Can Users Predict Relative Query Effectiveness?

Oleg Zendel
RMIT University
Melbourne, Australia
oleg.zendel@student.rmit.edu.au

Melika P. Ebrahim
RMIT University
Melbourne, Australia
melika.ebrahim@rmit.edu.au

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

ABSTRACT

Any given information need can be expressed via a wide range of possible queries. Recent work with such *query variations* has demonstrated that different queries can fetch notably divergent sets of documents, even when the queries have identical intents and superficial similarity. That is, different users might receive SERPs of quite different effectiveness for the same information need. That observation then raises an interesting question: do users have a sense of how useful any given query will be? Can they anticipate the effectiveness of alternative queries for the same retrieval need? To explore that question we designed and carried out a crowd-sourced user study in which we asked subjects to consider an information need statement expressed as a backstory, and then provide their opinions as to the relative usefulness of a set of queries ostensibly addressing that objective. We solicited opinions using two different interfaces: one that collected absolute ratings of queries, and one that required that the subjects place a set of queries into "order". We found that crowd workers are reasonably consistent in their estimates of how effective queries are likely to be, and also that their estimates correlate positively with actual system performance.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Query variations; query performance prediction

ACM Reference Format:
Oleg Zendel, Melika P. Ebrahim, J. Shane Culpepper, Alistair Moffat, and Falk Scholer. 2022. Can Users Predict Relative Query Effectiveness?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531893

1 INTRODUCTION

It has been known for some time that one of the sources of variability in IR evaluations arises with the users themselves, and that even when two users face exactly the same information need, they are likely to issue different queries when searching for information. For example, more than twenty years ago Buckley and Walz [4] considered the role of queries, and commented "*queries dealing with the same topic are extremely variable . . . ; and even short queries were rarely duplicated*"; and a further two decades prior to that Spärck Jones and Bates [20] observed "*variations over requests should be counteracted by the use of additional queries specifically designed to exhaust the relevant document set*".

More recently investigations have considered the role of query variations in terms of evaluation methodologies and consistency [1, 29]; collection design and construction [13, 14]; effectiveness metrics [15]; and as a device for increasing search performance [2, 3]. It is clear that query variations are an important – but perhaps under-appreciated – facet of system evaluation.

At the same time, there has been interest in *query performance prediction* (QPP). For example, a retrieval system might have a policy of early truncation for "easy" queries that it (somehow) knows will achieve good early effectiveness and hence user satisfaction, so as to be able to reallocate the saved resources to queries that it somehow knows will be "hard", seeking to avoid alienating one segment of its user base. Important work in this area includes that of He and Ounis [11], Carmel and Yom-Tov [5], and Hauff et al. [9]. As another example, Craswell et al. [6] give a per-query analysis comparing traditional IR systems with a learned model.

At the intersection of query performance prediction and user query variations we thus have an interesting question: do users have a sense of how useful any given query will be? That is, can users anticipate the effectiveness of alternative queries for the same retrieval need, and hence provide guidance (or even training data) to automatic techniques for query performance prediction?

To examine those questions we designed and carried out a crowd-sourced user study in which we asked subjects to consider an information need statement expressed as a backstory, and then provide their opinions as to the relative usefulness of a set of queries ostensibly addressing that objective. We solicited opinions using two different interfaces: one that collected absolute ratings of queries via a "stars" mechanism, with the queries presented together but judged individually; and one that required that the subjects assess the provided queries as a set, placing them in order from "*best*" to "*worst*", where "*the best query is the one that you think is most*

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07. . . $15.00
https://doi.org/10.1145/3477495.3531893

likely to generate useful results". Our experiments revealed three interesting relationships:

- Crowd workers are reasonably consistent in their evaluation of query quality;
- The "rating" and "ranking" interfaces yielded consistent outcomes; and
- Crowd worker evaluations of query usefulness correlated with actual effectiveness from a typical retrieval system when SERPs (search engine results pages) were evaluated using NDCG@10.

Section 2 describes the structure of the experiments we carried out and the findings we obtained; Section 3 summarizes a range of related work; and then Section 4 presents our conclusions and identifies areas for possible future work.

