Where Do Queries Come From?

Marwah Alaofi
RMIT University
Melbourne, Australia

Luke Gallagher
RMIT University
Melbourne, Australia

Dana McKay
RMIT University
Melbourne, Australia

Lauren L. Saling
RMIT University
Melbourne, Australia

Mark Sanderson
RMIT University
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Damiano Spina
RMIT University
Melbourne, Australia

Ryen W. White
Microsoft Research
Redmond, WA, USA

ABSTRACT

Where do queries – the words searchers type into a search box – come from? The Information Retrieval community understands the performance of queries and search engines extensively, and has recently begun to examine the impact of query variation, showing that different queries for the same information need produce different results. In an information environment where bad actors try to nudge searchers toward misinformation, this is worrisome. The source of query variation – searcher characteristics, contextual or linguistic prompts, cognitive biases, or even the influence of external parties – while studied in a piecemeal fashion by other research communities has not been studied by ours. In this paper we draw on a variety of literatures (including information seeking, psychology, and misinformation), and report some small experiments to describe what is known about where queries come from, and demonstrate a clear literature gap around the source of query variations in IR. We chart a way forward for IR to research, document and understand this important question, with a view to creating search engines that provide more consistent, accurate and relevant search results regardless of the searcher's framing of the query.

CCS CONCEPTS

• **Information systems → Information retrieval query processing**.

KEYWORDS

Information retrieval, query formulation, adversarial information environments

ACM Reference Format:
Marwah Alaofi, Luke Gallagher, Dana McKay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. 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, 13 pages. https://doi.org/10.1145/3477495.3531711

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

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531711

1 INTRODUCTION

How do you cook Beef Stroganoff for a dinner party? When 108 experimental subjects reported the initial query each used to answer this information need, even after case normalization, stemming, and spell correction, 19 different ways of wording the query were recorded.[1] Each form resulted in different effectiveness scores. Our offline test collections typically record one wording per topic, even though it has been shown that the words a searcher chooses will not only affect the accuracy of their search, but the type of results that are returned [84, 105]. Questions of dinner party food may seem trivial, but what if a searcher was looking for something more consequential: someone to hire, a political view, or a medical treatment? Our SIGIR community researches ranking algorithms, user interaction, search domains (e.g., Web, enterprise, etc.), and the normalization of queries, but do we consider why a query is worded in a particular way and how that wording affects search?

We contend that this topic is of increasing importance:

- There is a decline in trust in public institutions,[2] which means many more people research (and consequently search) for their own answers, often seeking to reinforce or refute existing views.
- By and large, people trust search engine results [82].
- The queries people use are getting longer [108], and longer queries will by their very nature have more varied forms.
- Search Engines (SEs) are operating in an *adversarial information environment.* Strongly held but briefly expressed views, blatant misinformation, and commercially motivated sentiments proliferate. There is evidence that others try to influence how we search [98].

Recent work has demonstrated that information encounters on social media are the genesis of a significant proportion of personally important view changes, and search comes later in the process [75]. There is evidence that the words used to present information influence search terms as well as many other contextual factors. Ultimately, the language in which results are expressed further influence searchers' engagement with search results [112].

Given the importance of the actual words typed in a search box, it is perhaps surprising that this is a perspective of the search process that has had little attention from the Information Retrieval (IR) community. User interaction with search has been researched extensively, but the focus is on understanding query behavior when interacting with a SE such as query reformulation in search sessions [26, 49], or query auto-completion [24]. There has been some

[1] The data for the work by Bailey et al. [10] can be downloaded at https://researchdata.edu.au/from-uqv100-test-query-variability/1307620
[2] https://www.un.org/development/desa/dspd/2021/07/trust-public-institutions/

work on query formation in Library and Information Science (LIS) [70, 96], much conceptual in nature. There is some empirical work asking participants to generate queries in response to a stated information need (e.g., [57]), the work is limited to artificial information needs, an approach that results in less user curiosity about the results and therefore likely changed behavior [54]. A handful of test collections were created that examine *query variations* [9, 23], but that research has focused on the impact of the variation on search results, rather than a query's origin. While such research has inspected particular factors affecting query formulation, there remains a gap in understanding connections across the informational, social, technical, and cognitive factors that affect query term selection. In short, a key question remains unanswered: *where do queries come from?*

In this perspectives paper, we first discuss why this question is important now, highlighting the rise of adversarial information environments, and an increased focus on equity in technology generally. We then examine the question from a variety of perspectives. We examine existing research on variability (Section 3) and demonstrate that – even for relatively simple needs and in homogeneous cohorts – there is high variability (Section 4). As a means of outlining the potential for work in this topic area, we examine variability from different perspectives to describe what is already known about query formation, and demonstrate that there is an important gap (Sections 5–7). Finally, we point to future research directions to address this gap (Section 8), and draw conclusions about the importance of this topic in Section 9.

2 WHY NOW

Query variability has always been a feature of search, query formation has been discussed in the LIS literature, but there is a long-standing disjunct between LIS and IR [41]. So why has the ways in which queries are generated become important *now*? There are two reasons: the increasingly adversarial nature of the information environment in which we operate, and increased understanding of the impact of human diversity on technology use.

One major change to the information environment is the potential for adversarial actors, whether in a commercial or political space, to deliberately word documents, presentations, or social media postings to nudge searchers to use particular words or phrases in their query. It has been suggested that such actions could lead searchers to the website of particular organizations which may then disseminate information that is in the interest of that organization, particularly for niche topics, so-called 'data voids' [45]. Tripodi [98], inspired by a qualitative study of communities with a conservative Christian worldview, described the detailed research conducted by that community in order to determine the veracity of content encountered online. However, the way queries were constructed could be influenced by the particular wording of that very content. Tripodi showed queries where the switching of one word could alter retrieval results from emphasizing right-wing to left-wing content [98]. Searchers have long been known to trust Web search engine results [82]. A searcher unknowingly influenced to query for a particular view may not seek alternates. Given the potential for information encounters to engender view changes [75], and the sheer volume of misinformation online, search engines can and must play a role in supporting searchers' quests for accurate information. To do this, though, we need to understand how searchers go from encountered information to search terms.

Another rationale for re-evaluating the importance of query generation is the new recognition of the importance of diverse voices in constructing IR systems and test collections. Much of the existing work on relevance has been done by field experts [86], who do not represent the diversity of searchers who actually use search systems. Sometimes this lack of diversity has resulted in negative representation of minority groups by search systems [79]. Beyond just representation, we know that when a group is not considered in design and testing, the resulting product will often not reflect their needs [66, 85]. This may mean that certain groups get search results that are less well suited to their needs, simply by virtue of the way they and their community search. Apart from the work detailed in Section 3, incorporating query variability in test collections is a priority that our community has largely ignored. As popular as test collections are, there remains a substantial gap in the predictions made by such offline datasets and the online realities of evaluation with a working system. While this is pure speculation on our part, it could be that the one reason for this disconnect between offline and online evaluation is the lack of query variability in test collections, partly because these test collections are based on a homogeneous sample of searchers. Understanding how different searchers construct queries could help make search engines more effective for those searchers.

Query variations give us a lens onto some of the factors that may be at play in query formulation: if we can tie query variations to people and information, we may understand how individual factors and background information affect what ultimately gets typed into a search box. In the next section, we address what is already known about query variability in IR.

3 QUERY VARIABILITY

It has been established that different searchers formulate different queries from the same underlying information need. This is referred to as *query variability*. The reasons for this variability remain an open question. Research to date has primarily been examined from the perspective of test collection creation, and analyses of retrieval experiments subsequently carried out using those collections.

3.1 Test collections created for query variability

The effect of multiple query variations on system effectiveness analysis was first investigated in the TREC-8 Query Track [23], where five participating groups generated queries based on a set of fifty topic statements. The conclusion was that topics – and specific queries that deal with the same topic – lead to extreme variability in system effectiveness scores, in contrast to different retrieval systems which were only somewhat variable. However, the track was not pursued beyond its initial two years of running.

Bailey et al. [9] studied whether variation in query formulation alters retrieval effectiveness, creating a new collection based on 180 TREC topics from different tracks, chosen to represent different levels of search task complexity. The topic statements were re-written into backstories, which were shown to crowdsourced workers who were asked to report what their first search query

would be. Analysis of system effectiveness demonstrated that the level of variation arising from different searcher query formulations was substantially higher than variation from differences in either search topics, or retrieval algorithms.

