ineffectiv* OR inefficient OR wasteful* OR overus* OR
↪ unnecess* OR/wrong OR unacceptable OR disinvest*) AND (treatment OR therapy OR intervention OR care OR diagnosis OR test
↪ OR screening OR procedure OR surgery OR operation OR referral OR 'health technolog*' OR practice*)
5 ( low-value OR 'low value' OR 'low added value' OR ineffectiv* OR inefficient OR wasteful* OR overus* OR unnecess* OR
↪ unacceptable OR disinvest*) AND (treatment OR therapy OR intervention OR care OR diagnosis OR test OR screening OR
↪ procedure OR surgery OR referral OR 'health technolog*' OR peration/OR/referral/OR/'health/technolog*'/OR/practice*)
6 ( low-value OR 'low value' OR 'low added value' OR ineffectiv* OR inefficient OR wasteful* OR overus* OR unnecess* OR
↪ unacceptable OR disinvest*) AND (treatment OR therapy OR intervention OR care OR diagnosis OR test OR screening
↪ OR referral OR 'health technolog*' OR procedure/OR/surgery/OR/referral/OR/'health/technolog*'/OR/practice*)
7 ( low-value OR 'low value' OR 'low added
↪ value'/OR/ineffectiv*/OR/inefficient/OR/wasteful*/OR/overus*/OR/unnecess*/OR/unacceptable/OR/disinvest*) AND
↪ (treatment OR therapy OR intervention OR care OR diagnosis OR test OR screening OR referral OR 'health technolog*' OR
↪ practice*)
8 ( low-value OR 'low value' OR 'low added value') AND (surgery OR procedure OR operation OR treatment OR therapy OR
↪ intervention OR care OR diagnosis OR test OR screening OR referral OR 'health technolog*' OR practice*)
9 (/low-value/OR/'low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR was
↪ teful* OR overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) NEANDR/4
↪ (surdiagery/OR/procedure/OR/operation/OR/treatment/OR/therapy/OR/intervention/OR/care/OR/dias OR test OR screeningnos
↪ is/OR/test/OR/screening//OR/referral/OR/'health/technolog*'/OR/practice*)
10 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) ANDEAR/4
↪ (diagnosis/OR/test/OR/screatment OR therapy OR ingtervention OR care )
11 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND
↪ (dtreatment/OR/therapy/OR/intervention/OR/cargnosis OR test OR screening )
12 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND (diagnosis/OR/test/OR/screatmenting
↪ OR therapy OR intervention OR care )
13 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND
↪ (treatment/OR/therapocedure OR surgery OR intervention/OR/careration )
14 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND (procedure/OR/surgery/OR/operation/)
15 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND (r'hefrralth technolog*' )
16 ('low value' OR 'low added value' OR harmful OR ineffectiv* OR inefficient OR outmode* OR underuse* OR wasteful* OR
↪ overus* OR misus* OR unnecess* OR wrong OR unacceptable OR poor OR disinvest*) AND ('health technolog*'/)
```

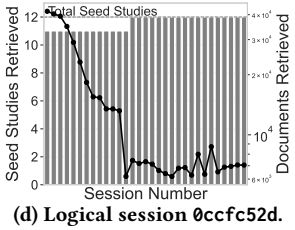
(a) Addition and removal history for logical session `bd24f1c7`.



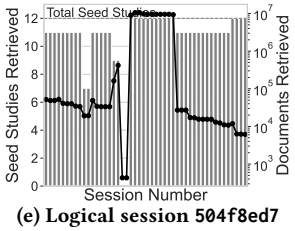
(b) Logical session `bd24f1c7`.



(c) Logical session `4529ed03`



(d) Logical session `0ccfc52d`.