2 EXPERIMENTAL DESIGN AND EXECUTION

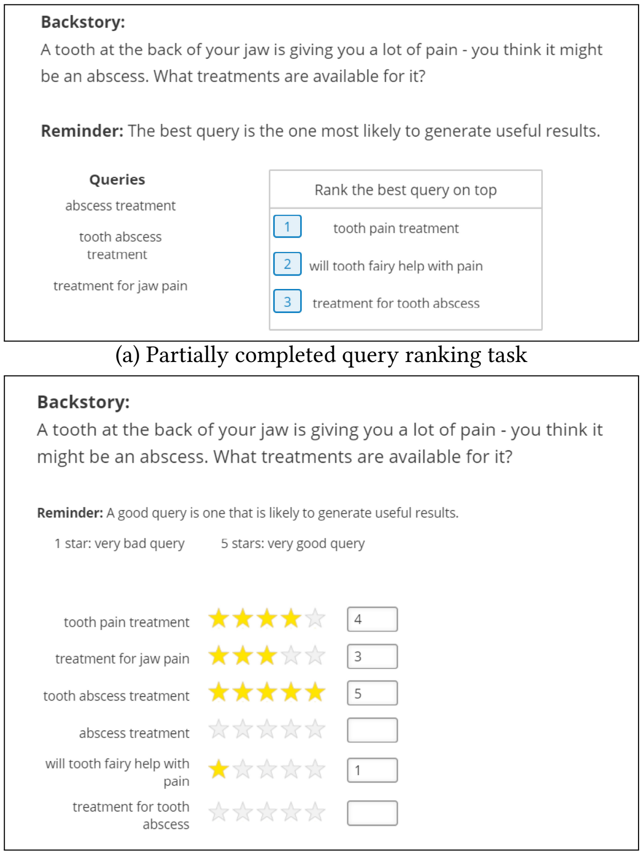
We now describe the development and structure of the data collection activity, and present the results that were obtained from it.

Topics, backstories, collection, and system. We selected a subset of twelve of the UQV100 topics and backstories [1], and five UQV100 query variations for each, choosing both the topics and then the queries to be broadly representative of the entire UQV100 query pool. The UQV100 dataset also contains relevance judgments against the ClueWeb12 Category B collection,[1] formed via pooling across the query variations. Effectiveness was measured using NDCG@10, in part as a consequence of the findings of Sakai and Zeng [18], who compared a range of metrics via the lens of whole-SERP satisfaction ratings assigned by users.

The system used was a Terrier PL2 divergence from randomness implementation,[2] employing sequential term dependencies with a window size of five (the default), with stop words removed and Krovetz stemming applied, and with query expansion disabled [16, 17]. We regard this as being a typical IR system, and hence expect that our results would then be broadly applicable to other systems as well. In future work we will also explore a range of further retrieval systems, and using more topics and queries.

Crowd workers and payment. We used the Amazon Mechanical Turk system, with workers required to be English-speakers and permitted to undertake the Human Intelligence Tasks (HITs) only if they had more than 3,000 previous HITs approved, and attained a 90% or better task acceptance rate. In accordance with our local research ethics requirements (and out of our respect for human dignity) workers were paid the equivalent of the minimum adult wage applicable in Australia, based initially on our self-measurement as we previewed the HITs, and then refined via preliminary experiments in which the time taken by workers to complete the HIT was recorded. In the main experiment (described below) workers were paid US$1.40 for each HIT they submitted.

HIT structure. Two distinct types of HIT were used: "*ranking*" and "*rating*". Each HIT in both categories involved a fixed set of four of the twelve topics, and hence there were three different ranking HITs, three different rating HITs, and six different HITs in total.



Backstory:
A tooth at the back of your jaw is giving you a lot of pain - you think it might be an abscess. What treatments are available for it?

Reminder: The best query is the one most likely to generate useful results.

Queries
abscess treatment
tooth abscess treatment
treatment for jaw pain

Rank the best query on top
1 tooth pain treatment
2 will tooth fairy help with pain
3 treatment for tooth abscess

(a) Partially completed query ranking task

Backstory:
A tooth at the back of your jaw is giving you a lot of pain - you think it might be an abscess. What treatments are available for it?

Reminder: A good query is one that is likely to generate useful results.
1 star: very bad query 5 stars: very good query

tooth pain treatment	★★★★☆	4
treatment for jaw pain	★★★☆☆	3
tooth abscess treatment	★★★★★	5
abscess treatment	☆☆☆☆☆	
will tooth fairy help with pain	★☆☆☆☆	1
treatment for tooth abscess	☆☆☆☆☆	

(b) Partially completed query rating task

Figure 1: Screen shots showing the two data collection modes, captured part way through each task. The work flow was structured so that each worker only encountered "query ranking" HITS (top) or only encountered "query rating" HITS (bottom). Five of the queries in each set of six are from the UQV100 collection; one of the six is an additional quality control query devised by us. The quality control query in the example is "*will tooth fairy help with pain*", and workers were expected to give it the lowest rank or rating.