Moffat et al. [78] studied the impact of query variability on the completeness of relevance judgments. Test collections are typically constructed using pooling [101]. Introducing query variability potentially adds to the number of documents in the pool, since each participating system will generate candidate documents in response to multiple queries per topic. The analysis demonstrated that diversity in the documents that need to be judged that arises from query variability is at least as substantial as that arising from system variability, and critically, that previous test collections were problematic for the rigorous study of query variability, since running new variant queries for a topic would introduce many new unjudged documents into the evaluation process.

Consequently, the UQV-100 collection was developed [10]. Information need backstories were written for 100 topics from TREC Web tracks, which were shown to crowd workers. They were asked to indicate what query they would use, resulting in an average of 58 normalized queries per backstory. Corresponding relevance judgments were obtained for the top-10 (as a minimum) pooled documents for each query variation. The collection was used to study the *consistency* of search systems in the presence of query variations [11]. Further test collections have since been enhanced with user query variations: Benham et al. [18] created query variations for 250 topics used in the TREC Core 2017 Track.

3.2 Experiments on query variation collections

Evidence that combining results from query variations boosts retrieval effectiveness [11, 16, 19], as well as the impact that query variability has on evaluation, led to interest in generating query variations automatically. Benham et al. [17] explored the use of a weighted random sampling process to generate query variations that, when fused and combined with indexing techniques, provide a better efficiency-effectiveness profile than relevance models. Breuer et al. [22] presented an evaluation framework for analyzing the extent to which simulated user query variations match real queries, based on factors such as query term similarities, shared task utility, and relative retrieval performance. The authors also proposed a new parameterized query variation simulation approach, and demonstrate that it has a higher fidelity to real queries than previous approaches. Human-generated query variations were compared with variations created using the Bing search engine's click-graph by Liu et al. [67], whose analysis showed that, while the system effectiveness of both approaches is comparable, the two types of variations display other subtle differences (e.g., the ranked lists of documents that each approach retrieves).

Penha et al. [84] explored the robustness of retrieval pipelines in the presence of query variation, motivated by the observation that recent neural-based retrieval approaches in particular may be brittle when evaluated using variant queries. Their analysis showed an average fall in effectiveness of 20% in nDCG@10 compared to performance on the original (single) queries available in the test collections. The work also grouped query variation into six categories: misspellings, naturality, ordering, paraphrasing, aspect changes, and generalizations/specializations, supporting more nuanced analysis into the impact of these different types of variations.

The impact of query variations on system evaluation was considered by Zuccon et al. [115], who proposed a mean variance evaluation framework that explicitly considers and separates query variation and topic variation, and supports the incorporation of risk sensitivity. They found that the new framework ranks systems differently from more standard test collection-based effectiveness evaluation approaches.

Throughout the development of test collections since the 1990s, there has been a steady stream of works examining the way that retrieval systems behave on those collections. Using ANOVA based methodologies, researchers have considered which factors most impact the effectiveness score of a retrieval algorithm. The sophistication of the ANOVA models has grown in the last few years [40]. It was long assumed that the way different topics were worded was the key driver of differences in scores between systems: a search for the origins of coffee beans results in different effectiveness scores from a search for ideal retirement locations in north west America. Culpepper et al. [32] showed that the variability observed between topics is not topic variability but variability in the way that the query is expressed. This is a key observation as it means that we have been focused on only one type of variability when building test collections. We have assumed that we need to study a large number of topics when in fact we need to study both a large number of topics and those topics expressed in a large variety of ways.

3.3 Conclusions so far

Query variability, where different query instances are used to represent the same underlying information need, can have a substantial impact on search systems, both in terms of system performance, as well as the effectiveness evaluation framework itself. The research so far has established that query variability substantially affects the accuracy of a SE, yet the number of test collections we have in our community to study this effect can be counted on one hand. The work has established that query variability exists, but it has not explained why it exists, what factors may influence such variability, or fully explored how search can alleviate or potentially even exploit such variability. Understanding where the variability comes from will likely give us insight into how to better design SEs to account for it.

4 EXPERIMENTS WITH VARIABILITY

While existing papers have detailed that the effectiveness of retrieval algorithms is susceptible to query variability, the variability has not yet been studied on more recently described supervised retrieval such as dense and learned-sparse methods, neither has it been shown on a commercial SE. We detail experiments examining query variability in these two situations.

4.1 Query variability in supervised retrieval

The purpose is to gain a preliminary understanding of retrieval consistency for a set of query variations, where the same query is compared across different inverted indexes – one traditional and one augmented via a learned sparse representation. A second aspect

is to identify if, by default, a learned sparse index promotes or rescinds the ability for a query variation to return relevant documents that were not previously identified by any other query variation for the given topic. This may serve to guide future research efforts towards a framework that combines user query variations and fairness-aware retrieval techniques within adversarial information environments. To explore these properties, we conduct a preliminary study using a recent learned sparse representation model HDCT [33]. HDCT is an extension of DeepCT [34] which uses contextualized embeddings from a transformer model with a regression layer to learn term weights based on query terms that are present in relevant documents.

Experimental setup. The ClueWeb12C corpus was used [33], i.e., a random 10% subset of the ClueWeb12B corpus containing 5,249,243 documents. The topic set used is the UQV-100 collection [10], consisting of query variations for each of the 100 topics from the TREC 2013–2014 Web tracks. In addition to the ClueWeb12C corpus, judged documents from the UQV-100 collection were also indexed. Two indexes were used: (i) a traditional term-frequency based index (TF), where the source documents were indexed without any further processing, and; (ii) a HDCT index (HDCT-Title) that utilized the title field in source documents in a similar manner to that described by Dai and Callan [33]. The process for HDCT segmented documents into 16,527,770 passages of 300 terms suitable as input for DeepCT. The document `title` field was used as a query pseudo-signal for the training phase of learning paragraph term weights. Training was performed for 100,000 steps with a batch size of 8 and learning rate of 2e−5. Passage term weights were aggregated with the `sum` method and a scaling coefficient $N = 10$ was applied. Indexing and retrieval was performed using the Anserini toolkit [111]. Indexes were constructed with Krovetz stemming and stop words retained. The following parameter configurations for retrieval were used: BM25 with $k_1 = 0.9$ and $b = 0.4$; QL with $\mu = 1000$; SDM with $w_t = 0.85$, $w_o = 0.1$, $w_u = 0.05$.

Results. Figure 1 displays the consistency in retrieved documents for various system-index combinations using Rank-Biased Overlap (RBO) [104]. RBO ($p = 0.9$) was computed to reflect the depth of the judgment pool and the traditional first Search Engine Result Page (SERP). In general, when running the same query across system-index combinations, the results show low consistency with relatively few exceptions. This gives some indication of the extent to which index variation plays a role in candidate results for learned sparse representations and may be an important factor when considering new frameworks that include a user query variation component. This presents another dimension of variability, and for continually changing adversarial environments any system change has a potential impact on users.

Figure 2 aims to give insight to the retrievability aspect of each retrieval method. The results show the number of variations in which a query variation (within a topic) retrieved a relevant document that was not retrieved by any other variation (within a topic). As can be seen, there are many instances of documents being returned by one variation that no other variations could obtain. Of interest here is that, in the case of initial query formation, users are led down distinct information seeking paths of document exposure. We hypothesize there exists an equivalent potential for retrieval of

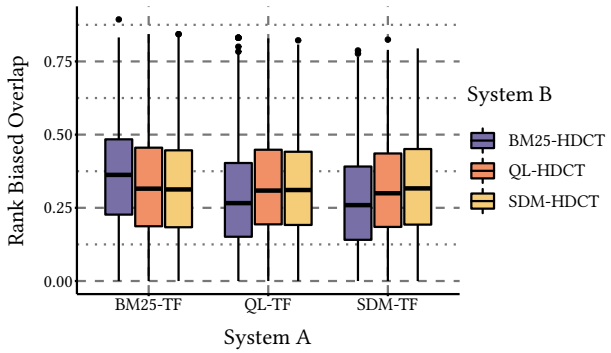

Figure 1: Evaluation results for RBO ($p = 0.9$). Each retrieval method on the traditional inverted index (TF) is compared against all other retrieval methods on the learned sparse representation index (HDCT-Title).

