Query Variation Performance Prediction for Systematic Reviews

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

Leif Azzopardi
Strathclyde University
Glasgow, Scotland
leif.azzopardi@strath.ac.uk

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

ABSTRACT

When conducting systematic reviews, medical researchers heavily deliberate over the final query to pose to the information retrieval system. Given the possible query variations that they could construct, selecting the best performing query is difficult. This motivates a new type of query performance prediction (QPP) task where the challenge is to estimate the performance of a set of query variations given a particular topic. Query variations are the reductions, expansions and modifications of a given seed query under the hypothesis that there exists some variations (either generated from permutations or hand crafted) which will improve retrieval effectiveness over the original query. We use the CLEF 2017 TAR Collection, to evaluate sixteen pre and post retrieval predictors for the task of Query Variation Performance Prediction (QVPP). Our findings show the IDF based QPPs exhibits the strongest correlations with performance. However, when using QPPs to select the best query, little improvement over the original query can be obtained, despite the fact that there are query variations which perform significantly better. Our findings highlight the difficulty in identifying effective queries within the context of this new task, and motivates further research to develop more accurate methods to help systematic review researchers in the query selection process.

1 INTRODUCTION

Systematic reviews form the cornerstone of evidence based medicine, aiming to answer complex medical questions based on all evidence currently available [20]. The process of creating a systematic review follows a strict protocol, where medical health professionals, alongside information specialists such as librarians, are typically required to formulate a search strategy *a priori* that defines the criteria for: (*i*) what will be included and excluded in the review (i.e. relevance criteria); (*ii*) what sources will be used (i.e. databases that will be searched); and (*iii*) what queries will be issued to each source. Depending on the protocol, the queries are often submitted for review to a panel of peers to ensure that the queries are appropriate, exhaustive, and not biased. However, this process is somewhat subjective, mainly driven by expertise rather than evidence. In this paper, we consider the task of *Query Variation Performance Prediction* (QVPP) as a means to help quantify the quality of queries formulated for systematic reviews. It is envisaged that QVPP could help both the medical health professionals in selecting effective queries, and the panel when reviewing such queries. Unlike the traditional Query Performance Prediction (QPP) task [5], where the performance of queries across *different* topics is estimated, QVPP attempts to estimate the performance of queries for the *same* topic. Thus, it is an open question how well current performance predictors are suited in this new context. In this paper, we investigate the applicability of QPPs for estimating the effectiveness of query variations in the context of medical systematic reviews. We examine 16 performance predictors (12 pre-retrieval, and 4 post-retrieval) to determine which QPPs provide the best estimates of effectiveness. Our experiments are conducted using the CLEF 2017 Technology Assisted Review (TAR) track collection [10]. We compare the ability of predictors to predict the performance across topics (i.e. QPP task) and effectiveness of predictors for identifying query variations that are better than the original query (i.e. QVPP task). For this study, we felt it was appropriate to use the title as the seed query, rather than the Boolean query associated with each topic, because the QPP methods employed were not designed to handle Boolean semantics. We leave this direction for further work.

2 BACKGROUND AND RELATED WORK

2.1 Search for Systematic Reviews

The processes of compiling, maintaining and updating a systematic review is lengthy and methodical [22]. Once a research question has been formed, the reviewers define search strategies outlining what should and should not be included in the review. This strategy is used to define the Boolean query to retrieve citations (i.e. the title, abstract, and meta-data) from medical databases. It is common for systematic reviews to use Boolean queries which are formulated iteratively with the help of information specialists (i.e. librarians). Once an appropriate search strategy is formulated, the reviewers screen all of the retrieved citations to determine which ones can potentially be included in the review. The query is central to a systematic review as it defines how many citations the reviewers must screen for inclusion, impacting the time and cost of the review [11, 20]. The selection of a query that optimises retrieval effectiveness can reduce the time and effort needed for reviewers to spend screening citations. Most work has focused on devising retrieval methods specific to systematic reviews to minimise the number of citations to screen [4, 11, 19]. In the context of systematic