In both types of HIT (ranking and rating) workers were presented with an information need statement expressed as a backstory [1, 15] and asked to respond to three initial statements via a five-point Likert item from "*strongly disagree*" to "*strongly agree*":

- I am familiar with this topic;
- this topic is interesting to me; and
- it should be easy to find relevant information for this topic.

The users were next requested to write a query that they would use to search for information on that topic.

The rankings HITs then presented a set of six different queries for each of the topics, see Figure 1(a) for one such topic and the corresponding backstory. User were required to drag the queries from the list at left and place them into the box on the right, with rank labels appearing as they did so. Queries could also be reordered within the box on the right after they had been dragged and dropped the first time.

[1] See http:www.lemurproject.org/clueweb12.php/.
[2] See http://www.terrier.org.

The ratings HITs used the same query set, but asked the workers to instead assign a "star" rating to each query, either by clicking on the list of stars, or by typing a count into the box to the right of the stars (Figure 1(b)). Each rankings HIT and corresponding ratings HIT contained the same four topics and same 24 queries, but the topics were ordered randomly for each user, and within each topic the queries were also ordered randomly.

A total of 100 workers were permitted to complete each HIT, with users unable to complete the rating HITs if they commenced any of the ranking HITs, and vice versa. That is, workers were either able to undertake up to three HITs of the rankings type, or up to three HITs of the ratings type. All data collection took place in the period 10–14 February 2022. In total, and including the preliminary experimentation, we spent US$1,177 collecting our data.

Post-work filtering. As can be seen in Figure 1, one of the six queries for each topic was inserted as a deliberately off-topic *quality control* filter. Workers were expected to give that query the lowest ranking or rating, and their work was flagged if they failed to do so more than twice. Workers were also asked to provide their own query for the topic, and these were scrutinized for appropriateness. If they did not correspond to the topic, those HITs were rejected. The HITs that had been flagged were also checked manually, and as a result, another 46 HITs were "approved" in terms of payment, but not included in the analysis. The overall HIT approval rate was 93.8%. After these filters had been applied the three rankings HITs had 90, 90, and 93 valid responses; and the three ratings HITs had 73, 84, and 88 responses that were used. Those 518 HITs had been undertaken by 267 workers, doing an average of 1.94 HITs each.

One unexpected outcome from the data collection process was that the Qualtrix survey instrument that was used as the within-HIT tool collected "star" ratings that included 0 as a valid value – the interface allowed "0" to be typed into the text box, even though that value could not be selected using the "clickable array". The small number of 0 values that were entered were all transformed to "one star" ratings for consistency in the subsequent analysis.[3]

Data interpretation. In the case of the rankings data, each user assigned a value between "1" (top-ranked) and "6" to each query. These were transformed by: (a) removing the quality control query, and condensing the five remaining rankings to the range 1 . . . 5; and then (b) subtracting from six, to get rankings from 5 (top-ranked) down to 1. The ratings were already from 5 (top-rated) down to 1.

Both rankings and rating are ordinal scale data. To allow numeric analysis, we mapped the ordinal categories to their numeric counterparts. For example, a "two star" rating was treated as the number 2.0 for the purposes of computing averages and standard deviations. We acknowledge that other numeric assignments would lead to different average scores, but do not believe that the overall conclusions would be substantially altered.

Analysis: Worker consistency. Figure 2 shows the pattern of ratings and rankings returned for two of the topics, numbers 213 and 286. The UQV100 backstory for Topic 213[4] is:

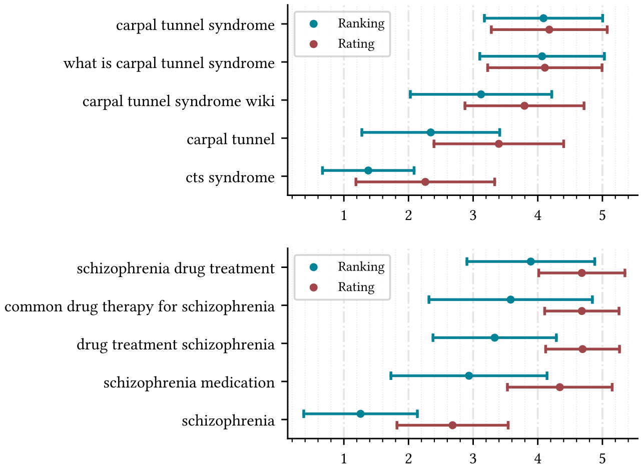


Figure 2: Means and standard deviations of mapped worker responses, with queries presented in order of decreasing mean mapped ranking, for Topic 213 (top) and Topic 286 (bottom). The first and third queries of Topic 286 are very similar, and would give exactly the same output in a bag-of-words retrieval system.