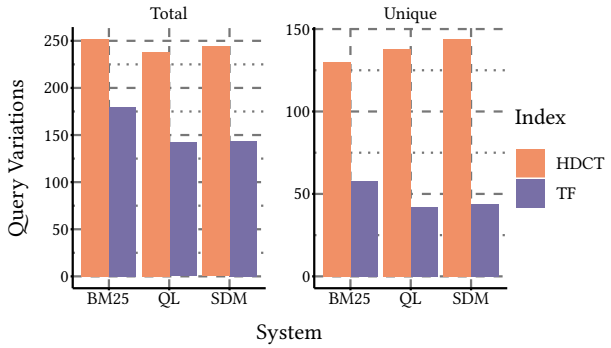

Figure 2: Number of query variations returning a relevant document that was not retrieved by any other variation within the same topic. Left panel displays the total occurrences; right panel shows occurrences that are unique to a given index, by a system.

unique adversarial documents given a query variation and future work may need to consider adversarial labeled datasets with user query variations [45].

4.2 Variability and a commercial search engine

In this second experiment, we aimed to understand how query variations affect modern SEs. Using the UQV-100 dataset [9], we submitted the query variations of 100 topics to a commercial SE and measured the consistency of search results using RBO.

Experimental setup. We used the query variations provided in the UQV-100 test collection and issued the 5,744 variations of the one hundred topics with an average of 57.44 query variations per topic to a widely used commercial SE: Google Search.[3] For each submitted query, we captured the retrieved result URLs and their corresponding rankings from the first SERP. Only URLs of Web documents, videos, and top stories were captured. Advertisements and other less common SERPs items such as social media posts

[3] https://www.google.com

were excluded from the collected URLs. Every query was submitted using a new browsing instance, to prevent user profiling.

Evaluation was performed with RBO ($p = 0.9$) to measure the retrieval consistency across query variations for the SE employed in our experiment. The RBO calculation is performed by comparing both sets of result URLs that were returned by a pair of query variations. Each topic RBO score is reported as an average RBO score of all topic variation pairs.

To understand the effect of query variation type on retrieval consistency – for example, how consistent the model is given a variation in query naturality (keyword vs. natural language queries) as opposed to a variation in word ordering – we used the six query variation categories proposed by Penha et al. [84]: *aspect change*, *specialization or generalization*, *misspelling*, *paraphrasing*, *naturality* and *word ordering*. The authors manually annotated 650 randomly selected query variation pairs and assigned them to one or more categories. As UQV-100 query variations are spell-corrected, the *misspelling* category was not included in the annotated set.

Submitting the *same query* multiple times to a commercial SE may generate different sets of results, e.g., due to dynamic updates. Therefore, it is important to consider the possibly inconsistent nature of SEs when interpreting our findings, i.e., the question of "how much of the inconsistency, if observed, is attributed to query variations rather than the inconsistency of SEs" becomes crucial. To quantify inconsistency, we randomly selected 100 queries from all topics and ran them 10 times at 30-second intervals, giving five minutes total run-time for each query, which exceeded the average time needed to run all query variations for a given topic.

Results. Figure 3 shows the distribution of RBO scores across variation categories, as well as the average RBO score distribution of all query variations across UQV-100 topics. The distribution of same-query RBO scores is also shown for comparison. Results show a clear reduction in retrieval consistency across variation categories when compared to the measured consistency for same-query search result sets. That is, the average topic RBO ranges from 0.09 to 0.66 with a mean of 0.33, which is well below the same-query average RBO score of 0.97.

When considering variation categories, the lowest consistency in RBO is observed in *aspect change* and *specialization/generalization* variations. This is not surprising given the expected change in query semantics. Retrieval consistency is, however, still low in other variation categories. For example, the pair of query variations *"drug treatment for schizophrenia"* and *"schizophrenia drugs treatments"* – categorized as a change in query naturality – have an RBO score of 0.83 (the overall category RBO median is 0.42). Such minor variations in queries generally have no impact on retrieval consistency in traditional IR. Our preliminary results, however, suggest that a modern commercial SE is likely to be more sensitive to minor changes in queries, and that more emphasis should be given to maintaining consistency across query variations.

4.3 Summing up

The results of these initial analyses demonstrate that query variability has a notable impact on the retrieval effectiveness and consistency of state-of-the-art retrieval algorithms and, as of early 2022, a widely used commercial SE, showing the presence of the issues that this paper considers in both modern commercial and research retrieval systems.

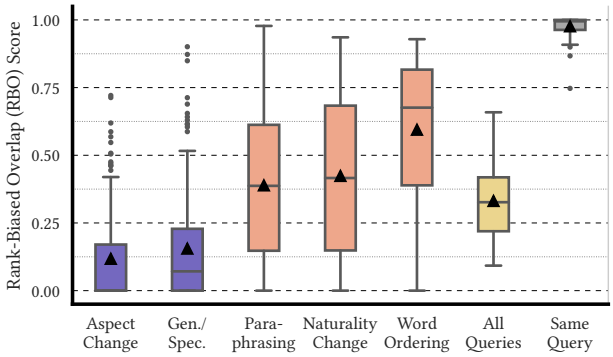

Figure 3: RBO distribution given semantic (purple) and syntactic (orange) variation categories, as well as all query variations within a topic (yellow) and *same query* runs at different intervals (gray). Lines and triangles represent medians and means, respectively.

5 INFORMATION SEEKING MODELS

One place we might anticipate discussion on where queries come from is in the literature on information seeking. Most models center on search, and therefore queries are an important part of these models. It is perhaps surprising, then, how little most of them have to say about query formation.

Of the major models (Ellis [39], Marchionini [71], Kuhlthau [61], and Wilson [109]), only Marchionini mentions query formulation specifically. This mention is made descriptively, however: it is noted that until the late 1990s, much query formulation was done through the use of intermediaries, and that the emergence of what were, at the time, new search tools meant that users would be formulating their own queries. Marchionini and White [72] expand on this discussion in a later paper, which formalizes the encoding of a query as the 'express' process, and note that there can be semantic differences between the query as a representation of user needs, and the needs themselves. Again, though, the issue of how a user makes the leap from semantic representation to query is skirted.

Ellis' and Kuhlthau's models do not explicitly mention search, although search is clearly a part of each of them – Ellis discusses choosing search sources during monitoring [39], and Kuhlthau talks about 'gathering information' [61], which includes search. Each of these models, however, describes a process moving from needing information through interacting with a system or systems, to having what is needed. Kuhlthau offers some insight into factors that may affect query term choice by noting that information seekers are affected by their emotions and cognitive biases as they move through the information seeking process, and that it can provoke feelings of both anxiety – where information seekers feel they do not know what they are doing – and relief – when they feel they have an answer. Wilson's model expands on this, pointing to demographics, the information seeker's role in the search, psychological and environmental factors as affecting the information seeking process [109]. It seems likely that these factors will affect the ways in

which information seekers formulate queries, as well. This remains an open question, though, and one worthy of investigation.

Perhaps the model with the most to say on query formulation is Belkin et al.'s ASK model [15]. This model tracks information seeker needs from the moment they recognize an anomaly in their state of knowledge, until that anomaly is resolved. It notes that one of the core problems of IR is to 'represent what a user doesn't know'. That paper also notes that there is no 'real or true representation of knowledge' in IR, but that there are a range of representations of knowledge – including that of document authors and that of information seekers, which may not be the same [14]. It is again surprising, then, that they do not suggest uncovering how people arrive at the particular representation of knowledge or anomaly that they issue to a system, instead focusing on generating a dialogue between system and user where the representation of this gap might be collaboratively generated.

It is clear, then, that the issue of how queries are formed has been at least considered in IR for over 30 years. In a time where information intermediaries are decisively a relic of the past, though, how people choose the actual words they type into a SE (or issue to a spoken conversational system) is a question overdue for consideration. All of the models presented here rely on repeated interactions between user and system, and the opportunity to browse, see results that were not searched for, and expand information horizons as part of the information seeking process. With these capabilities, the exact search terms an information seeker types in for their first query need not be a complete limit on the types of material that will answer their question. This is not the reality of modern search, though, which is conducted on a SE which focuses on relevance, and where information seekers typically examine few results [92]. The search terms entered, then, are more important than ever. It is for this reason that there have been calls for information interfaces generally [74], and search tools specifically, that present diverse results, or results from diverse viewpoints [46]. While these approaches may be a partial solution, understanding how people arrive at the query they select will allow us to both improve on these approaches, and better understand when they are necessary.