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210078

reviews, however, there has been no work that explored identifying potentially better queries (or, in the case of this work, variations), though this has been explored within other search tasks. For example, among many others, Ozertem et al. [16] investigated a learning to rank approach for query suggestion for web queries. In the medical domain, Koopman et al. have explored the effectiveness of query variations with respect to the task of clinical trial retrieval and used regression on QPPs to determine the effectiveness of queries [13]. Previous work by Karimi et al. [11] has shown that non-Boolean queries can outperform typical Boolean queries (i.e those used in systematic reviews). Additionally, because it is unclear how to apply current QPP methods to Boolean queries, our investigation focuses on the effectiveness of QPPs to identify potential query variations within the context of searching systematic reviews with non-Boolean queries.

2.2 Query Performance Prediction

Broadly speaking, QPP methods can be considered as either (*i*) pre-retrieval, or (*ii*) post-retrieval. **Pre-retrieval predictors** use statistics about queries and the collection in order to make a prediction. **Post-retrieval predictors** use the results, such as the retrieval status value and rank of documents to make a prediction about the effectiveness of a query. Both pre-retrieval and post-retrieval QPPs are evaluated in the same way: the scores of the QPPs are compared against a retrieval effectiveness evaluation score such as mean average precision (MAP). In standard QPP evaluation, the linear coefficient correlation (Pearson's r) is computed between the scores of a QPP measure and a retrieval evaluation measure. Strong correlations lead to the hypothesis that they are effective. Hauff et al. [6] have found that this correlation is often insufficient and does not accurately reflect the performance of a query for a retrieval system, or that the correlation is high only for specific tasks; however, no solutions to this problem have been proposed. While the main task of this study is to use QPPs to identify and rank effective queries, we include the correlation for comparison to previous studies and to show the drawbacks of using it as an evaluation strategy for the QVPP task.

3 METHOD

The goal of this study is to evaluate query performance predictors in the task of QVPP — and to determine whether it is possible to identify better performing queries over the original variation seed queries. As part of our study, we employ sixteen commonly used pre-retrieval and post-retrieval QPPs. Query variations are generated for each topic to provide the pool of possible queries. The QVPP task can be broken down into three subtasks: (S1) *ranking* the query variations according to estimated retrieval effectiveness, (S2) identifying a *better* query variation than the seed query, (S3) identifying the *best* query variation. These three subtasks will identify how correlated QPPs are with actual retrieval effectiveness, and, crucially, how useful they are in predicting the most effective queries.

3.1 Data and Materials

The CLEF 2017 TAR collection was used [10]. This collection consists of 50 topics, where each topic is based on a diagnostic test accuracy systematic review. Each topic contains the title of the systematic review (which we have used as the query for the QPP task, and the seed query for the QVPP task), the (Boolean) query used to retrieve citations in the systematic review, and a list of retrieved PubMed document identifiers (associated with relevant and non-relevant citations). We indexed a subset of 198,366 PubMed[1] citations with Lucene 6.2 using the publicly available Lucene4IR toolkit[2][2], where we employed Porter stemming and stopping. In line with CLEF TAR protocol, the subset of PubMed was obtained by retaining only citations that appear in the qrels of the TAR collection. For retrieval, we used language modelling with Dirichlet smoothing. In line with other studies on QPP, we set the smoothing parameter to $\mu = 1,000$ [6, 7, 18, 21]. Our retrieval experiments were performed using a field comprising a concatenation of the title, abstract, authors, and journal. For our experiments, we used MAP (as in previous work on QPP [6, 7, 18, 21]) as the evaluation metric. This aligns with the goals of the medical researchers who value precision and recall, and was used as part of the CLEF 2017 TAR track [10].

