Modality Effects When Simulating User Querying Tasks

Joel Mackenzie
The University of Melbourne
Melbourne, Australia
joel.mackenzie@unimelb.edu.au

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

ABSTRACT

When faced with an information need, different users are likely to issue different queries. A renewed interest in these *user query variations* has resulted in a number of collections, tasks, and studies which utilize multiple queries for each topic. The most commonly applied technique for generating query variations is to show a short *backstory* to a pool of crowdworkers, and ask each of them to provide a keyword query that they would expect to provide more information pertaining to the backstory. In this short paper we explore whether the length of the backstory and the mode in which the backstory is conveyed to crowdworkers affect the resulting queries. Our experiments demonstrate that both the length of the backstory and the mode in which the backstory is delivered influence the resultant query variations; and that there is a consequent downstream implication in terms of forming the judgment pools necessary to assess systems in the presence of query variations.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Crowdsourcing**; **Query log analysis**; *Search interfaces*; *Personalization*.

KEYWORDS

User query variations; crowdsourcing

ACM Reference Format:
Joel Mackenzie and Alistair Moffat. 2021. Modality Effects When Simulating User Querying Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3471158.3472244

1 INTRODUCTION

The recent CC-News-En collection [20] joins other activities that have taken place in the last few years that create *user query variations*, that is, sets of queries that have the same underlying information need, but have been expressed using different words and have originated from different users. For example, Bailey et al. [3] describe a collection of approximately 5,000 queries in connection with 100 topics, and then obtain relevance judgments from experienced judges, allowing (amongst many other questions that can be considered) comparative effectiveness of queries to be assessed [4].

Common to these activities is the use of crowdworkers to create the query variations. Workers are typically presented with a brief *backstory*, motivating the topic, and asked to "imagine" that they wanted to find out more in connection with what they are told via the backstory. For example, one of the UQV100 backstories was [3]:

> *Having heard of the pristine environment in Maryland, you have long dreamed of taking a fishing holiday there. However, you think that you may need a fishing license in Maryland. How do you get one?*

Once the backstory has been presented, the workers are then asked what their first query would be, and requested to enter that into a text box that is captured by the crowdsourcing interface.

The recent publicly-available CC-News-En collection [20] was built following that general pattern. But it made use of backstories of three different lengths, and delivered them to the workers via two quite different modalities – both written and oral. The data that was collected from crowdworkers thus allows us to examine the following research question: *How do backstory properties affect crowdworker query generation?* In particular, we are able to investigate whether the *length of the backstory* and the *presentation mode of the backstory* impact the diversity of the queries generated by crowdworkers, and whether they then affect the range of documents that those queries then retrieve via a search system. Our results indicate that showing textual backstories to crowdworkers may increase the homogeneity of the queries, with workers more likely to adopt query terms directly from the backstory.

2 BACKGROUND

The importance of user query variability has been known for decades, with early studies examining differences in query formulations [27], including the rankings that arise from variations [5], and including ways of combining those rankings to form more effective and robust rankings [6]. The TREC-8 Query Track [12] examined many issues of query variability, including how and why systems perform differently under variations of the same query; whether topic difficulty can be predicted over query variations; and how test collections can be created with query variations.

Renewed recent interest in user query variations has led to exploration in a range of areas, including crowdsourcing query variations [3, 20, 24, 30]; enhancing search effectiveness through the use of query variations [4, 7, 8, 19, 29]; efficient processing of query variations [10, 13]; query performance prediction across variations [14, 28, 31, 35]; evaluating search across multiple query variations for a single topic [2, 22, 23, 36]; automatic generation of query variations [9, 18]; and investigation of the differences between humans

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472244

and systems for generating clinical queries from verbose patient narratives [16].

Crowdsourcing User Query Variations. In important early work, Belkin et al. [5] employed experienced searchers to generate user query variations based on existing TREC topic descriptions, each of which included information about the problem, a description of desired document contents, and a narrative describing how relevance would be assessed. A similar approach was used in the TREC-8 Query Track [12], where various experts (students or academics) generated query variations based on topic descriptions. Recently, Benham et al. [8] used this approach to generate query variations for the TREC CORE track.

Instead of using the somewhat verbose TREC topic statements, Bailey et al. [2, 3] crafted a backstory for each topic of interest, aiming to make the information statements "less abstract and more engaging" [2]. To this end, the backstories were intentionally written so that crowdworkers might empathize with the information need. To assist with that goal, the backstories often included fictional family members or friends, to make them more natural to crowdworkers than the plain topic descriptions would be.