5.1 Potential research contributions

The information seeking models presented here are the basis for much of our understanding of how people go about understanding their need to find information, and then finding it. While the relevance of these models has been questioned (given when they were developed), sometimes research into ongoing validity concludes that the model is still valid, as happened for Kuhlthau et al.'s model in 2008 [62]. In other cases, the models have been expanded to include new behaviors or to model previously unexplored parts of the information seeking process, as Meho and Tibbo did with Ellis' model [76]. In this paper, we are calling for the latter: a re-imagining of the models, with a clearer focus on the specifics of query formulation. This re-imagining must be based in empirical research, both large and small scale, qualitative and quantitative. We argue that it is vitally important, in an age where information is so widely weaponized, to understand what elements of the preceding and subsequent stages of each of the major models play a role in the words that are typed into a search box, and whether there are any other features we can systematize in this way. Creating a model of query formulation specifically would serve not just IR researchers and practitioners, but also information scientists, social scientists, practicing librarians developing information literacy who could improve querying skills [87], and those developing policies and frameworks about issues including information interfaces [46], algorithmic bias [8], and misinformation [29].

6 CONTEXT AND COGNITION

The existence of query variations in the context of similarly described information needs shows us that factors other than the information need are at play when searchers generate queries. What are those factors, though, how do they relate to one another? In this section we address some of the potential factors that may be in play in query formulation, thus driving some of the variability identified by IR researchers. We review some of the library and information science research that has examined query formulation in the context of these factors.

6.1 Context

Contextual, psychological, and demographic factors can drive variability in how users formulate queries [43]. Variation in query formulation can substantially affect search result quality and consequent user satisfaction; however, little research was located that explicitly addresses these factors. Contextual factors including task complexity, topic specialization, and topicality have been investigated [5]. However, other potentially significant contextual factors have not been investigated, such as the nature of the information encounter that drives a searcher's information need, or the agenda that motivates a searcher's query. Key psychological factors that may influence query formulation include an individual's pre-existing knowledge of – and beliefs about – the search topic, cognitive style, personality, and cognitive biases.

Of these, recent research has addressed the role of cognitive biases in information seeking and retrieval [5, 58, 105]. Cognitive biases are systematic errors in thinking that are used to simplify judgments and decision-making but may instead undermine the quality of these processes. In the context of information search, cognitive biases can derail a search process by, for instance, encouraging users to frame queries in a positive rather than negative way, and to accept information that accords with prior beliefs. However, little is known about the role of such biases in formulating the first query posed to a SE. There are also some examples of research addressing predictors of retrieval effectiveness, but not query formulation. For instance, Ford et al. [43] found that high retrieval effectiveness was predicted by male gender, high self-efficacy, low topic complexity, and an image-based cognitive style. Demographic factors including age, gender, and political orientation are also likely to impact on query formulation but no research was located that has explicitly addressed this.

Other contextual factors that may affect searcher and SE performance include the input method for the search, e.g., conversational queries are likely to differ from typed queries, and queries typed on a mobile phone from those typed on a keyboard. Similarly, engagement with results is likely to be different based on device, but also on the level of attention a searcher is able to give their task.

Searchers demonstrably perform better when they are focused on their task [110], for example. Understanding how device, input method, and other task factors affect query formulation could improve IR performance dramatically, especially where much of the context is machine detectable.

6.2 Functional fixedness

One cognitive bias (largely unexplored in IR) that may play an important role in query formulation is *functional fixedness* [36], where people are limited to using an object in the way that it is traditionally used. Functional fixedness applies to a range of physical tools and digital tools, from hammers to search systems (it is this phenomenon that is described by the phrase 'to someone with a hammer, everything looks like a nail'). It has been shown to limit human creativity and delay problem solving [1, 20]. Experiments on functional fixedness provide a way to understand an object's psychological structure – the behavioral possibilities and affordances associated with it [113]. Fixedness is not present in young children [44], who have not yet developed mental models of how objects should be used, and may be affected by the nature of previous experience, with one study showing that more variation in prior experiences leads to less fixedness [42].

For SEs, experienced users may consider that they can only be used a certain way and formulate queries accordingly (e.g., as short key phrases with precise terminology [3]). In several studies, novice and experienced searchers have been shown to exhibit significant differences in their search behavior and search outcomes, including in the formulation of queries [3], usage of search tactics [47], and overall task performance [64].

6.3 Language and cognition

When considering language, it is important to ask 'what is a question?'. When MacKay asked this in the 1960s [70], he concluded that a question had to have an 'organizing function', i.e., a way of expressing either what the searcher does not understand (indicatively meaningful), or have meaningful outcomes that change dependent on the answer (interrogatively meaningful). Both question types are important, but those that are indicatively meaningful but do not match the facts in the world – for example, 'where does the sun go at night? The sun doesn't go anywhere, it's the earth that turns'. Both reflect the highest level of confusion on the part of the searcher, and are more likely to be prone to linguistic and conceptual variation. This also means that in this state, the searcher is particularly suggestible, whether by a helpful (though potentially biased) reference librarian [96], or by a malicious information adversary. Early work on cognitive models and information transfer [14] note that when people try to answer questions using a system, the system acts as an intermediary between the questioner and the authors of texts that may answer the question, and this has implications for language – on the part of both system and searcher – and trust in the intermediary that are under explored.

As noted above, queries may vary significantly in the language used by the searcher [7, 13]. Queries may be formulated using natural language, keywords, domain-specific language, and computer-based languages. Search queries may also vary in their valence: in other words, a given query may be posed positively or negatively [65]. For instance, "are vaccines safe?" is likely to deliver very different results to "are vaccines dangerous?" Contextual or individual factors may tacitly or explicitly influence both the valence and language of a query. At times users may consciously choose one query form over another in order to obtain particular results. At other times, query variation is accidental. Users are typically not aware that the system is highly sensitive to even slight variations in query formulation. For example, White and Hassan [106] showed that the presence of specific query terms – e.g., "help" (more likely to match pages affirming the effectiveness of a treatment), "can" (denoting possibility), "cure" (often matching spurious content) – reduces search result accuracy.

The review by Wacholder [103] on query formulation argues for more study of the cognitive and linguistic aspects of query formulation, and of the contexts in which query formulation may occur. Wacholder describes the difficulty in studying cognition as a significant bottleneck. Some of the linguistic literature, such as the importance of vocabulary and syntax, has been summarized by Wacholder [103], who points to work by (e.g.) Vakkari et al. [99], noting that vocabulary is a major problem when people are searching outside their domain of expertise. Studies of multilingual searchers also note that searchers may struggle when searching in a second (or third) language [93]. Taken together, these elements suggest that people unfamiliar with a topic may be more inclined to use readily available search terms, such as words in a piece of encountered information – but we do not know this for sure. Finally, the context in which people are searching – e.g., what prompted an information need, the physical environment in which they are searching, the device they are using, and their time, needs and preferences – are all pointed to by Wacholder as important (echoing Wilson's model [109]), but understudied.

6.4 Potential research contributions

The research survey here shows that queries may vary along a number of dimensions, including linguistic, contextual, and cognitive. While Wacholder proposed a basic model of query formulation, there are still many unanswered questions: how does an information seeker move from an 'anomalous state of knowledge' [15] to typing in a query? What factors in the search influence that process, and are they open to abuse by bad actors promoting misinformation? Most importantly, can we correct for any of this influence? These questions appear un- or under-examined in the IR literature.

One way we might address these questions is to isolate some of these factors by controlling the information encounter, and resulting information need. Examining the queries produced in different cognitive, contextual and linguistic situations may show us consistent variations between groups, and thus tell us something about the role of these non-information factors in where queries come from. Query variation will be documented as a function of individual differences (including demographic and psychological factors) as well as contextual factors (time of day, pre-existing knowledge and beliefs about the topic, etc.). Controlling for the search aim and information encounter will enable identification of determinants of variation in the way that queries are formulated. Further tweaks to this research could involve the use of eyetrackers to understand how people engage with different types or elements of information,

and relating the ways people scan information to the queries they ultimately construct.

Once we have an understanding of the factors that affect formulation, variation in search results as a function of different query formulations will be measured in order to distinguish innocuous variation from variation that produces misleading results. Reducing system sensitivity to slight changes in query formulation will maximize a user's search experience and agency by more reliably producing high quality search results.

Finally, learning from the query formulation strategies of search experts to benefit everyone has long been argued for [107]. However, given the presence of functional fixedness, there may be cases where novice searchers, unconstrained by past experiences (i.e., little or no fixedness) with SEs may formulate more effective search queries than seasoned experts. If those queries or query classes can be reliably identified algorithmically, one might up-weight the querying activity of these novices when training algorithms to, say, generate query suggestions or query auto-completions for matching queries; more study on this is required. Given the effects of the number of pre-utilization functions [42] (i.e., how many prior uses a person has for an object), the nature of people's longitudinal search behavior may also contribute to the extent of functional fixedness that they experience. Repetition of queries over time has been found to be common [97] and searchers with a low variance in their historical queries may be more susceptible to fixedness effects and require support to help overcome it.