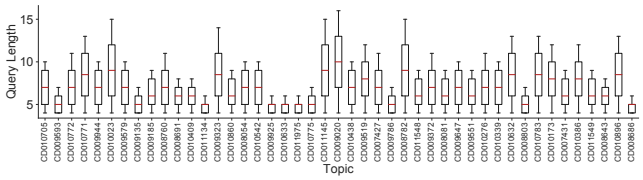
3.2 Query Performance Predictors

Our methodology uses the following pre-retrieval QPPs: **Query Length** (QL) [15], **Term Length** (TL) [9], **Character Length** (CL); which we define as: $\text{CL} = \sum_{w \in Q} |w|$, **Inverse Document Frequency** [5] (we use the sum (SIDF), average (AIDF), max (MIDF), and standard deviation (SDIDF) of each of the terms in the query for our idf-based predictors as has been common practice in previous work [5, 8, 25]), **Inverse Collection Term Frequency** (ICTF) [14], **Query Scope** (QS – ω) [17], **Simplified Clarity Score** (SCS) [9], and **Collection Query Similarity** [25]. Our methodology also uses the following post-retrieval QPPs: **Weighted Information Gain** (WIG) [26], **Weighted Expansion Gain** (WEG) [12], **Normalised Query Commitment** (NQC) [21], and **Clarity Score** (CS) [5]. The pre- and post-predictors have been chosen due to their common usage in the already established QPP task.

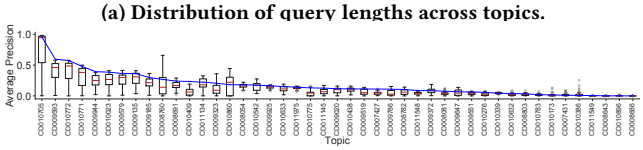
3.3 Query Variations

For each topic, the topic title was used as a seed query to generate query variations. We build upon the approach described by Azzopardi [1] whereby variations are generated by a process of query reduction. First, stop words were removed to create a query of n terms. Given this query of n terms, queries of lengths $n - i$ were extracted, where $i = 1$ to $i = n - 3$ (i.e. queries between 3 and n terms in length). We then used stratified sampling to randomly select a subset of the queries, based on the lengths. 50 queries per topic were then randomly sampled. We did not control for term redundancy between selected queries. Five of the topics were excluded because the title was less than 4 terms, and thus did not enable the generation of more than 50 queries of length 3 or more. The distribution of the length of variations across topics is reported in Figure 1a (ordered by seed query performance), while the distribution of the performance (AP) of the query variations over topics is reported in Figure 1b. In the latter, we also report the AP of the seed queries for comparison. This figure shows that for each topic, there is generally at least one query variation that outperforms the seed query. This motivates the task of QVPP.

[1]Downloaded on the 23rd of August 2017 comprising 26,759,399 citations.
[2]https://github.com/lucene4ir/lucene4ir



(a) Distribution of query lengths across topics.



(b) Distribution of AP for query variations (boxplots) and seed queries (in blue) across topics.

Figure 1: Distribution of query lengths and AP scores over each topic. Both sub figures are ordered by AP score.

3.4 Experimental Procedure

For each topic, query variations were issued to the retrieval system, their AP recorded, and the QPPs computed. To evaluate the QPP task and subtask S1 for the QVPP, the linear correlation (r) between predictors and AP was computed using Pearson's correlation, as in previous work [6, 9, 12, 18, 21]. Pearson's r measures the strength of the linear relationship between two variables. We then averaged over all topics and report both the mean and standard deviation. We further computed Kendall's and AP Correlation. Kendall's τ uses pairwise comparisons to determine how different one ranked list is from another. We use the τ_b variant which accounts for ties in the ranking. AP Correlation (τ_{ap}) [24] is conceptually similar to Kendall's τ, however it assigns more importance to rank disagreement at the top of the list. As we are interested in identifying orderings for high performing queries, τ_{ap} provides a more robust indication than τ. In this work, we use the $\tau_{ap,b}$ variant which accounts for ties in the ranking [23]. The inclusion of τ_b and $\tau_{ap,b}$ in our analysis is used to demonstrate the ability of a predictor to rank queries for the QVPP tasks — and in particular, to rank effective queries at the top ($\tau_{ap,b}$). For QVPPs subtask S2, we recorded the number of topics for which the predictor identified a variation with a higher AP than the seed query. For subtask (S3), we selected the query variation that is predicted to be the most effective for each topic. Given the "best" query variation for each topic, we then computed the AP and compared it against the seed query.

