Query Variability and Experimental Consistency: A Concerning Case Study

Lida Rashidi
The University of Melbourne
and RMIT University
Melbourne, Australia
rashidi.l@unimelb.edu.au

Justin Zobel
The University of Melbourne
Melbourne, Australia
jzobel@unimelb.edu.au

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

ABSTRACT

In offline experimentation, the effectiveness of a search engine is evaluated using a document collection, a set of queries against that collection, a set of relevance judgments connecting the documents and the queries, and an effectiveness metric. This measurement pipeline is used as a surrogate for user satisfaction – the extent to which the system provides useful information to the users that are issuing the queries. But queries are responses to information needs, or topics, and there can be a wide variety of ways in which any given information need can be expressed as a query. That one-to-many relationship suggests that, in an IR experiment, use of any single query to represent a topic may be insufficient. In this case study, we demonstrate that this practice is indeed a weakness, by showing that the TREC 2013 and 2014 Web track queries, which are regarded as being indicative of specific information needs, are not necessarily representative of crowd-generated queries for the same underlying needs, and can give rise to inconsistent system relativities when compared to user-generated queries. From this instance we must thus note an element of concern: that current test collection design strategies can lead to effectiveness results that are at odds with those experienced by typical non-expert users.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Test collections**; **Relevance assessment**; **Retrieval effectiveness**.

KEYWORDS

Evaluation; significance testing

ACM Reference Format:
Lida Rashidi, Justin Zobel, and Alistair Moffat. 2024. Query Variability and Experimental Consistency: A Concerning Case Study. In *Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24), July 13, 2024, Washington DC, DC, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3664190.3672519

1 INTRODUCTION AND BACKGROUND

In offline IR evaluation, the quality of the results retrieved by a search engine is assessed using a collection of documents, queries, their corresponding relevance judgments, and an effectiveness measure; see Sanderson [21] for more information. While all aspects of this framework are necessary, the queries play an especially important role. If they are to be a realistic test they should represent the way in which one or more users would formulate the retrieval task in response to a given information need. Since this is a human process, for any given information need there is no single query that can be regarded as being definitive; for this reason, there has been a range of work that has examined user query variability [2–4, 7, 18, 22, 27], including the effect that query variations have on pooling costs [17] and ways in which query variations might be exploited to boost retrieval performance [5]. While a range of techniques such as query suggestion, substitution, and expansion are used to refine and, to a certain extent, homogenize the initial query in order to produce better retrieval results [8, 15], there is nevertheless the potential for considerable variability in the effectiveness experienced by different users.

The significance and impact of query variations is at odds with the design of many of the ⟨documents, queries, judgments⟩ corpora that have been widely used by the IR community. For example, many of the experimental comparisons carried out using the TREC Ad Hoc and Web Tracks over the last three decades of measurement have relied on the use of a single query associated with each information need, making the implicit assumption that this single query is canonical. The relevance judgments have been guided by a narrative that provides additional and nuanced information about what is required in order for a document to be an answer, but the TREC-supplied queries – normally one per topic, and sometimes denoted as being the topic "title" – are often the only ones used to compare system effectiveness. Furthermore, not only is the use of just a single query of concern, but in some corpora the query used is created as a topic descriptor – a quite different function to, say, choosing a phrase believed to be a "typical" query for that topic.

It is these various concerns that prompted the investigation reported in this case study. First we consider a very simple question: *if one user perceives some retrieval system (System A, say) to be better than another (System B) over a set of queries, will a different user with the same set of information needs, but expressed via different queries, also perceive System A as being better*? That is, we examine the extent to which system comparisons are agnostic to the query formulations used to measure them.

Reassuringly, using the TREC-supplied queries for the 2013 and 2014 Web track (for which a large pool of query variants is available from a different study) we find that the systems that perform well using one query variant for each topic are also likely to perform well if presented with a different set of user-generated queries. That

ICTIR '24, July 13, 2024, Washington DC, DC, USA

ACM ISBN 979-8-4007-0681-3/24/07.
https://doi.org/10.1145/3664190.3672519

is, systems that appear to be "good" to one user do indeed tend to also appear "good" to other users.