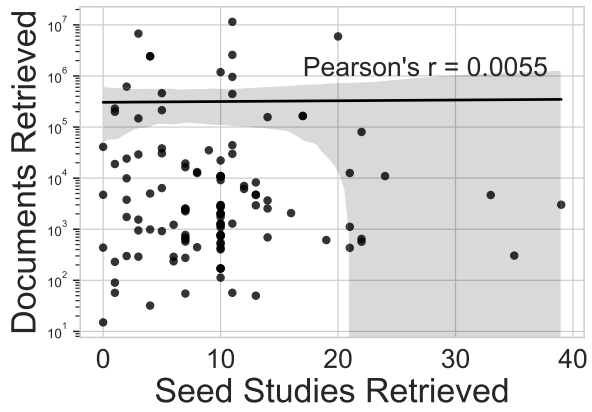


(e) Logical session `504f8ed7`

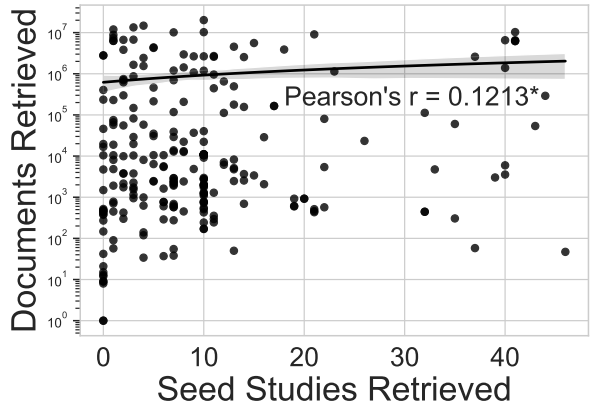
```
3 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Blackheads[tiab]/OR/Whiteheads[tiab]/OR/Pimples[tiab]/OR/Vulgaris[tiab] OR Lesion[tiab]) AND ("Phototherapy"[Mesh] OR
↪ "Blue light"[tiab] OR Phototherapy[tiab] OR Phototherapies[tiab] OR "Photoradiation therapy"[tiab] OR "Photoradiation Therapies"[tiab] OR "Light
↪ Therapy"[tiab] OR "Light Therapies"[tiab] OR LED[tiab] OR Diode[tiab]))
4 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Vulgaris[tiab] OR Lesion[tiab]) AND ("Phototherapy"[Mesh] OR "Blue light"[tiab] OR Phototherapy[tiab] OR
↪ Phototherapies[tiab] OR "Photoradiation therapy"[tiab] OR "Photoradiation Therapies"[tiab] OR "Light Therapy"[tiab] OR "Light Therapies"[tiab] OR LED[tiab] OR
↪ Diode[tiab]))
5 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Vulgaris[tiab] OR/Lesion[tiab])/AND ("Phototherapy"[Mesh] OR "Blue light"[tiab] OR Phototherapy[tiab] OR
↪ Phototherapies[tiab] OR "Photoradiation therapy"[tiab] OR "Photoradiation Therapies"[tiab] OR "Light Therapy"[tiab] OR "Light Therapies"[tiab] OR LED[tiab] OR
↪ Diode[tiab]))
6 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Vulgaris[tiab]) AND ("Phototherapy"[Mesh] OR "Blue light"[tiab] OR Phototherapy[tiab] OR Phototherapies[tiab] OR
↪ "Photoradiation therapy"[tiab] OR "Photoradiation Therapies"[tiab] OR "Light Therapy"[tiab] OR "Light Therapies"[tiab] OR LED[tiab] OR Diode[tiab]))
7 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Vulgaris[tiab]) AND ("Blue light"[tiab] OR Phototherapy"[Mesh]/OR/"Blue/light"[tiab] OR
↪ Phototherapies[tiab] OR "Photoradiation therapy"[tiab]/OR/Phototherapies[tiab] OR "Photoradiation Therapies"[tiab] OR "Light Therapy"[tiab] OR
↪ "LPhotoradiaghtion Therapies"[tiab]/OR/"Light/Therapy"[tiab]/OR/"Light/Therapies"[tiab]/OR/LED[tiab]/OR/Diode[tiab]))
8 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Vulgaris[tiab]) AND ("Blue light"[tiab] OR Phototherapy[tiab] OR Phototherapies[tiab] OR "Photoradiation therapy"[tiab]
↪ OR "Photoradiation Therapies"[tiab] OR "Light Therapy"[tiab] OR "Light Therapies"[tiab]))
9 (("Acne Vulgaris"[Mesh] OR Acne[tiab] OR Blackheads[tiab] OR Whiteheads[tiab] OR Pimples[tiab] OR Vulgaris[tiab] OR Lesion[tiab]) AND
↪ ("Phototherapy"[Mesh] OR "Blue light"[tiab] OR Phototherapy[tiab] OR Phototherapies[tiab] OR "Photoradiation therapy"[tiab] OR "Photoradiation
↪ Therapies"[tiab] OR "Light Therapy"[tiab] OR "Light Therapies"[tiab] OR LED[tiab] OR Diode[tiab]))
```

(f) Addition and removal history for logical session `4529ed03`.