Understanding the contextual, linguistic, and cognitive factors that affect query formulation will give us new understanding of the issues described in the next section.

7 QUERY (RE)FORMULATION

There is some existing research on how queries are formed. Some of this work comes from the library science domain, and is based on asking searchers what query they would generate given an information need scenario (this research is summarized in [103]). We already know that people are less invested in artificially constructed information than they are in their own information needs [54], so these studies, while interesting, offer limited insight into the true query formulation experience. We can supplement this understanding by examining query logs, understanding query elicitation, and looking at query reformulation.

7.1 Examination of query logs

One approach to understanding the queries information seekers generate is through transaction logs, though these logs of course cannot reveal searcher intention, nor searcher experience. Many log studies show that information seekers typically type in short, generic queries the majority of the time [63, 92]. This has usually been attributed to poor search skills, but more recent work demonstrating that searchers change their behavior in response to the underlying SE – even where the interface remains the same – suggest that this might be an economical, rather than a naïve strategy [73]. This study approach, while it reveals a little about query formation, does so without an understanding of the information needs of the searchers studied – only the representation they make of them. The challenge with query log studies is that they usually use a click graph to find equivalences in queries. Documents that are both retrieved and clicked on by users are assumed to be a means of identifying queries in common. Apart from one exception we are aware of [67], this largely unstudied topic would offer valuable examination in the context of query variability. One could examine if using a click graph is reliable enough to characterize the extensive variability of searchers, and the consequent impact on retrieval effectiveness. Such a study would need to combine both query log data and sets of query variations generated by other means.

7.2 Query elicitation from auto-completion

Query Auto-Completion (QAC) is one of the mechanisms SEs provide to assist users in formulating their queries [24, 35, 60]. As soon as a user starts typing into the query box, the QAC component suggests possible ways of completing the query. QAC aims to save time and effort (e.g., keystrokes) to the user by efficiently suggesting queries from a large set of query logs.

QAC relies on a number of features coded in query logs created by previous users' interactions [52]. These signals include: query frequency [12]; performance prediction [69]; query reformulations [53]; click-through features [25]; and context-aware features such as time [90], location [89], or spoken conversations [102].

A known side-effect of QAC is that this process can reinforce stereotypes and unintended biases present in query logs [56, 79, 81]. In the worst case, this can create harm or discrimination. But it may also lead to drift from the initial intent the user had: the queries suggested by QAC are a direct prompt of how to instantiate an information need [72]. Another side-effect of QAC is that it converges to less variation in the queries submitted to the SE, narrowing the query space: as users are more likely to use one of the suggested queries – already existing in the query logs – they are less likely to create a new query variation. This also impacts on the retrievability of documents [6], as new query variations may lead to the retrieval of documents that have not been retrieved before. Our experiments in Section 4.1 corroborate the latter (see Figure 2).

7.3 Query elicitation from librarians

That queries are a representation of something we do not know [15], and are difficult to generate [21, 94], has been seen in a range of works from IR and LIS. Early work suggested this problem might be addressed by systems interacting with information seekers to understand their information needs [15, 30, 80]. To this end, there were many studies of reference interviews to understand 'query elicitation' – the process of a reference librarian asking questions to understand how to best encode an information need as a search [30, 80]. The predominant findings of these studies are that information seekers initially specify what they need very loosely, and that a conversation with a reference librarian can increase the specificity of needs such as the topic of the search, acceptable sources for an answer, or how recent results must be. Of course, the biases of the intermediary in this situation become an additional layer over the query that is ultimately generated, potentially changing the information that users access [96].

7.4 Reformulation

A common search experience is one composed of a series of interactions with a SE, involving a *session* of query reformulations (w.r.t. a single information need) [49, 55]. Query log analysis has repeatedly shown that a significant fraction of users reformulate their initial query [83, 92]. Reformulated queries are unlike initial queries, however: by constructing an initial query searchers have surmounted one of the greatest difficulties in search (describing something they do not have [21]). The search results then give searchers new knowledge, both in terms of the vocabulary they might use and the type of results they might get.

Searchers' frequent query reformulation prompted research efforts such as the TREC Session Track [55] which helped to identify some of the inherent limitations (and unresolved system challenges [31, §5.3.2]) of the single query test collection view.

Query reformulation has been extensively investigated in many forms including user behavior analysis [28, 37, 50], models of cognition [59] and predictive models for various information prompting strategies such as query suggestion [91], query intent prediction [4, 27], query understanding [26], and more recently proactive reformulation [88]. It reflects a shift in cognition as user interaction unfolds, whereby the underlying information need is refined in order to elicit a different (but related) system response.

The way in which users reformulate their queries is an integral aspect of many query reformulation studies. Reformulations are often characterized by syntactical changes (e.g., word retention and removal) and/or intentional changes (e.g., specification and generalization). Huang and Efthimiadis [48] analyzed query logs and prior work to synthesize a taxonomy of transformation strategies. Their proposed strategies are mainly based on syntactical changes such as word reordering, removal, and use of abbreviations/acronyms. The authors evaluate the different reformulations using interaction metrics derived from click data and identify which reformulations are more effective (i.e., likely to cause clicks, given useful and not useful result sets).

Session-level factors such as task type, previous queries and user preferences and cognition styles have shown a potential effect on query reformulations. In a controlled lab experiment, Liu et al. [68] observed a significant effect of task type and objective, and subjective difficulty, on reformulation behavior. No statistical difference was observed between reformulation strategies and the cognitive abilities of participants. In another study, Kinley et al. [59] showed that users' query reformulation behavior is affected by their cognitive styles. For example, *analytic* users are more likely to add, remove and replace terms in query reformulations, compared to *wholist* users (i.e., users who retain an overall view of information) who tend to compose new reformulated queries. Jiang and Ni [51] investigated the factors that influence word changes in query reformulations. Their findings suggest that task and user characteristics may not directly affect users' decisions on word changes, but their previous queries may do. For example, users are less likely to remove a word that appeared frequently in previous queries.

While the literature has thoroughly studied different query reformulation strategies, factors influencing their use and their impact on the user's overall search experience, only a limited body of research has investigated the effect of modern SEs on query reformulations. Most of the available work has focused on studying the effect of SERP snippets and landing pages [37, 38, 91]. The question of how other SERP components such as direct answers, suggested searches and related entities influence user reformulation decisions remain under-investigated.

Sloan et al. [91] observed similarity between reformulated queries and the preceding query's clicked snippets and pages. Similar findings were observed by Eickhoff et al. [38]. In a follow-up eye-gaze tracking study, Eickhoff et al. [37] found that though 43% of all added query terms have occurred in previously visited SERPs and pages, the number of the added terms users actually paid attention to is much lower (21%). Chen et al. [28] attempted to understand fine-grained reformulation behaviors in the context of modern SEs. They examined the utility of reformulation entries such as the search input box, suggested queries, related entities and hot queries. Their findings showed that most reformulations are submitted using the input box (83.64%) and hot queries (10.95%) with a low utilization of other reformulation entries over session iterations. As an attempt to identify the inspiration source for each reformulation, users were asked to specify the source as being a landing page, snippets or other SERP components. Their findings suggest that 58.73% of reformulations were adopted from neither, with other SERP components contributing to 17.19% of the submitted reformulations, followed by landing pages (12.88%), and snippets (11.12%).

Though Chen et al. [28] initiated an interesting line of research, the explicit nature of their data collection, which relies on user-supplied input given within a 2-day window, may not precisely capture the source of inspiration for query reformulations. Users may find it difficult to articulate the source of their reformulations particularly as more time passes. Query reformulation is also a complex process and requires information synthesis, which may not be reflected at the term level, which is the the main signal being captured in existing work. More signals – explicit and implicit – could be utilized to gain better insights into the source of reformulations in the context of modern SEs.

7.5 Potential research contributions

Studies of query logs, query elicitation, and query reformulation can offer us some information about where queries come from, but each approach has its limitations.

The challenge with query log studies is that they usually use a click graph to find equivalences in queries. Documents that are both retrieved and clicked on by users are assumed to be a means of identifying queries in common. Apart from one exception we are aware of [67], this largely understudied topic would offer valuable examination of query variability. One could examine whether using a click graph is reliable enough to characterize the extensive variability of searchers, and the consequent impact on retrieval effectiveness. Such a study would need to combine both query log data and sets of query variations generated by other means.