We then consider a related question: *is the same consistency observed if one of the "users" is in fact the set of TREC-supplied queries?* The answer is concerning, as our results show that the TREC-provided queries are exceptions to the general pattern. Relationships between systems that are established via the TREC-provided queries are often *not* supported by user-generated queries that address the same information need. That is, this second experiment demonstrates that the TREC queries, upon which so much reliance is placed in terms of IR experimental methodology, may *not* be reliable predictors of the system–versus–system relationships that might be experienced by users formulating their own queries.

To summarize: there has been an implicit assumption in the use of many of the TREC test corpora that the TREC-provided query for each topic is canonical, and thus is an invariant. Using the one corpus for which suitable user-generated queries and relevance judgments are available, we describe an experiment that tests this widely-held evaluation assumption and show that the TREC-provided queries have different statistical properties to worker-generated queries. Our case study is limited to a single corpus. Nevertheless it provides a counter-example to that implicit assumption, and hence raises clear concerns that require careful consideration.

The notion of query variability and its impact on system performance has also been studied by Culpepper et al. [12], who demonstrate that relative system performance can vary depending on the choice of query, and hence that the inclusion of query variations can affect the influence that "topic difficulty" has in IR evaluation. Query performance prediction (QPP) using query variations has also been explored [24], with the relative performance of the QPP methods found to be impacted by the effectiveness of the queries used to represent the underlying need, and with ensemble methods across multiple variations providing better predictions.

2 EXPERIMENTS

We now describe the experiments that were carried out and present their outcomes. To address our questions we require a standard test collection spanning documents, topics, and judgments. In addition, we also require a set of query variations for each topic; and we need to be confident that relevance judgments provide sufficiently comprehensive coverage of those variants that different systems can be fairly compared. The only corpora that meet that full set of requirements are the TREC 2013 and 2014 Web tracks, described shortly. We also require a selection of retrieval systems, so that it is possible to determine whether relative performance is preserved when different query variants are used.

Queries and Systems. As noted, we make use of the TREC 2013 and 2014 Web track resources. The queries used in these tracks were from commercial search engines, and selected as being representative of typical search tasks. For each of TREC 2013 and 2014, a set of 50 such queries was collated, taking a mix of broad and specific information needs. To prevent ambiguity, we refer to these as being the *seed queries* associated with the corpus for these two tracks. Some of the seed queries are structured in a multiple-subtopic paradigm, while others are focused on a single subtopic [10, 11]. Relevance judgments relative to the seed queries and their possible subtopics were created by NIST assessors using a six-point scale [10, 11], but we did not use those judgments in these experiments.

Bailey et al. [2] subsequently created query variants, known as the UQV100 queries, together with relevance judgments derived from document pools formed by executing each of those query variants using five different systems. That process commenced with a set of 100 information need statements (referred to as *backstories*) developed from the 100 TREC-supplied seed queries for the 2013 and 2014 Web tracks, selecting (where there were multiple options available) one identified subtopic. The backstories can thus be thought of as inferred topic statements, akin to the topic descriptions that were used in earlier TREC rounds. For consistency with previous work, we continue to refer to each of these inferred information needs as a topic.

Those backstories were then presented to crowd-workers, who were asked what query they would use in response to each. That process led to a total of 10,835 query variants being collected, averaging (after data validation and filtering) 108 queries per topic and 57.6 distinct query variants per topic [2]. High levels of diversity in user-generated queries in response to backstories has been a key finding of work in this area [2, 18].

We next checked the 100 sets of user-generated UQV100 queries to see if the corresponding TREC-provided seed queries had been suggested as a query by the crowd workers. There were 77 queries for which that had happened. Because we were interested in comparing the seed queries and user-generated queries, we selected that subset of 77 for use in our experiments. That is, each of the 77 topics employed in the experiments described shortly has a user-generated query set that includes the corresponding TREC seed query, and hence has relevance judgments for which the TREC seed query was a "first class contributor". That filtering process led to a total of 4,218 distinct queries, or 54.8 per topic.