Figure 3: Exemplar sessions where one of the queries in the session is more effective than the first query in the session. Figures 3a and 3f present the retrieval history for the two sessions (green indicates additions, red with strikeout indicates removals). The right hand side plots have two shared y-axes: the left y-axis indicates the number of seed studies retrieved and is represented as a bar plot; the right y-axis indicates the total number of documents retrieved and is represented as a line plot. A horizontal dashed line also indicates the total number of seed studies. Figures 3d and 3e present the retrieval history of two longer sessions that could not be included for space reasons.



(a) Subset of 118 logical sessions.



(b) All 274 logical sessions.

Figure 4: Relationship between number of documents retrieved and number of seed studies retrieved for different 'final queries' in each session. Figure 4a are the 118 queries from logical sessions where the query is 'more effective' than the first query in the session. Figure 4b are the last queries submitted in a session across all 274 logical sessions.

session. On closer inspection, the queries in these sessions either do not change, or there are only superficial changes, like adding or removing a set of outermost braces, which does not impact the search. It is unclear what the intention of these sessions are, and we leave further investigation to future work.

Focusing now on the five refinement procedures, the first procedure (DecreaseRet-MaintainSeed, Figure 5a) is what we expected the majority of sessions to look like. Here, a query that already retrieves all of the seed studies is issued and is refined such that the total number of documents retrieved only decreases or stays the same. We consider this procedure to capture the essence of query refinement in the truest sense, as the effectiveness of the query (at least in terms of precision) only increases while maintaining recall (given the weak signal of relevance that a seed study provides).

The next procedure (IncreaseRet-IncreaseSeed, Figure 5b) is interesting to us as it essentially achieves the opposite effect of the previous procedure. In this procedure, the user begins with a query that almost achieves total recall over the seed studies. Then, through several iterations, the user modifies the query such that the total number of documents retrieved increases in an attempt to increase the number of seed studies retrieved (effectively increasing recall while trading off precision). Why this procedure is interesting is that the sessions end almost immediately once more or all of the seed studies are found. We believe this procedure to be more exploratory, where the intent is to determine an upper bound on how many studies must be retrieved to find the most seed studies.

In contrast, IncreaseDecreaseRet-IncreaseSeed (Figure 5c) could be considered the combination of IncreaseRet-IncreaseSeed followed by the DecreaseRet-MaintainSeed. The session begins with a query that does not retrieve all of the seed studies, followed by a large spike in the number of retrieved studies to retrieve more seed studies, and ending with a gradual decrease of retrieved studies while maintaining the newly retrieved seed studies.

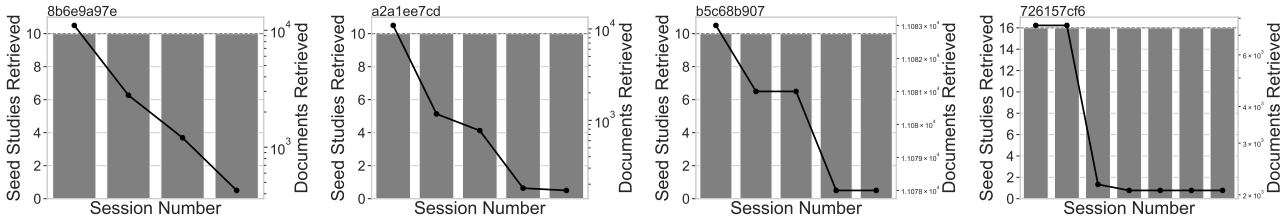
The fourth procedure we identified (ExploreRet, Figure 5d) is more unusual and less clear why users undertake it. Unlike the first three procedures where there is an intuitive goal to the procedure, here, we observe the total number of documents fluctuating up and down, but between one or two points. The recorded number of retrieved documents changes either periodically or stochastically, i.e., either every second query returns to the previously recorded retrieved documents, or there may be stretches of queries of similar effectiveness before returning to a previous number of retrieved documents. One explanation for this fluctuation in retrieved documents may be that the user submits the initial query multiple times within the same session and applies some minor change each time.

The last procedure that we identified (CombineProcedures, Figure 5e) is one where we believe a user to be iteratively combining different procedures. We note that such combination procedures only ever occur in longer sessions, where there is room for more complex combinations of procedures to arise. Contrast these long sessions with the relatively shorter sessions discussed above.