Visual Prompting. Use has also been made of rich media such as photographs or videos to motivate crowdworkers to generate queries. Stanton et al. [30] explored this idea in the domain of medical queries, with crowdworkers shown an image (for example, a red rash on an arm) or a short video clip (perhaps a person with a severe cough). The workers were then told to imagine that they had the affliction that was depicted, and asked to provide three distinct search queries with which they might hope to find useful information. The same approach was used to generate medical query variations for the CLEF eHealth Evaluation task [15, 25].

Oral Prompting. Most recently, Mackenzie et al. [20] created backstories by generating summaries of *target articles* which could be conveyed to crowdworkers in order to elicit query variations, with the resultant queries a form of *known-item finding* task [1]. As well as displaying the plain textual backstory (as an image of the text, a standard deterrent against `Copy+Paste` functionality), use was made of *spoken audio clips*, simulating a "breaking news" summary which might be heard over the radio or on TV. Both of these extensions – summaries of different lengths, and aural cueing – were included with the aim of generating greater query diversity.

Our Contribution. We make use of that CC-News-En resource, first to explore whether the length or modality of the backstory influences the variance of the resulting query variations, and then to establish the implications of the variance in querying on downstream evaluation and pooling tasks.

3 INTRINSIC ANALYSIS

We first conduct an intrinsic analysis over a set of user query variations and the corresponding backstories, to determine how different query generation approaches impact the resulting queries generated by crowdworkers.

Crowdsourced Queries. We employ the UQVs from the recent CC-News-En resource [20], consisting of over 10,000 query variations for 173 unique news topics, compiled in early 2020 (pre-pandemic). To collect these query variations, crowdworkers were provided with a backstory outlining a news topic, with backstories of three different lengths (Title < Short < Long) constructed using automatic summarization tools. Either an image of text (Text) or an automatically generated spoken audio clip (Audio) of the backstory was provided to each user. Hence, for each topic, there were six distinct ways of conveying the backstory: {Title, Short, Long} × {Text, Audio}. For example, the set of three backstories associated with Topic 50 (related to global warming in the Arctic) were:

- Title: *Arctic heating up at twice as fast as rest of globe.*
- Short: *The Arctic is heating up twice as fast as the rest of the world – triggering a "massive decline in sea ice and snow," according to a new federal report.*
- Long: *The Arctic is heating up twice as fast as the rest of the world – triggering a "massive decline in sea ice and snow," according to a new federal report. On Tuesday, the National Oceanic and Atmospheric Administration released its 11th annual Arctic Report Card, which compiles data from 61 scientists in 11 countries. "Rarely have we seen the Arctic show a clearer, stronger or more pronounced signal of persistent warming and its cascading effects on the environment than this year," Jeremy Mathis, director of NOAA's Arctic Research Program, said in a statement.*

The processing pipeline that led to these three lengths meant that the Short backstory was always a prefix of the Long backstory [20]. Each collected UQV has both raw and spell-corrected instances available, alongside associated keystroke-level interaction data including timestamps. Full details of the acquisition process and interface are provided by Mackenzie et al. [20]; and Benham et al. [11] have employed the CC-News-En keystroke timing data in a visualization that illustrates how crowdworkers author their queries.

Query Length. First, we examine the lengths of the query variations compared to the length of the corresponding backstory. Intuitively, longer backstories might be expected to result in lengthier queries, as crowdworkers have more seed information to work with. Figure 1 shows the observed query length for both Text and Audio modes stratified by backstory length. Query length is relatively stable and largely independent of the backstory length, although very short and very long backstories seemed to result in slightly shorter queries. Interestingly, the Audio mode resulted in shorter queries than the Text mode, for each backstory length. We hypothesize that this is a consequence of the ephemeral nature of aural backstories – they are no longer available to the worker once they start typing, making it cognitively harder to generate each query. Indeed, recent work from Leroy and Kauchak [17] shows that information presented over a textual channel is more likely to be retained than information presented over an audio channel.

Query Term Variance. Next, we investigate how backstory length and mode impact the *variance* of the query terms generated. We follow the approach of Liu et al. [18] and use *Jaccard similarity* as a measure of variance, treating each backstory, and each worker-generated query, as a bag-of-words.

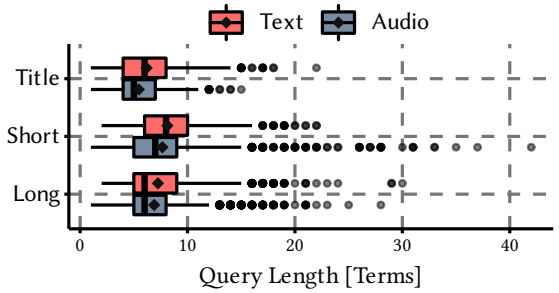

Figure 1: Query length, stratified by backstory length and presentation mode. Queries collected via audio recordings tend to be slightly shorter than those collected via textual images.