Those queries were then passed to a suite of fifteen different retrieval systems, and retrieval runs prepared. The fifteen systems used three different ranking models: BM25, a probabilistic retrieval model based on bag-of-words [20]; QLD, query likelihood with a Dirichlet smoothed language model [25]; and SPL, an information theoretic model [9]. Other variants of these three core systems were created via an RM3 query expansion model [1] and via axiomatic reranking algorithms [13], using different parameter settings. We used the Anserini toolkit [23] to formulate these search engines with their different extensions and parameter settings.[1]

We make use of several effectiveness metrics and hence a range of different corresponding user models, to cover the spectrum of possible evaluation scenarios. The five metrics employed are average precision (AP), a top-weighted recall-sensitive mechanism [6]; normalized discounted cumulative gain (NDCG) [14], which is also recall-sensitive but less heavily top-weighted; precision at depth 10 (Prec@10), modeling users who always look at exactly ten results; rank-biased precision (RBP) with a parameter $\phi = 0.85$, simulating users who on average view the top approximately seven answers in each results listing [16] but may also look at fewer or more; and reciprocal rank (RR), modeling users that search until they find a first relevant document. The last three have associated

[1] https://github.com/castorini/anserini

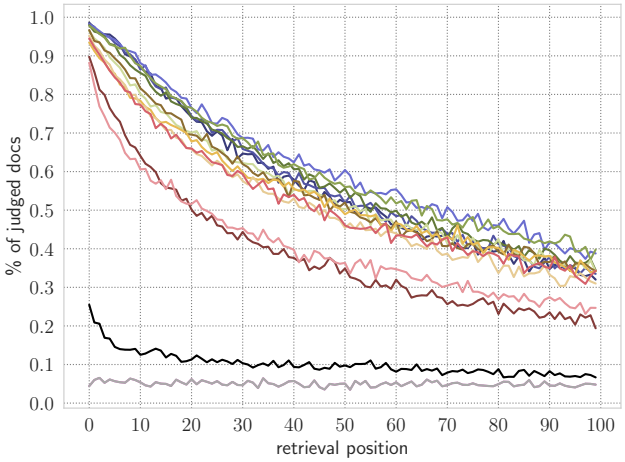


Figure 1: Fraction of judged documents at each position in runs, averaged over 4,218 queries associated with 77 topics. Each line represents one of the fifteen systems constructed.

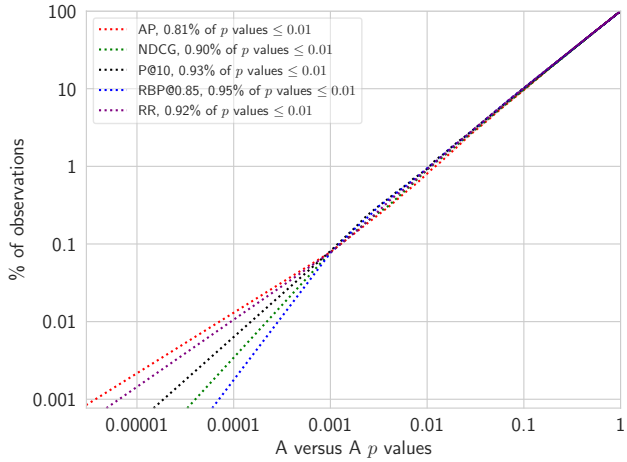


Figure 2: Comparing a system with itself (Experiment 1). A p value is generated by comparing the 77 per-topic scores observed by user α and user β when they both use the same system to process different queries that address the same original information need. The annotations in the legend show that for all five effectiveness metrics the number of false positives is below the computed p value when summed for all user pairs for which $p \le 0.01$.

user models that do not require the user to be aware of the total volume of possible answer documents in the collection [16].

Experiment 0: Validation of Systems and Judgments. The relevance judgments for the UQV100 collection were generated by pooling the runs for each topic (over an average of 57.6 distinct queries per topic), with all runs generated at first by an Indri/BM25 system [2], and then later augmented by runs from four other retrieval systems.[2] Multiple rounds of judging were undertaken: first, uniform pooling was applied to depth 10 on each run of the original Indri/BM25 system, yielding 21,895 judgments (covering all 100 topics); then a further 5,501 documents were judged based on a "weighted coverage" basis [2]; and finally further judgments were undertaken when the query runs from another four research retrieval systems were included. When restricted to the 77 topics selected for our experiments, this total set of 55,587 judgments was reduced to 39,478 judgments (71.0% of the original set), of which 8,140 (20.6%) were relevant at grade one or above.

To validate the judgments, and to verify that they were suited to the fifteen systems we employed, we first computed, for each system, the fraction of judged documents that were surfaced by the query variations at each position in the run. Figure 1 shows the results. Thirteen of the systems exhibit what we would regard as "normal" behavior, with high average fractions of judged documents at early positions in the corresponding runs, tapering downward as the depth in the run increases.