4 RESULTS AND DISCUSSION

4.1 Query Performance Prediction

We first describe the effectiveness of the predictors in the traditional QPP task; that is, the task of predicting the performance of a single query for a topic. The results of this task are reported in Table 1. Most predictors exhibited a weak positive or negative correlation, while TL exhibited the strongest correlation for all three correlation measures. These results indicate that in the context of predicting the effectiveness of queries for systematic reviews, general purpose QPPs such as the ones used in this work are not sufficient.

4.2 Query Variation Performance Prediction

Next, we investigate the QVPP task. For subtask S1, we examine the correlation values reported in Table 1.

QPP	QPP Task			QVPP Task		
	r	τ_b	$\tau_{ap,b}$	Mean r	Mean τ_b	Mean $\tau_{ap,b}$
AICTF	-0.092	-0.108	-0.106	0.208 (±0.320)	0.157 (±0.226)	0.072 (±0.206)
AIDF	-0.082	-0.075	-0.086	0.231 (±0.308)	0.168 (±0.214)	0.070 (±0.199)
AQL	0.162	0.088	0.122	0.323 (±0.249)	0.274 (±0.134)	0.151 (±0.144)
ASCQ	-0.126	-0.139	-0.143	0.216 (±0.287)	0.145 (±0.194)	0.045 (±0.193)
CL	0.161	0.088	0.120	0.323 (±0.249)	0.274 (±0.134)	0.152 (±0.144)
CS	-0.065	-0.054	-0.058	0.309 (±0.356)	0.234 (±0.294)	0.139 (±0.274)
MIDF	0.053	0.060	0.002	0.266 (±0.399)	0.236 (±0.321)	-0.098 (±0.293)
MSCQ	0.079	0.074	0.075	0.365 (±0.377)	**0.314** (±0.318)	-0.026 (±0.291)
NQC	-0.128	-0.149	-0.105	0.180 (±0.398)	0.147 (±0.312)	0.078 (±0.268)
QS	-0.080	-0.109	-0.049	-0.091 (±0.206)	-0.109 (±0.145)	-0.081 (±0.138)
SCS	-0.135	-0.162	-0.141	-0.029 (±0.311)	-0.030 (±0.226)	-0.039 (±0.192)
SDIDF	0.027	0.016	-0.021	**0.382** (±0.319)	0.309 (±0.170)	**0.175** (±0.170)
SSCQ	0.149	0.065	0.122	0.360 (±0.272)	0.295 (±0.149)	0.167 (±0.171)
TL	**0.178**	**0.118**	**0.126**	0.321 (±0.256)	0.284 (±0.145)	0.101 (±0.175)
WEG	-0.021	-0.016	0.012	0.375 (±0.338)	0.291 (±0.249)	0.173 (±0.241)
WIG	-0.107	-0.095	-0.055	0.364 (±0.375)	0.287 (±0.289)	0.173 (±0.267)

Table 1: Results from the QPP and QVPP tasks. The r and $\tau_{ap,b}$ is reported for the QPP task, and the mean r and $\tau_{ap,b}$ for each query variation in each topic is reported for the QVPP task, as well as the standard deviation.

- SDIDF exhibited the strongest r and $\tau_{ap,b}$. We further note that WIG, WEG, MSCQ and SSCQ also exhibited similar correlations in terms of r and τ_b compared to SFDIDF.
- MSCQ showed very weak $\tau_{ap,b}$ correlation, suggesting that the query variations predicted at the top of the ranking are not the most effective.

For subtasks S2 and S3, Table 2 reports the effectiveness of using predictors to select a query variation (the one predicted to be the most effective) for each topic in place of the seed query.

- SDIDF was not as effective compared to other predictors at identifying better query variations and was only able to correctly identify the best variation for one of the 45 topics.
- WEG, WIG, CS, and NQC (all post-retrieval predictors) correctly identified the best variations for more than two topics.
- On average, QPPs were not only unable to predict the best possible query variation (most performed worse than a random ranking, reported in the table for comparison), but were consistently unable to select variations that outperformed the seed query.