It might be tempting to consider QAC as the solution to varying result quality caused by the query variation identified by the research detailed above. However, the experiments on commercial SEs suggest that the breadth of variation that searches are displaying goes beyond the normalization that QAC provides. The completion

often uses query logs as a means of identifying the types of query variations that could be harmonized in the completion. As detailed earlier, it is not clear that query logs are a sufficient record capable of capturing the broad variability that searches appear to be displaying. A means of better identifying query variations will be required, and then that extensive variation fed into QAC systems.

Besides QAC and query reformulation, there are other mechanisms to support the formulation of information needs. For instance, asking clarifying questions [2, 114], query rewriting for conversational search [77, 100], or auto-completion of voice queries [95].

The study of reformulation is the study of queries during a session. One line of research is to investigate whether within-session query changes reflect the same characteristics as query variability from distinct users for the same information need, and whether the interactions and information encounters or prompts that take place have similar or different influence on a users next interaction.

Query variability and session variability may have commonalities, so it would be interesting to explore whether session variability can help to inform us of the more general query variability. A crowd sourced session variability dataset may help to better understand some of these questions. Pairwise reformulation judgments could be a data driven way to reveal new patterns that extend the work of Penha et al. [84]. Further understanding of whether different patterns exist in different domains – for example, legal and medical search, where the cost of being misinformed is greater than a casual ad hoc query – is also important.

Each of these types of study, though, give us only part of the picture of how individual users with their own information needs form queries. We have some insight, but we still do not know, for example, the ways in which distraction and reformulation interact to result in the particular query a searcher will issue in response to an information need. This type of holistic understanding would allow us to design SEs to support searchers in meeting their actual information needs, and in preventing the worst damage caused by searches returning (and searchers using) misinformation.

8 WHERE NEXT?

The intention of this paper is to establish the following problem: we do not have a comprehensive picture of how searchers select the terms they type into a search box. We know that there are likely to be cognitive, linguistic, informational, and contextual factors at work in this process, but how they fit together remains unclear.

We can see traces of the impact of all of these factors in research on formulation conducted in library science, in query reformulation studies, and in studies of query elicitation. The largest traces seen in IR, though, are in the emerging stream of work on query variation: we now know that searchers will generate a wide range of variations for a single stated information need, and that those variations affect what is returned by search systems. In an adversarial information environment, it is a matter of urgency to ensure search systems cope with query variations that have been 'nudged' toward misinformation by bad actors. It is also important to ensure SEs are performing equitably: that searchers are getting equally useful results regardless of factors such as race or gender.

Understanding how these various factors fit together will require a mixed-methods approach of qualitative and quantitative work to narrow down the impact of various personal and environmental factors, and generate new models of query formulation. This will require information scientists, psychologists, human-computer interaction researchers, and IR specialists to work together, as Fidel has called for [41]. For the IR community, the ultimate goal will be to generate test collections reflecting the query variability we know exists, but with an understanding of *how* that variability comes to be, so that search systems can respond to underlying problems.

The generation of such test collections will result in further interesting challenges for the IR community: it cannot be ignored that adding a substantial number of query variations for each topic in a test collection will increase the amount of work required in generating the relevance judgments for that collection. Means of alleviating that work is itself a source of research opportunity [78]. Query variations could be formed through generative methods, such as those examined recently by Penha et al. [84].

With such collections in place, a range of studies can be conducted. Different query normalization techniques beyond current stemming or QAC approaches could be tested under a much wider set of variations. One could understand from the variant data which are the more or less popular variations that are tried, studying the relationship between popularity and effectiveness. With a rich collection of query variations included in test collections, there is also the potential for retrieval algorithms to identify the most successful variations and then harmonize the results with poorer variations, so that less well-formulated query variations can be rewarded with results from the most effective variation.

9 CONCLUSIONS

This perspectives paper has outlined a new challenge to the IR community: *Where do queries come from?*

New areas of IR research often arise from new types of document collections or algorithms. While various aspects of this question have been investigated, many of these are under-studied, and there is no overall understanding or framework of how users ultimately come up with the precise query that they choose to submit to a search system. We have outlined a range of different research perspectives, and identified distinct means by which this topic can be studied in future.

Potential research directions were outlined to identify approaches from the qualitative (e.g., psychological studies involving user experiments in a range of different contexts involving a broad range of individuals) as well as quantitative (e.g. large-scale algorithmic examinations of query logs, as well as retrieval algorithms) perspectives, all with the goal of improving retrieval effectiveness.

It is our hope that this perspectives paper spurs the community into examining this topic more extensively. The history of IR research has shown significant gains in improvement of search systems by starting with the query. Imagine how much further we could go if we understood the ways in which that query was created. This will require a new stream of research, one that begins with an Anomalous State of Knowledge and ends with a query.

ACKNOWLEDGMENTS

This work is partially supported by the Australian Research Council (CE200100005, DE200100064, DP190101113).

REFERENCES

[1] Robert E Adamson. 1952. Functional fixedness as related to problem solving: A repetition of three experiments. *Journal of Experimental Psychology* 44, 4 (1952), 288. https://doi.org/10.1037/h0062487

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484. https://doi.org/10.1145/3331184.3331265

[3] Anne Aula. 2003. Query formulation in web information search. In *Proceedings of the International Conference WWW/Internet*. 403–410.

[4] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM CIKM International Conference on Information and Knowledge Management*. 829–838. https://doi.org/10.1145/2661829.2661912

[5] Leif Azzopardi. 2021. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 ACM SIGIR CHIIR Conference on Human Information Interaction and Retrieval*. 27–37. https://doi.org/10.1145/3406522.3446023

[6] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM CIKM Conference on Information and Knowledge Management*. 561–570. https://doi.org/10.1145/1458082.1458157

[7] Judith L. Bader and Mary Frances Theofanos. 2003. Searching for cancer information on the internet: Analyzing natural language search queries. *Journal of Medical Internet Research* 5, 4 (2003), e31. https://doi.org/10.2196/jmir.5.4.e31

[8] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[9] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 625–634. https://doi.org/10.1145/2766462.2767728

[10] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A test collection with query variability. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 725–728. https://doi.org/10.1145/2911451.2914671

[11] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval consistency in the presence of query variations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404. https://doi.org/10.1145/3077136.3080839

[12] Ziv Bar-Yossef and Naama Kraus. 2011. Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*. 107–116. https://doi.org/10.1145/1963405.1963424

[13] Cory Barr, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of English web-search queries. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 1021–1030. https://aclanthology.org/D08-1107

[14] Nicholas J. Belkin. 1984. Cognitive models and information transfer. *Social Science Information Studies* 4, 2 (1984), 111–129. https://doi.org/10.1016/0143-6236(84)90070-X

[15] Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks. 1982. Ask for information retrieval: Part I. Background and theory. *Journal of Documentation* 38, 2 (1982), 61–71. https://doi.org/10.1108/eb026722

[16] Rodger Benham and J. Shane Culpepper. 2017. Risk-reward trade-offs in rank fusion. In *Proceedings of the 22nd Australasian Document Computing Symposium*. Article 1, 8 pages. https://doi.org/10.1145/3166072.3166084

[17] Rodger Benham, J. Shane Culpepper, Luke Gallagher, Xiaolu Lu, and Joel Mackenzie. 2018. Towards efficient and effective query variant generation. In *Proceedings of the First Biennial Conference on Design of Experimental Search and Information Retrieval Systems*. 62–67. http://ceur-ws.org/Vol-2167/paper4.pdf

[18] Rodger Benham, Luke Gallagher, Joel Mackenzie, Tadele Tedla Damessie, Ruey-Cheng Chen, Falk Scholer, Alistair Moffat, and J. Shane Culpepper. 2017. RMIT at the 2017 TREC Core track. In *Proceedings of the 26th Text REtrieval Conference*. https://trec.nist.gov/pubs/trec26/papers/RMIT-CC.pdf

[19] Rodger Benham, Joel Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting search performance using query variations. *ACM Transactions on Information Systems* 37, 4 (2019), 41:1–41:25. https://doi.org/10.1145/3345001

[20] Herbert G Birch and Herbert S Rabinowitz. 1951. The negative effect of previous experience on productive thinking. *Journal of Experimental Psychology* 41, 2 (1951), 121. https://doi.org/10.1037/h0062635

[21] Christine L. Borgman. 1996. Why are online catalogs still hard to use? *Journal of the American Society for Information Science* 47, 7 (1996), 493–503.