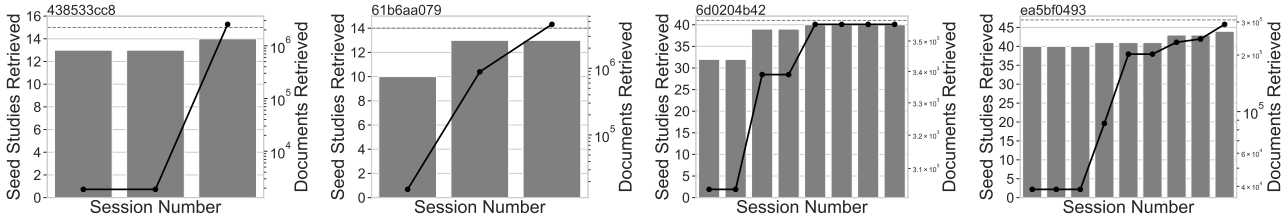
These procedures reveal the many nuanced tasks users have in searchrefiner. The diverse range of tasks users undertake using this tool demonstrates the difficulty of developing automatic query refinement methods for systematic review literature search. However, we believe that these procedures may be used to develop automatic methods that mimic or approximate them.

4 CONCLUSION

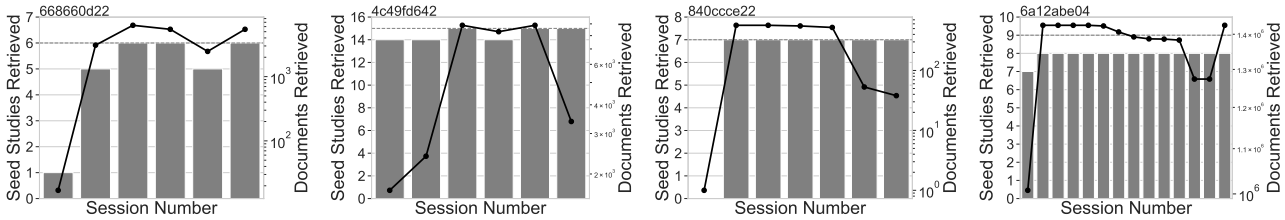
The refinement of complex Boolean queries for a systematic review literature search is highly nuanced. The behaviours observed in these logs are envisioned to inform and develop automatic methods to assist in query refinement. For example, the general strategy of removing and adding terms to achieve a desired result is a clear direction for future work in predicting the terms to add or remove for a query. The other important finding from this study is that the most effective query in terms of seed studies and total documents retrieved may not be the most suitable query. Determining when to stop refining a query is key for automatic refinement methods, and automatically predicting when to stop is another area for future



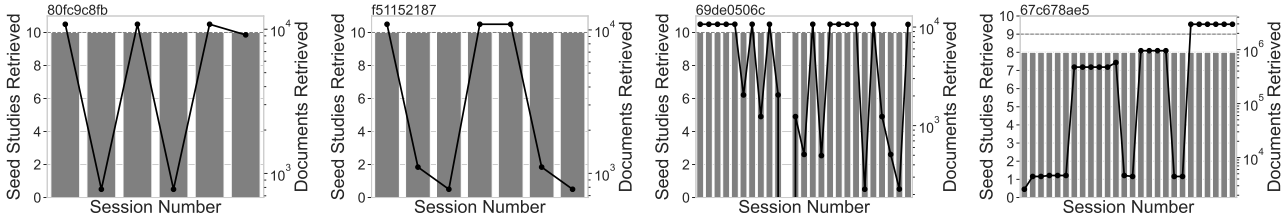
(a) DecreaseRet-MaintainSeed. Iteratively reduce the total documents retrieved while maintaining seed studies.



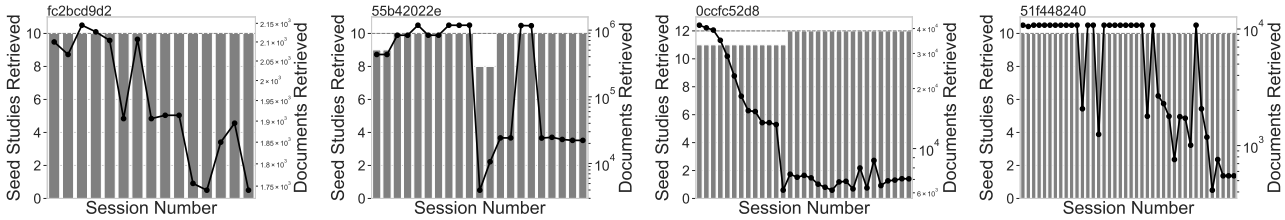
(b) IncreaseRet-IncreaseSeed. Iteratively retrieve more documents to retrieve more seed studies.