Finally, we examine the retrieval effectiveness (MAP) of the best query variations, as selected by each predictor (Table 2). The MAP of the seed queries was 0.180 and the MAP of the best possible query variations ("oracle") was 0.235 (statistical significant improvement over seed query: two-tailed paired t-test $p = 1.90e-05$). We found that none of the predictors selected variations which improved over the seed query, and all predictors except WEG were statistically significantly worse ($p < 0.05$). Nevertheless, we found that most predictors outperformed the random ranking and were also able to select query variations that were better than the median variations.

4.3 Comparison of QPP and QVPP

As reported in Table 1, predicting the performance across topics (QPP) appears more difficult than within topics (QVPP): in fact, stronger correlations are observed for the QVPP task than for the QPP task, though there is high variance across topics. Comparing the QPP task to the QVPP task, predictors in the query length family (i.e. TL, AQL, and CL) exhibit the strongest positive correlations in both the QPP and QVPP tasks. Additionally, these QPPs are able to identify a higher number of queries that are more effective than

both the seed query and median query baseline. TL exhibits the strongest correlation in the QPP task and is able to identify the highest number of query variations that are better than the seed. SCS and QS does not have strong correlations in the QPP task or the QVPP task; and the results of the QVPP task reflect the difficulty these predictors have in identifying better query variations. It is surprising to see predictors that exhibit the strongest relative correlations in the QVPP versus the QPP task perform worse when used to identify better variations. As a result, it is unclear what specific features contribute to identifying more effective query variations, and how these features can be applied to the Boolean queries used in medical systematic reviews.

5 CONCLUSION AND FUTURE WORK

While most predictors tend to be weakly correlated with MAP, the predictors themselves are poor at identifying the most effective query variations. The predictor that is most correlated with retrieval effectiveness is only able to predict the best query variation for one out of 45 topics. Four predictors (WEG, CS, WIG and NQC) are able to select more than two queries, with WEG able to select four. However the best predictors are only slightly better than a random ranking of query variations. Additionally, for the task of identifying better queries, there does not appear to be a strong relationship between r, τ_b, or $\tau_{ap,b}$ and the ability to select queries better than the baselines. We conclude by conjecturing that these popular predictors from the literature are insufficient in predicting query effectiveness of query variations, particularly within the context of ranking and selecting text queries for searching medical literature for systematic reviews.

Future work is required in this area to facilitate more effective query selection using predictors. First, we highlight the need for domain-specific query performance predictors that consider features relating to systematic reviews and the retrieval of relevant medical literature, e.g., the PICO framework, which is already used effectively to enhance retrieval in this domain [4, 19]. Secondly, this work can be extended to study the complex Boolean queries used in systematic reviews to determine how query performance predictors can be adopted. Furthermore, this investigation can also be expanded to query variations outside of the systematic review domain, e.g., [3, 27]. This initial work provides the foundations for the QVPP task and contextualises it for systematic reviews, where the outcomes have the potential for saving hundreds and even thousands of hours of search effort if more effective queries exist and can be recommended.

QPP	Better Seed Predictions	Better Median Predictions	Best Predictions	MAP
Random	12	13	2	$0.128^{1.51e-04}$
Median	12	—	0	$0.143^{6.48e-06}$
AICTF	14	12	2	$0.128^{8.80e-04}$
AIDF	11	13	0	$0.141^{3.81e-04}$
AQL	19	23	0	$0.156^{8.77e-03}$
ASCQ	7	11	1	$0.122^{3.03e-05}$
CL	19	23	0	$0.156^{8.77e-03}$
CS	17	22	3	$0.154^{3.05e-02}$
MIDF	9	11	0	$0.111^{2.62e-03}$
MSCQ	12	14	1	$0.147^{3.13e-02}$
NQC	12	18	3	$0.146^{4.63e-03}$
QS	9	13	0	$0.123^{1.49e-05}$
SCS	13	11	2	$0.131^{1.27e-03}$
SDIDF	18	29	1	$\mathbf{0.169}^{2.39e-02}$
SSCQ	19	**31**	1	$\mathbf{0.169}^{2.74e-02}$
TL	**21**	29	1	$0.167^{3.22e-02}$
WEG	15	22	**4**	$0.163^{7.12e-02}$
WIG	12	19	3	$0.157^{2.50e-02}$

Table 2: Results for QVPP tasks S2 ("better query") and S3 ("best query"). For S2, we also report for how many topics the variation was better than the median variation. For S3, we also report the MAP of the query variation ranked highest by each predictor. Statistical significance (two-tailed paired t-test) between the MAP values of seed queries and that of the highest ranked variation is reported in superscript.