[22] Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating simulations of user query variants. In *Proceedings of the 44th European Conference on Information Retrieval Research*. 80–94. https://doi.org/10.1007/978-3-030-99736-6_6

[23] Chris Buckley and Janet A. Walz. 1999. The TREC-8 Query track. In *Proceedings of the 8th Text REtrieval Conference*. http://trec.nist.gov/pubs/trec8/papers/qtrack.pdf

[24] Fei Cai and Maarten de Rijke. 2016. A survey of query auto completion in information retrieval. *Foundations and Trends in Information Retrieval* 10, 4 (Sep 2016), 273–363. https://doi.org/10.1561/1500000055

[25] Fei Cai, Ridho Reinanda, and Maarten De Rijke. 2016. Diversifying query autocompletion. *ACM Transactions on Information Systems* 34, 4, Article 25 (Jun 2016). https://doi.org/10.1145/2910579

[26] Yi Chang and Hongbo Deng. 2020. *Query Understanding for Search Engines*. Springer. https://doi.org/10.1007/978-3-030-58334-7

[27] Jia Chen, Yiqun Liu, Jiaxin Mao, Fan Zhang, Tetsuya Sakai, Weizhi Ma, Min Zhang, and Shaoping Ma. 2021. Incorporating query reformulating behavior into web search evaluation. In *Proceedings of the 30th ACM CIKM International Conference on Information and Knowledge Management*. 171–180. https://doi.org/10.1145/3459637.3482438

[28] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of The Web Conference 2021*. 743–755. https://doi.org/10.1145/3442381.3450127

[29] Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. 2020. Overview of the TREC 2020 Health Misinformation Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference (NIST Special Publication, Vol. 1266)*. https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf

[30] Andy Crabtree, Michael B. Twidale, Jon O'Brien, and David M. Nichols. 1997. Talking in the library: Implications for the design of digital libraries. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*. 221–228. https://doi.org/10.1145/263690.263824

[31] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34–90. https://doi.org/10.1145/3274784.3274788

[32] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2021. Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems* 40, 1, Article 19 (Sep 2021). https://doi.org/10.1145/3470563

[33] Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*. 1897–1907. https://doi.org/10.1145/3366423.3380258

[34] Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting For first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1533–1536. https://doi.org/10.1145/3397271.3401204

[35] Giovanni Di Santo, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2015. Comparing approaches for query autocompletion. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 775–778. https://doi.org/10.1145/2766462.2767829

[36] Karl Duncker and Lynne S Lees. 1945. On problem-solving. *Psychological Monographs* 58, 5 (1945), i. https://doi.org/10.1037/h0093599

[37] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 13–22. https://doi.org/10.1145/2766462.2767703

[38] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan T. Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. 223–232. https://doi.org/10.1145/2556195.2556217

[39] David Ellis. 1989. A behavioural model for information retrieval system design. *Journal of Information Science* 15, 4-5 (1989), 237–247. https://doi.org/10.1177/016555158901500406

[40] Nicola Ferro, Yubin Kim, and Mark Sanderson. 2019. Using collection shards to study retrieval performance effect sizes. *ACM Transactions on Information Systems* 37, 3, Article 30 (Mar 2019). https://doi.org/10.1145/3310364

[41] Raya Fidel. 2012. *Human Information Interaction: An Ecological Approach to Information Behavior*. MIT Press.

[42] John H Flavell, Allan Cooper, and Robert H Loiselle. 1958. Effect of the number of pre-utilization functions on functional fixedness in problem solving. *Psychological Reports* 4, 3 (1958), 343–350. https://doi.org/10.2466/PR0.4.3.343-350

[43] Nigel Ford, David Miller, and Nicola Moss. 2001. The role of individual differences in internet searching: An empirical study. *Journal of the American Society for Information Science and Technology* 52, 12 (2001), 1049–1066. https://doi.org/10.1002/asi.1165

[44] Tim P German and Margaret Anne Defeyter. 2000. Immunity to functional fixedness in young children. *Psychonomic Bulletin & Review* 7, 4 (2000), 707–712. https://doi.org/10.3758/bf03213010

[45] Michael Golebiewski and danah boyd. 2019. Data voids: Where missing data can easily be exploited. *Data & Society* (2019).

[46] Natali Helberger. 2011. Diversity by design. *Journal of Information Policy* 1, 1 (2011), 441–469. https://doi.org/10.5325/jinfopoli.1.2011.0441

[47] Ingrid Hsieh-Yee. 1993. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science* 44, 3 (1993), 161–174.

[48] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM CIKM Conference on Information and Knowledge Management*. 77–86. https://doi.org/10.1145/1645953.1645966

[49] Bernard J. Jansen, Amanda Spink, Chris Blakely, and Sherry Koshman. 2007. Defining a session on Web search engines. *Journal of the Association for Information Science and Technology* 58, 6 (2007), 862–871. https://doi.org/10.1002/asi.20564

[50] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 607–616. https://doi.org/10.1145/2600428.2609633

[51] Jiepu Jiang and Chaoqun Ni. 2016. What affects word changes in query reformulation during a task-based search session?. In *Proceedings of the 2016 ACM CHIIR Conference on Human Information Interaction and Retrieval*. 111–120. https://doi.org/10.1145/2854946.2854978

[52] Jyun-Yu Jiang and Pu-Jen Cheng. 2016. Classifying user search intents for query auto-completion. In *Proceedings of the 2016 ACM ICTIR International Conference on the Theory of Information Retrieval*. 49–58. https://doi.org/10.1145/2970398.2970400

[53] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 445–454. https://doi.org/10.1145/2600428.2609614

[54] Matthew Jiwa, Patrick S Cooper, Trevor T-J Chong, and Stefan Bode. 2021. Choosing increases the value of non-instrumental information. *Scientific Reports* 11, 1 (2021), 1–11. https://doi.org/10.1038/s41598-021-88031-y

[55] Evangelos Kanoulas, Paul D. Clough, Ben Carterette, and Mark Sanderson. 2010. Overview of the TREC 2010 Session track. In *Proceedings of the 19th Text REtrieval Conference*. https://trec.nist.gov/pubs/trec19/papers/SESSION.OVERVIEW.2010.pdf

[56] Stavroula Karapapa and Maurizio Borghi. 2015. Search engine liability for autocomplete suggestions: Personality, privacy and the power of the algorithm. *International Journal of Law and Information Technology* 23, 3 (07 2015), 261–289. https://doi.org/10.1093/ijlit/eav009

[57] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Investigating users' query formulations for cognitive search intents. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 577–586. https://doi.org/10.1145/2600428.2609566

[58] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The meant, the said, and the understood: Conversational argument search and cognitive biases. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 5 pages. https://doi.org/10.1145/3469595.3469615

[59] Khamsum Kinley, Dian Tjondronegoro, Helen Partridge, and Sylvia Edwards. 2012. Human-computer interaction: The impact of users' cognitive styles on query reformulation behaviour during web searching. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 299–307. https://doi.org/10.1145/2414536.2414586

[60] Unni Krishnan, Alistair Moffat, and Justin Zobel. 2017. A taxonomy of query auto completion modes. In *Proceedings of the 22nd Australasian Document Computing Symposium*. Article 6, 8 pages. https://doi.org/10.1145/3166072.3166081

[61] Carol Collier Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371.

[62] Carol C Kuhlthau, Jannica Heinström, and Ross J Todd. 2008. The 'information search process' revisited: Is the model still useful. *Information Research* 13, 4 (2008), 13–4. http://InformationR.net/ir/13-4/paper355.htm

[63] Eng Pwey Lau and Dion Hoe-Lian Goh. 2006. In search of query patterns: A case study of a university OPAC. *Information Processing and Management* 42, 5 (2006), 1316–1329. https://doi.org/10.1016/j.ipm.2006.02.003

[64] Ard W Lazonder, Harm JA Biemans, and Iwan GJH Wopereis. 2000. Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society for Information Science* 51, 6 (2000), 576–581. https://doi.org/10.5555/358242.358267

[65] Binh Le, Damiano Spina, Falk Scholer, and Hui Chia. 2022. A crowdsourcing methodology to measure algorithmic bias in black-box systems: A case study with COVID-related searches. In *Proceedings of the Third Workshop on Bias and Social Aspects in Search and Recommendation (Bias @ ECIR 2022)*. 13 pages.