(c) IncreaseDecreaseRet-IncreaseSeed. Retrieve documents to find seed studies, then reduce documents and maintain seed studies.



(d) ExploreRet. Fluctuate between low and high numbers of retrieved documents.



(e) CombineProcedures. Iteratively combines multiple procedures.

Figure 5: Different procedures identified by analysing the changes in number of documents retrieved and number of seed studies retrieved in similar sessions. As in the same plots as Figure 3, the dashed horizontal line represents the total number of seed studies. Each plot has a shared y-axis that corresponds to the number of seed studies retrieved (left, represented as bars) and the total documents retrieved (right, represented as a line). The logical has for each of the sessions is also visible in the top right of each plot, if readers are interested in further analysing these specific sessions in the collection.

work. The naïve approach taken in this paper demonstrates that this is a difficult challenge.

In addition to the findings and observations of this study being able to feed into automatic methods, it is envisioned that they also find their way into tools for practitioners to continue to use. Indeed, the general strategy for refining queries tends to be the removal and addition of terms. A tool that suggests which terms to remove or add in order to achieve a desired result is a clear direction for future work. Many sessions in this study indicate that for this task, simply finding all seed studies and the least amount of documents is not acceptable for 'total recall'. Users are often not satisfied to stop once they find a query that retrieves all seed studies. Instead, they continue to refine their query, which may now retrieve fewer or more documents. Estimating when to stop refining queries is an important area of research for automatic query refinement and for tools such as the one used in this study to assist users in deciding when to stop refining.

Overall, our observations indicate that the use of seed studies is clearly important for the formulation of queries as it provides a weak signal of relevance to the user. However, in terms of when to stop refining the query, we find that seed studies bear little on this decision (Section 3.2). Digging deeper into the refinement process by analysing a handful of sessions where there is a more effective query than what was initially submitted revealed an almost chaotic process whereby the effectiveness of a query (as measured by seed studies) would fluctuate dramatically (Section 3.3). In addition to identifying a general strategy that is used throughout the sessions for refining queries, i.e., removing and adding terms, we also identified several general refinement procedures that users took to achieve certain broad goals such as increasing precision or recall (Section 3.4). Understanding these procedures allow us to identify the different use cases that users may use the tool for, and may reveal further insight for developing automatic methods that mimic these procedures.

In addition to the insights that have been gleaned from this paper that can feed into future methods and tools, there is also a clear direction for future work to perform lab-based user studies that survey users to understand not only more nuanced strategies for refining queries, and to better understand and identify new procedures that users use to refine their queries. Longer term lab-based studies would also allow follow-up studies to allow one to evaluate the suitable query developed by the information specialist in terms of the included studies, or strong relevance assessments (as opposed to the weak relevance assessments of seed studies), to measure the true effectiveness of the query. One such IR collection that not only contains topics (i.e., queries, relevance assessments, etc.) but also seed studies [24] already provides early insights into the relationship between seed studies and effective queries. However, there is much more work that can be done in this space to better exploit and understand the uses of seed studies in the context of systematic review literature search.

ETHICS & ACKNOWLEDGEMENTS

Ethics for this study was approved by The University of Queensland, Faculty of Engineering, Architecture and Information Technology, Low & Negligible Risk Ethics Sub-Committee, approval number 2019002743.

This research is partially funded by the Australian Research Council Discovery Projects program DP210104043.

We wish to thank the anonymous reviewers for their insightful and constructive feedback on this paper. We would also like to thank Ahmed Mourad and Joel Mackenzie for suggesting improvements to early drafts of this paper.

REFERENCES

[1] Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.
[2] Justin Clark. 2013. Systematic Reviewing. In *Methods of Clinical Epidemiology*, Gail M. Williams Suhail A. R. Doi (Ed.).
[3] Justin Clark, Paul Glasziou, Chris Del Mar, Alexandra Bannach-Brown, Paulina Stehlik, and Anna Mae Scott. 2020. A Full Systematic Review Was Completed in 2 Weeks Using Automation Tools: A Case Study. *Journal of clinical epidemiology* 121 (2020), 81–90.
[4] Justin Michael Clark, Sharon Sanders, Matthew Carter, David Honeyman, Gina Cleo, Yvonne Auld, Debbie Booth, Patrick Condron, Christine Dalais, Sarah Bateup, et al. 2020. Improving the Translation of Search Strategies Using the Polyglot Search Translator: A Randomized Controlled Trial. *Journal of the Medical Library Association* 108, 2 (2020), 195.