Figure 2, top, plots the Jaccard similarity between each query and its associated backstory. Similarity decreases with the increased length of the backstories, because the denominator of the similarity computation is larger for longer backstories. Within each backstory length the Audio mode tends to result in queries with a lower similarity to the seeding backstory than does the corresponding Text mode. We hypothesize that this is again due to the ongoing availability (while the crowdworker is typing) of the Text backstories, making it less cognitively demanding to develop a query from the textual backstory. A high-similarity tail can be observed for the Short Audio data; detailed analysis of those queries suggested that this was caused by crowdworkers either copying subsets of the audio word-for-word, or using automatic text-to-speech tools while undertaking the task [11, 20].

Figure 2, bottom, shows the Jaccard similarity between *all pairs* of query variations for each topic, again stratified by backstory length and mode. This computation measures the *intra-topic similarity* [18], and provides a measure of similarity of the set of query variations for each topic. The Title backstories provide the highest intra-topic similarity, presumably because there is less information available to the crowdworkers as they are generating their queries. Interestingly, the Short and Long backstories are not markedly different in terms of their similarity, which may suggest that there is a saturation point with respect to the amount of information retained from a backstory. An alternative (and perhaps more cynical, but perhaps also more pragmatic) explanation is that the crowdworkers typed their query just as soon as they thought they had a sufficient understanding of the topic, and ignored the further details available in the Long backstories. Once again, the Audio mode resulted in less intra-topic similarity than the associated Text mode.

Typographical Errors. Another factor that arises in real-world querying is that of typographical errors. For example, Benham et al. [11] note that audio tasks can result in users misunderstanding certain terms (like mistaking the term *"Obama"* for the phonetically similar phrase *"A bomber"*). Similarly, we expect that users who are provided Audio backstories are more likely to make typographical errors, as the correct spellings are not immediately available to the workers. Table 1 reports the number of queries in which a spelling correction was made (based on comparing the default query to the spellchecked version of the query), and the magnitude of the error using edit distance. The results show that as the length of the backstory increases, so too does the percentage of queries with a typographical error (with the exception of Short Text).

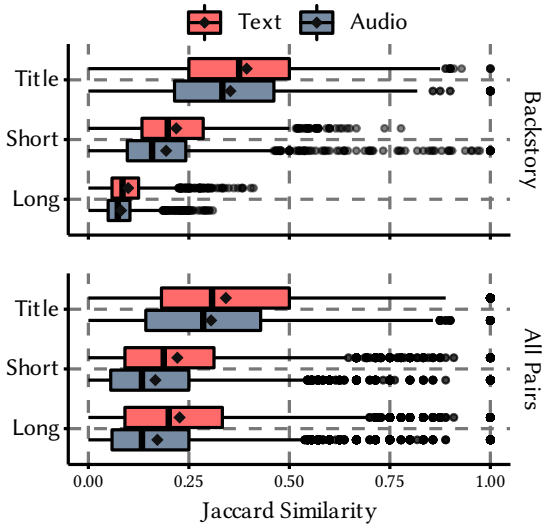

Figure 2: Jaccard similarity between query variations and corresponding backstories (top); and between all pairs of queries for each topic (bottom).

Table 1: Percentage of query variations with typographical errors, and the average edit distance between the misspelled and corrected queries, with the latter computed across the subset of queries with one or more typographical errors.

Length	Text			Audio		
	Title	Short	Long	Title	Short	Long
% w. Typo	13.47	11.13	14.05	13.88	14.78	22.22
Mean Edit	1.25	1.35	1.34	1.40	1.36	1.56

4 VARIANCE IN RANKINGS

The previous section considered the sets of terms present in the backstories and generated queries, showing that, in general, the length and delivery mode of the backstory has an influence on query diversity. We now explore whether these effects result in measurable differences in terms of ranked retrieval outcomes. Recent work by Liu et al. [18] demonstrates that both human and machine generated query variations can be highly effective, even though markedly different in both the terms used and the rankings produced by those variations. A different thread of work has examined the variance in judgment pools due to both system and query variations [2, 22, 23, 36]. Given that previous context, this section explores these ideas from the perspective of *how* the query variations were generated, examining whether the diversity observed in the queries affects what occurs during retrieval tasks.