REFERENCES

[1] L. Azzopardi. 2009. Query side evaluation: an empirical analysis of effectiveness and effort. In *SIGIR*.
[2] L. Azzopardi, Y. Moshfeghi, M. Halvey, R. S. Alkhawaldeh, K. Balog, E. Di Buccio, D. Ceccarelli, J. M. Fernández-Luna, C. Hull, S. Mannix, and S. Palchowdhury. 2016. Lucene4IR: Developing Information Retrieval Evaluation Resources Using Lucene. *SIGIR Forum* (2016).
[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A Test Collection with Query Variability. In *SIGIR*.
[4] F. Boudin, J. Nie, and M. Dawes. 2010. Clinical Information Retrieval using Document and PICO Structure. In *HLT*.
[5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2002. Predicting query performance. In *SIGIR*.
[6] C. Hauff, L. Azzopardi, and D. Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *ECIR*.
[7] C. Hauff, D. Hiemstra, and F. de Jong. 2008. A survey of pre-retrieval query performance predictors. In *CIKM*.
[8] B. He and I. Ounis. 2004. Inferring query performance using pre-retrieval predictors. In *SPIRE*.
[9] B. He and I. Ounis. 2006. Query performance prediction. *IS* (2006).
[10] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. 2017. Technologically Assisted Reviews in Empirical Medicine. In *CLEF*.
[11] S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC* (2010).
[12] A. Khwileh, A. Way, and G. J. F. Jones. 2017. Improving the Reliability of Query Expansion for User-Generated Speech Retrieval Using Query Performance Prediction. In *CLEF*. CLEF.
[13] B. Koopman, L. Cripwell, and G. Zuccon. 2017. Generating clinical queries from patient narratives: A comparison between machines and humans. In *SIGIR*.
[14] K. L. Kwok. 1996. A New Method of Weighting Query Terms for Ad-hoc Retrieval. In *SIGIR*.
[15] J. Mothe and L. Tanguy. 2005. Linguistic features to predict query difficulty. In *SIGIR*.
[16] U. Ozertem, O. Chapelle, P. Donmez, and E. Velipasaoglu. 2012. Learning to suggest: a machine learning framework for ranking query suggestions. In *SIGIR*.
[17] V. Plachouras, I. Ounis, C. J. van Rijsbergen, and F. Cacheda. 2003. University of Glasgow at the Web Track: Dynamic Application of Hyperlink Analysis using the Query Scope.. In *TREC*.
[18] F. Raiber and O. Kurland. 2014. Query-performance prediction: setting the expectations straight. In *SIGIR*.
[19] H. Scells, G. Zuccon, B. Koopman, A. Deacon, L. Azzopardi, and S. Geva. 2017. Integrating the framing of clinical questions via PICO into the retrieval of medical literature for systematic reviews. In *CIKM*.
[20] I. Shemilt, N. Khan, S. Park, and J. Thomas. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Syst. Rev.* (2016).
[21] A. Shtok, O. Kurland, and D. Carmel. 2009. Predicting query performance by query-drift estimation. *ECIR* (2009).
[22] E. Tacconelli. 2010. Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis* (2010).
[23] J. Urbano and M. Marrero. 2017. The Treatment of Ties in AP Correlation. In *ICTIR*. 4.
[24] E. Yilmaz, J. A. Aslam, and S. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *SIGIR*. 8.
[25] Y. Zhao, F. Scholer, and Y. Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. *ECIR* (2008).
[26] Y. Zhou and W. B. Croft. 2007. Query performance prediction in web search environments. In *SIGIR*.
[27] G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, and A. Deacon. 2016. The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In *CLEF*.