[5] Su Golder, Yoon K Loke, and Liliane Zorzela. 2014. Comparison of Search Strategies in Systematic Reviews of Adverse Effects to Other Systematic Reviews. *Health Information & Libraries Journal* 31, 2 (2014), 92–105.
[6] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From Search Session Detection to Search Mission Detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*. Citeseer, 85–92.
[7] Elke Hausner, Charlotte Guddat, Tatjana Hermanns, Ulrike Lampert, and Siw Waffenschmidt. 2015. Development of Search Strategies for Systematic Reviews: Validation Showed the Noninferiority of the Objective Approach. *Journal of clinical epidemiology* 68, 2 (2015), 191–199.
[8] Elke Hausner, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. 2012. Routine Development of Objectively Derived Search Strategies. *Systematic reviews* 1, 1 (2012), 19.
[9] Jorge R. Herskovic, Len Y. Tanaka, William Hersh, and Elmer V. Bernstam. 2007. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *Journal of the American Medical Informatics Association* 14, 2 (March 2007), 212–220.
[10] Rezarta Islamaj Dogan, G. Craig Murray, Aurélie Névéol, and Zhiyong Lu. 2009. Understanding PubMed User Search Behavior through Log Analysis. *Database* 2009 (Jan. 2009), bap018.
[11] Youngho Kim, Jangwon Seo, and W Bruce Croft. 2011. Automatic Boolean Query Suggestion for Professional Search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[12] Ian A. Knight, Max L. Wilson, David F. Brailsford, and Natasa Milic-Frayling. 2019. Enslaved to the Trapped Data: A Cognitive Work Analysis of Medical Systematic Reviews. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, Glasgow Scotland UK, 203–212.
[13] Grace E. Lee and Aixin Sun. 2018. Seed-Driven Document Ranking for Systematic Reviews in Evidence-Based Medicine. In *Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 455–464.
[14] Hang Li, Harrisen Scells, and Guido Zuccon. 2020. Systematic Review Automation Tools for End-to-End Query Formulation. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25–30.
[15] Zhiyong Lu, W. John Wilbur, Johanna R McEntyre, Alexey Iskhakov, and Lee Szilagyi. 2009. Finding Query Suggestions for PubMed. *AMIA Annual Symposium Proceedings* 2009 (2009), 396–400.
[16] Melissa L Rethlefsen, Ann M Farrell, Leah C Osterhaus Trzasko, and Tara J Brigham. 2015. Librarian Co-Authors Correlated with Higher Quality Reported Search Strategies in General Internal Medicine Systematic Reviews. *Journal of clinical epidemiology* 68, 6 (2015), 617–626.
[17] Tony Russell-Rose and Philip Gooch. 2018. 2dSearch: A Visual Approach to Search Strategy Formulation. In *Proceedings of the 1st Biennial Conference on Design of Experimental Search and Information Retrieval Systems*.
[18] Margaret Sampson and Jessie McGowan. 2006. Errors in Search Strategies Were Identified by Type and Frequency. *Journal of Clinical Epidemiology* 59, 10 (2006), 1057–e1.
[19] Harrisen Scells and Guido Zuccon. 2018. Generating Better Queries for Systematic Reviews. In *Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484.

[20] Harrisen Scells and Guido Zuccon. 2018. Searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews. In *Proceedings of the 27th International Conference on Information and Knowledge Management*. 1939–1942.
[21] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean Query Refinement for Systematic Review Literature Search. In *Proceedings of the 28th World Wide Web Conference*. 1646–1656.
[22] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2020. A Comparison of Automatic Boolean Query Formulation for Systematic Reviews. *Information Retrieval Journal* (2020), 1–26.
[23] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. 2007. Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*. ACM Press, Amsterdam, The Netherlands, 151.
[24] Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[25] Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. 2022. Seed-Driven Document Ranking for Systematic Reviews: A Reproducibility Study. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 686–700. https://doi.org/10.1007/978-3-030-99736-6_46
[26] Steve Wedig and Omid Madani. 2006. A Large-Scale Analysis of Query Logs for Assessing Personalization Opportunities. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06*. ACM Press, Philadelphia, PA, USA, 742.
[27] Illhoi Yoo and Abu Saleh Mohammad Mosa. 2015. Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users. *JMIR Medical Informatics* 3, 3 (July 2015), e3740.