> "*Your elderly aunt has recently complained about pain in her hands. Having recently heard of carpal tunnel syndrome, you wonder if this might be what she's suffering from, and decide to find more information on the symptoms of this syndrome.*"

The two sections of Figure 2 show the two sets of five queries in decreasing order of mean mapped ranking score. As can be seen, there is reasonably good consistency in the workers' opinions of relative usefulness, and the standard deviation associated with each of the queries is relatively small. Confidence intervals about the mean values were also calculated, and for Topic 213 the largest of the ten confidence intervals was 0.49 wide, for the query "*cts syndrome*" and the rating interface. A broadly similar pattern of behavior also occurred across the other ten topics.

Analysis: Rankings versus ratings. Also apparent in Figure 2 is the agreement between the query orderings generated by the two interfaces. To quantify the overall agreement, a Pearson coefficient was calculated over the full set of sixty queries comparing the mean rankings and ratings scores. The resultant correlation of $r = 0.90$ confirms that the two interfaces have a high degree of agreement.

In the lower part of Figure 2 three of the five queries were judged plausible by the workers, illustrating a drawback of the ranking interface: because an ordering is always required, workers were unable to assign ties. That forced distinction led to the high standard deviations with the rankings approach, whereas the ratings interface allowed three high star ratings to be assigned, which, across the pool of workers, then resulted in a smaller standard deviation.

Analysis: Worker predictions. We can now address the question posed in our title: can users predict relative query effectiveness? All 12 × 5 = 60 queries were executed using the experimental retrieval system, and SERPs generated and scored. The resulting NDCG@10 scores were scatter-plotted against the mean mapped rankings and

[3] There is an interesting methodological question here. Should a "one star" to "five star" scale permit a "zero star" response if the worker does not want to assign any stars?
[4] See http://dx.doi.org/10.4225/49/5726E597B8376 for other backstories.

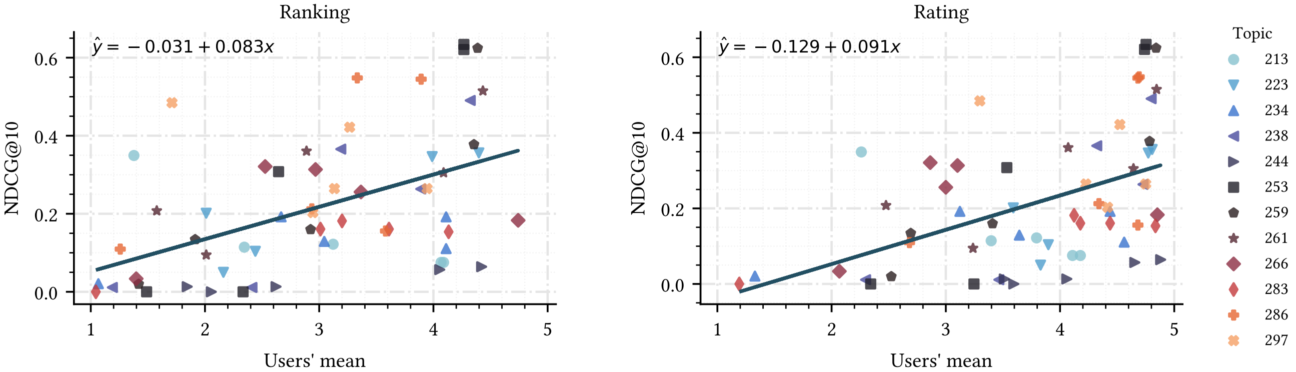


Figure 3: Scatter-plot of NDCG@10 scores (vertical axes) against crowd worker opinion in regard to query usefulness (horizontal axes), using rankings (left pane) and ratings (right pane). The line of best fit is overlaid for each set of 60 data points. The sixty points have Pearson correlation coefficients of $r = 0.50$ and $r = 0.49$ for rankings and ratings respectively; both with significance $p < 0.001$.

ratings generated by the workers for that query (shown in Figure 3) and Pearson coefficients computed.