Systems and Configurations. We indexed the CC-News-En document corpus with Anserini [34], and used two models for retrieval: a bag-of-words BM25 model with the default parameters $k_1 = 0.9$ and

Table 2: Mean RBO ($\phi = 0.99$), stratified by backstory length and mode. Each entry in the table is an average over all possible ranked list pairs for each topic, over all topics, and over two different retrieval systems.

		Text			Audio		
		Title	Short	Long	Title	Short	Long
Text	Title	0.30	0.18	0.19	0.26	0.14	0.14
	Short	—	0.21	0.21	0.15	0.15	0.15
	Long	—	—	0.23	0.17	0.16	0.16
Audio	Title	—	—	—	0.24	0.13	0.13
	Short	—	—	—	—	0.12	0.12
	Long	—	—	—	—	—	0.13

$b = 0.4$ [26]; and a proximity-based sequential dependency model (SDM) [21] with default weights of 0.85, 0.10, and 0.05 for terms, ordered windows, and unordered windows, respectively. We employed the query set already described in Section 3, and generated runs to a depth of $k = 1{,}000$.

Retrieval Similarity. Following Liu et al. [18], we use *rank-biased overlap* [32] to measure the similarity between ordered lists, setting the top-weightedness parameter $\phi = 0.99$, corresponding to an expected user viewing depth of 100. Table 2 reports mean RBO, faceted by backstory length and mode, with the averages computed across all pairs of ranked lists for each topic (from different queries), across all of the topics, and across the two retrieval systems. The six blue values denote *intra-mode* RBO values, with both length and modality held constant; whereas the three purple values denote RBO values that compare between the Text and Audio modes for each of the three backstory length.

The results in Table 2 are consistent with those in Section 3, and confirm that there is greater diversity (lower RBO scores) in the result lists obtained via the Audio backstories and resultant queries than for the Text-seeded queries. Furthermore, Short and Long backstories generate more result variance than Title ones.

Simulating Judgment Pools. The final experiment explores the number of documents that would need to be judged at a range of fixed pooling depths d, again seeking to isolate the effect of the length and modality of the backstory presentation. Like RBO, pool size at depth d is a way of quantifying the consistency (or lack thereof) of the retrieved documents, and in turn, of the query variations provided as input to the retrieval systems. Figure 3 plots the average (over topics) of the pool size as a function of the pool depth d, with the data again stratified by backstory length and modality, and with the pools for each topic formed as the union of the runs from the two systems (BM25 and SDM) and all associated query variations.

The overall trends shown in Figure 3 corroborate previous experimentation with pooling across query variations [2, 22, 23]. There is a clear relationship between pool depth and the number of unique documents included in the pool associated with each topic, with no evidence that the different queries rearrange the same sets of documents (which would result in a tapering off of the rate of growth of each line). In this regard Figure 3 confirms previous observations

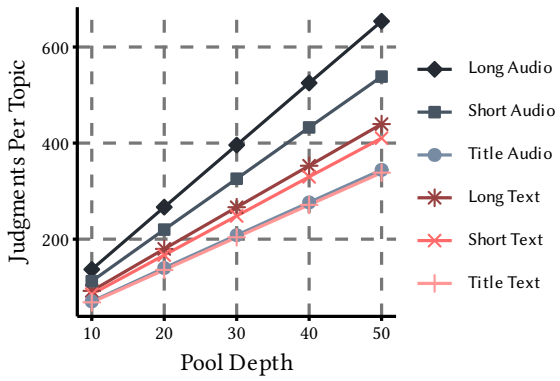

Figure 3: Growth in pool size, computed as the average number of unique documents surfaced per topic, averaged over all 173 topics, broken down by backstory length and backstory delivery mode, and shown as a function of pool depth d.

that query variations using a small number of systems is a source of candidates (that is, documents that might be relevant in response to the original information need captured by the backstory) that is at least as (or even more) useful than single-query runs generated by multiple different systems [22, 23].

Moreover, the length and modality of the backstory are also important factors. Backstories communicated as audio clips result in queries that retrieve a larger pool of documents than backstories presented as text; and long backstories (and the longer queries that they provoke) also generate more diverse pools. These various effects mean that designers of retrieval experiments must ensure that sufficient judgment capacity is available if their experiment is to demonstrate the effects they believe are present.