[66] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, et al. 2020. Indigenous Protocol and Artificial Intelligence position paper. (2020). https://doi.org/10.11573/spectrum.library.concordia.ca.00986506

[67] Binsheng Liu, Nick Craswell, Xiaolu Lu, Oren Kurland, and J. Shane Culpepper. 2019. A comparative analysis of human and automatic query variants. In *Proceedings of the 2019 ACM ICTIR International Conference on Theory of Information Retrieval*. 47–50. https://doi.org/10.1145/3341981.3344223

[68] Chang Liu, Jacek Gwizdka, and Nicholas J. Belkin. 2010. Analysis of query reformulation types on different search tasks. In *Proceedings of the 2010 iSchool iConference*. 477–485.

[69] Yang Liu, Ruihua Song, Yu Chen, Jian-Yun Nie, and Ji-Rong Wen. 2012. Adaptive query suggestion for difficult queries. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 15–24. https://doi.org/10.1145/2348283.2348289

[70] Donald M MacKay. 1969. *What Makes a Question?* 31–38.

[71] Gary Marchionini. 1997. *Information Seeking in Electronic Environments*. Number 9. Cambridge University Press.

[72] Gary Marchionini and Ryen White. 2007. Find what you need, understand what you find. *International Journal of Human Computer Interaction* 23, 3 (2007), 205–237. https://doi.org/10.1080/10447310701702352

[73] Dana McKay and George Buchanan. 2013. Boxing clever: How searchers use and adapt to a one-box library search. In *Augmentation, Application, Innovation, Collaboration, OzCHI '13*. 497–506. https://doi.org/10.1145/2541016.2541031

[74] Dana McKay, Stephann Makri, Shanton Chang, and George Buchanan. 2020. On birthing dancing stars: The need for bounded chaos in information interaction. In *Proceedings of the 2020 ACM CHIIR Conference on Human Information Interaction and Retrieval*. 292–302. https://doi.org/10.1145/3343413.3377983

[75] Dana McKay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the change that we seek: Information interactions during a change of viewpoint. In *Proceedings of the 2020 ACM CHIIR Conference on Human Information Interaction and Retrieval*. 173–182. https://doi.org/10.1145/3343413.3377975

[76] Lokman I. Meho and Helen R. Tibbo. 2003. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology* 54, 6 (2003), 570–587. https://doi.org/10.1002/asi.10244

[77] Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2021. Adaptive utterance rewriting for conversational search. *Information Processing and Management* 58, 6 (2021), 102682. https://doi.org/10.1016/j.ipm.2021.102682

[78] Alistair Moffat, Falk Scholer, Paul Thomas, and Peter Bailey. 2015. Pooled evaluation over query variations: Users are as diverse as systems. In *Proceedings of the 24th ACM CIKM International Conference on Information and Knowledge Management*. 1759–1762. https://doi.org/10.1145/2806416.2806606

[79] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. https://nyupress.org/9781479837243/algorithms-of-oppression

[80] Ragnar Nordlie. 1999. "User revealment" - A comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–18. https://doi.org/10.1145/312624.312618

[81] Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When are search completion suggestions problematic?. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4. Article 171. https://doi.org/10.1145/3415242

[82] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12, 3 (2007), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

[83] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*. 1–es. https://doi.org/10.1145/1146847.1146848

[84] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Proceedings of the 44th European Conference on Information Retrieval Research*. Springer, 397–412. https://doi.org/10.1007/978-3-030-99736-6_27

[85] Caroline Criado Perez. 2019. *Invisible Women: Data Bias in a World Designed for Men*. Abrams.

[86] Stephen Robertson. 2008. On the history of evaluation in IR. *Journal of Information Science* 34, 4 (2008), 439–456. https://doi.org/10.1177/0165551507086989

[87] Victoria L Rubin. 2019. Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation* 75, 5 (2019), 1013–1034. https://doi.org/10.1108/JD-12-2018-0209

[88] Procheta Sen, Debasis Ganguly, and Gareth J. F. Jones. 2021. I know what you need: Investigating document retrieval effectiveness with partial session contexts. *ACM Transactions on Information Systems* 40, 3 (2021). https://doi.org/10.1145/3488667

[89] Milad Shokouhi. 2013. Learning to personalize query auto-completion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 103–112. https://doi.org/10.1145/2484028.2484076

[90] Milad Shokouhi and Kira Radinsky. 2012. Time-sensitive query auto-completion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 601–610. https://doi.org/10.1145/2348283.2348364

[91] Marc Sloan, Hui Yang, and Jun Wang. 2015. A term-based methodology for query reformulation understanding. *Information Retrieval Journal* 18, 2 (2015), 145–165. https://doi.org/10.1007/s10791-015-9251-5

[92] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology* 52, 3 (2001), 226–234.

[93] Ben Steichen and Ryan Lowe. 2021. How do multilingual users search? An investigation of query and result list language choices. *Journal of the Association for Information Science and Technology* 72, 6 (2021), 759–776. https://doi.org/10.1002/asi.24443

[94] Hanna Stelmaszewska and Ann Blandford. 2004. From physical to digital: A case study of computer scientists' behaviour in physical libraries. *International Journal on Digital Libraries* 4, 2 (2004), 82–92. https://doi.org/10.1007/s00799-003-0072-6

[95] Raphael Tang, Karun Kumar, Kendra Chalkley, Ji Xin, Liming Zhang, Wenyan Li, Gefei Yang, Yajie Mao, Junho Shin, Geoffrey Craig Murray, and Jimmy Lin. 2021. Voice query auto completion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic, 900–906. https://doi.org/10.18653/v1/2021.emnlp-main.68

[96] Robert S Taylor. 1968. Question-negotiation and information seeking in libraries. *College & Research Libraries* 29, 3 (1968), 178–194.

[97] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS 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*. 151–158. https://doi.org/10.1145/1277741.1277770

[98] Francesca Tripodi. 2018. *Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices.* Technical Report. Data & Society Research Institute. https://datasociety.net/research/media-manipulation/

[99] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management* 39, 3 (2003), 445–463. https://doi.org/10.1016/S0306-4573(02)00031-6

[100] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A comparison of question rewriting methods for conversational passage retrieval. In *Proceedings of the European Conference on Information Retrieval*. 418–424. https://doi.org/10.1007/978-3-030-72240-1_43

[101] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Vol. 63. MIT Press.

[102] Tung Vuong, Salvatore Andolina, Giulio Jacucci, and Tuukka Ruotsalo. 2021. Spoken conversational context improves query auto-completion in web search. *ACM Transactions on Information Systems* 39, 3, Article 31 (May 2021). https://doi.org/10.1145/3447875

[103] Nina Wacholder. 2011. Interactive query formulation. *Annual Review of Information Science and Technology* 45, 1 (2011), 157–196. https://doi.org/10.1002/aris.2011.1440450111

[104] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* 28, 4 (2010), 20:1–20:38. https://doi.org/10.1145/1852102.1852106

[105] Ryen W White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3–12. https://doi.org/10.1145/2484028.2484053

[106] Ryen W White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web* 8, 4 (2014), 1–33. https://doi.org/10.1145/2663355

[107] Ryen W White and Dan Morris. 2007. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 255–262. https://doi.org/10.1145/1277741.1277787

[108] Ryen W White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*. 135–136. https://doi.org/10.1145/2740908.2742769

[109] Thomas D. Wilson. 1997. Information behaviour: An interdisciplinary perspective. *Information Processing and Management* 33, 4 (1997), 551–572. https://doi.org/10.1016/S0306-4573(97)00028-9

[110] Jiun-Yu Wu and Chen Xie. 2018. Using time pressure and note-taking to prevent digital distraction behavior and enhance online search performance: Perspectives from the load theory of attention and cognitive control. *Computers in Human Behavior* 88 (2018), 244–254. https://doi.org/10.1016/j.chb.2018.07.008

[111] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality* 10, 4, Article 16 (Oct 2018). https://doi.org/10.1145/3239571

[112] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting civil discourse through search engine diversity. *Social Science Computer Review* 32, 2 (2014), 145–154. https://doi.org/10.1177/0894439313506838

[113] George D Yonge. 1966. Structure of experience and functional fixedness. *Journal of Educational Psychology* 57, 2 (1966), 115. https://doi.org/10.1037/h0022967

[114] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*. 418–428. https://doi.org/10.1145/3366423.3380126

[115] Guido Zuccon, João R. M. Palotti, and Allan Hanbury. 2016. Query variations and their effect on comparing information retrieval systems. In *Proceedings of the 25th ACM CIKM International Conference on Information and Knowledge Management*. 691–700. https://doi.org/10.1145/2983323.2983723