What is immediately apparent is that there is an observable correlation between the workers' perceptions as to query usefulness and the NDCG@10 scores for the corresponding SERPs. Note that at no stage did the crowd workers see the SERPs; nor is it likely that any of the workers executed the queries at a search service and carried out their own effectiveness assessment in order to try and "get it right". We also reiterate that these plots do not include the quality control queries, and that all of the queries that have contributed to the correlation are genuine variations arising from the backstory, as collected by Bailey et al. [1].

We can thus answer in the affirmative: users *can* predict relative query effectiveness, at least to some extent. This pattern is, however, variable across topics. For example, the five points associated with Topic 213 (light blue circles) form a negatively correlated pattern in both panes of Figure 3, as do the five points associated with Topic 297 (orange crosses). Hence another area for future exploration is to explore this data on a per-topic basis, with more query variations and more crowdworkers assigned, to understand whether these observed counter-trends are simply chance over small numbers of samples, or represent more fundamental issues associated with those two topics.

For perspective, an automatic query performance prediction based on NQC – a post-retrieval mechanism which measures the standard deviation of the retrieval scores of the top k retrieved documents for each query [19] – achieved a Pearson correlation against NDCG@10 of $r = 0.46$, with k chosen from $\{10, 25, 50, 75, 100\}$ to maximize the correlation.

3 RELATED WORK AND BACKGROUND

We now provide an overview of some of the related literature.

Query variations. As has already been noted in Section 1, query variations have received renewed interest through the last five years, with crowd workers used to create several resources [1, 12] and a number of implications then studied [2, 3, 13–15, 29]. As one part of that sequence of work, Thomas et al. [21] explore possible query syntactic features that might correlate to query effectiveness, with a number of moderate relationships emerging. Our work here can be viewed as a user-based complement to that study.

Query performance prediction. Section 1 also introduced some of the previous work in the area of query performance prediction (QPP). In particular, Carmel and Yom-Tov [5] summarize a wide range of prior work, including the early Clarity predictor of Cronen-Townsend et al. [7] and the work by Zhao et al. [26] and Zhou and Croft [27, 28]; and Hauff et al. [9] compare a total of 22 pre-retrieval predictors across multiple TREC document collections and topic sets. Since then Shtok et al. [19] have introduced their "NQC" method, based on the standard deviation of retrieval scores; it is still regarded as providing competitive results.

More recently, Zendel et al. [24] propose a framework to utilize query variants from the same topic as reference points to improve QPP quality, noting that the query chosen to represent each topic has a measurable effect on relative QPP quality; and Faggioli et al. [8] suggest the use of ANOVA analysis to predict rank difference errors. In related work Zendel et al. [25] compare automatic QPP methods between queries from different topics and the same topic (variants), concluding that most of the previously reported differences are due to differences in effectiveness (measured by AP). That is, queries from different topics are more likely to have significant retrieval effectiveness between them than queries from the same topic, an effect also noted by Thomas et al. [21].

User studies. Other investigations have examined users and their response to information need statements. For example, Wu et al. [23] undertook a qualitative user study with eight topics (five exploratory; three navigational), with 41 users asked to rate queries (one per information need). Pre- and post-retrieval evaluations were carried out, with users asked to rate queries before and after seeing each SERP. One interesting finding was that users tend to see longer queries as better; it is an area where we will be able to extend the analysis already reported here.

Hauff et al. [10] compare automatic QPP methods to users' ratings, using correlation to measure the relationship between user and system effectiveness, and Cohen's Kappa to measure inter-assessor agreement. Their experiments show that correlations are low when rating a single query per topic, and even worse when rating query suggestions (variants for the same topic), concluding that pre-retrieval QPP works poorly for rating suggestions, and that it might be easier to rank suggestions than to rate them.

Finally, we note that Turpin and Scholer [22] carried out experiments using manipulated rankings, finding no correlation between system effectiveness (measured by average precision) and the time taken by users to find a set of relevant documents. They also found that topic effect was a strong influence, with some topics easy for their subjects and others hard, but not in a manner that was connected to SERP effectiveness. The users themselves were also a measurable effect, with some significantly faster than others.

4 CONCLUSIONS AND FUTURE WORK

We have explored the question of whether users have a sense of how effective queries are likely to be. To do that we used two different modalities to gather opinions about the usefulness of queries, presenting crowd workers with multiple query variations for each of a sequence of search topics. The workers' ordinal responses were then mapped to numeric values for the purposes of analysis.