5 CONCLUSION AND FUTURE WORK

We have examined the effect that backstory modality and length have on the process of eliciting user query variations from crowdworkers. The CC-News-En collection includes backstories of three different lengths, which were provided to crowdworkers via either audio-only sound clips, or via textual images (that is, non-cut/pasteable written words) [20]. We have found that audio presentation of backstories consistently resulted in a broader spectrum of queries emerging, with less overlap to the original backstory text, less overlap to other workers' queries, and greater diversity of documents retrieved when the queries were evaluated using two standard retrieval mechanisms. There were also effects – less pronounced, but nevertheless observable – arising from the length of the backstory, with longer backstories tending to inspire longer queries that were more diverse in terms of words used and also in terms of documents retrieved. These effects need to be factored in when designing the structure (and anticipating the cost of) of retrieval experiments.

One avenue for future investigation is how variation in the *content* of the backstory can lead to higher (or lower) variance from crowdworkers. While the backstories in the CC-News-En collection have minor variations in their content [20] (mostly due to the information density of differing lengths), it would be interesting to

determine the extent in which *backstory variations* impact query generation. It would also be interesting to examine the relationship between crowdworker demographics [33] and submitted query variations, to develop a better understanding of alternative sources of variance in query formulation tasks.

Acknowledgments. We thank Guido Zuccon and Binsheng Liu for useful discussions related to this research. This work was supported by the Australian Research Council (project DP190101113).

REFERENCES

[1] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proc. SIGIR*, pages 603–604, 2006.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. SIGIR*, pages 625–634, 2015.

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.

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

[5] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect of multiple query variations on information retrieval system performance. In *Proc. SIGIR*, pages 339–346, 1993.

[6] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. & Man.*, 31(3): 431–448, 1995.

[7] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. Aust. Doc. Comp. Symp.*, pages 1.1–1.8, 2017.

[8] R. Benham, L. Gallagher, J. Mackenzie, T. T. Damessie, R.-C. Chen, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2017 TREC CORE track. In *Proc. TREC*, 2017.

[9] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. Towards efficient and effective query variant generation. In *Proc. DESIRES*, pages 62–67, 2018.

[10] 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.

[11] R. Benham, J. Mackenzie, J. S. Culpepper, and A. Moffat. Different keystrokes for different folks: Visualizing crowdworker querying behavior. In *Proc. CHIIR*, pages 331–335, 2021.

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

[13] M. Catena and N. Tonellotto. Multiple query processing via logic function factoring. In *Proc. SIGIR*, pages 937–940, 2019.

[14] 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.

[15] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth evaluation lab 2015. In *Proc. CLEF*, 2015.

[16] B. Koopman, L. Cripwell, and G. Zuccon. Generating clinical queries from patient narratives: A comparison between machines and humans. In *Proc. SIGIR*, pages 853–856, 2017.

[17] G. Leroy and D. Kauchak. A comparison of text versus audio for information-comprehension with future uses for smart speakers. *J. Am. Med. Inform. Assoc.*, 2 (2):254–260, 2019.

[18] B. Liu, N. Craswell, X. Lu, O. Kurland, and J. S. Culpepper. A comparative analysis of human and automatic query variants. In *Proc. ICTIR*, pages 47–50, 2019.

[19] X. Lu, O. Kurland, J. S. Culpepper, N. Craswell, and O. Rom. Relevance modeling with multiple query variations. In *Proc. ICTIR*, pages 27–34, 2019.

[20] 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.

[21] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR*, pages 472–479, 2005.

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

[23] 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.

[24] 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.

[25] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. F. Jones, M. Lupu, and P. Pecina. CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Proc. CLEF*, 2015.

[26] S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trnd. Inf. Retr.*, 3:333–389, 2009.

[27] T. Saracevic and P. Kantor. A study of information seeking and retrieving. III. Searchers, searches, and overlap. *J. Amer. Soc. Inf. Sc. Technol.*, 39(3):197–216, 1988.

[28] H. Scells, L. Azzopardi, G. Zuccon, and B. Koopman. Query variation performance prediction for systematic reviews. In *Proc. SIGIR*, pages 1089–1092, 2018.

[29] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. LambdaMerge: Merging the results of query reformulations. In *Proc. WSDM*, pages 795–804, 2011.

[30] I. Stanton, S. Ieong, and N. Mishra. Circumlocution in diagnostic medical queries. In *Proc. SIGIR*, pages 133–142, 2014.

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

[32] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Sys.*, 28(4):20.1–20.38, 2010.

[33] I. Weber and C. Castillo. The demographics of web search. In *Proc. SIGIR*, pages 523–530, 2010.

[34] P. Yang, H. Fang, and J. Lin. Anserini: Reproducible ranking baselines using lucene. *J. Data Inf. Qual.*, 10(4):1–20, 2018.

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

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