We found that users have reasonably good agreement with each other in terms of which queries are "likely to return useful results" for a given search topic, and which queries were not. More importantly, their aggregate opinions are also in broad agreement with the results generated by an actual search system when effectiveness is measured via NDCG@10. This is a rather pleasing result, since it means that query performance prediction techniques can also hope to obtain the same strong outcomes.

In future work we plan to investigate other facets of our data, including: analysis of the queries we collected from the workers at the time they provided their rankings or ratings; other effectiveness metrics, to determine the extent to which different metrics correlate with users' advance perceptions; and other systems, to determine the extent to which users' advance perceptions might be shaped by the characteristics of different similarity models.

Acknowledgements. We thank the reviewers for their comments. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (grant DP190101113). Code and data are available from https://github.com/Zendelo/cs-qpp.

REFERENCES

[1] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*, pages 395–404, 2017.

[3] R. Benham, J. Mackenzie, A. Moffat, and J. S. Culpepper. Boosting search performance using query variations. *ACM Trans. Inf. Sys.*, 37(4):41.1–41.25, 2019.

[4] C. Buckley and J. Walz. The TREC-8 query track. In *Proc. TREC*, 1999.

[5] D. Carmel and E. Yom-Tov. *Estimating the query difficulty for information retrieval.* Number 15 in Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool, 2010.

[6] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees. Overview of the TREC 2019 deep learning track. In *Proc. TREC*, 2020.

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. SIGIR*, page 299–306, 2002.

[8] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, and F. Scholer. An enhanced evaluation framework for query performance prediction. In *Proc. ECIR*, pages 115–129, 2021.

[9] C. Hauff, D. Hiemstra, and F. De Jong. A survey of pre-retrieval query performance predictors. In *Proc. CIKM*, pages 1419–1420, 2008.

[10] C. Hauff, D. Kelly, and L. Azzopardi. A comparison of user and system query performance predictions. In *Proc. CIKM*, page 979–988, 2010.

[11] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proc. SPIRE*, pages 43–54, 2004.

[12] J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, and A. Moffat. CC-News-En: A large English news corpus. In *Proc. CIKM*, pages 3077–3084, 2020.

[13] A. Moffat. Judgment pool effects caused by query variations. In *Proc. Aust. Doc. Comp. Symp.*, pages 65–68, 2016.

[14] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *Proc. CIKM*, pages 1759–1762, 2015.

[15] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.

[16] I. Ounis, G. Amati, P. V., B. He, C. Macdonald, and Johnson. Terrier information retrieval platform. In *Proc. ECIR*, pages 517–519, 2005.

[17] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. Incorporating term dependency in the DFR framework. In *Proc. SIGIR*, pages 843–844, 2007.

[18] T. Sakai and Z. Zeng. Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures. *ACM Trans. Inf. Sys.*, 39(2):14:1–14:35, 2021.

[19] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Sys.*, 30(2):1–35, 2012.

[20] K. Spärck Jones and R. G. Bates. Report on the design study for the "ideal" information retrieval test collection. Technical Report 5428, Computer Laboratory, University of Cambridge, 1977. British Library Research and Development Report.

[21] P. Thomas, F. Scholer, P. Bailey, and A. Moffat. Task, queries, and rankers in pre-retrieval performance prediction. In *Proc. Aust. Doc. Comp. Symp.*, pages 11.1–11.4, 2017.

[22] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. SIGIR*, page 11–18, 2006.

[23] W.-C. Wu, D. Kelly, and K. Huang. User evaluation of query quality. In *Proc. SIGIR*, page 215–224, 2012.

[24] O. Zendel, A. Shtok, F. Raiber, O. Kurland, and J. S. Culpepper. Information needs, queries, and query performance prediction. In *Proc. SIGIR*, page 395–404, 2019.

[25] O. Zendel, J. S. Culpepper, and F. Scholer. Is query performance prediction with multiple query variations harder than topic performance prediction? In *Proc. SIGIR*, pages 1713–1717, 2021.

[26] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. ECIR*, page 52–64, 2008.

[27] Y. Zhou and W. B. Croft. Ranking robustness: A novel framework to predict query performance. In *Proc. CIKM*, pages 567–574, 2006.

[28] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proc. SIGIR*, pages 543–550, 2007.

[29] G. Zuccon, J. Palotti, and A. Hanbury. Query variations and their effect on comparing information retrieval systems. In *Proc. CIKM*, pages 691–700, 2016.