But two of the systems were notably anomalous. Those two systems surfaced quite different document sets, with very low fractions judged, and hence with correspondingly low accuracy in any computed effectiveness scores. As a result of this step we removed those two systems and proceeded to our main experimentation using the remaining thirteen systems, satisfied that the UQV100 relevance judgments provide a reasonable fit. Note that interrogation of experimental outcomes in this way is a critical step prior to making inferences about results. There is no basis for assuming in experiments of this form that unjudged documents are irrelevant, and the presence of systems where there is high uncertainty in measured results would be a confound. On the other hand, removal of any particular system from the results does not introduce bias.

Methodology. To simulate a *pair of random users* we sample the set of 4,218 query variations on a per topic basis, selecting two different variants for each topic, assigning one to a user denoted α and the other to a user denoted β. This gives us a sequence of 77 queries for user α, and a set of 77 disjoint queries for user β, but with both α and β able to be regarded as seeking answers to the same set of 77 information needs. That is, user α and user β can be considered as a paired experiment for statistical testing purposes, to determine if α receives better quality responses from a system than does user β.

The sampling and statistical testing process can then be repeated many times, to develop an overall pattern of behavior, in a manner akin to the bootstrap test. In the experiments reported next a set of 10,000 repetitions were carried out, each involving a user α searching using a set of 77 queries, paired with a user β searching for the same information but via a different set of 77 queries.

Experiment 1: A versus A. In this experiment we suppose that user α and user β both make use of the same retrieval system to process their queries. We employ each of the 13 retrieval systems, and 10,000 drawings to form pairs of users; that is, we in effect carry out 130,000 experiments in which two different users are assumed to use the same system to search for the same information need.

Each of those trials generates paired vectors of 77 metric scores for each of the five effectiveness metrics, and hence can be further processed by a statistical test to determine a set of five p values.

[2]See http://dx.doi.org/10.4225/49/5726E597B8376, file README.txt.

Having always compared a system against itself, we expect to see very few findings of "there is a significant difference between the experience of user α and the experience of user β" (when "experience" is assessed via the corresponding metric score) that is, very few small p values. Figure 2 shows the results (with p values computed using a Student t-test) and exhibits the expected pattern of results. Each of the five lines represents cumulative fractions for one of the effectiveness metrics, and while they differ a little at the left-hand low-frequency section of the plot, all five are closely aligned through the region of primary interest between $p \approx 0.001$ and $p \approx 0.05$. For example, regardless of metric, approximately 1% of the experiments resulted in a p value less than 0.01.

This outcome validates the methodology we have developed, and demonstrates that the five metrics all display the expected behavior.

Experiment 2: A versus B. A frequent goal in offline IR evaluation is to compare two systems, evaluating an incumbent regarded as being a *champion* against a newer *challenger*. With two systems in play, referred to here as being "A" and "B", and two users α and β with the same set of information needs but different queries, as already described, we can ask the contingent question "*if user α observes better performance from System A than they do from System B, will user β observe the same relationship*?"

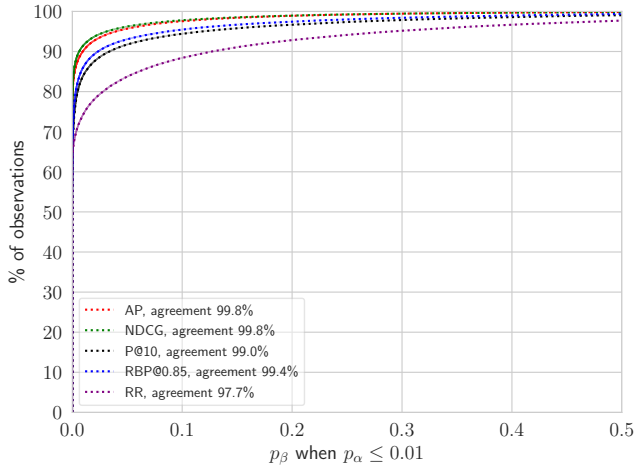
In the first part of the exploration we restrict our attention to system pairs A and B for which user α detects a strongly statistically significant outcome over their set of 77 queries in favor of System A, filtering the set of all possible (A, B, α, β) tuples to the subset for which user α calculates a Student t-test p_α value that is ≤ 0.01. Applying that filtering process on a per-metric basis reduced the 780,000 (A, B, α, β) (starting with 10,000 query sets α, and with 78 system pairs possible from 13 systems) combinations to between 191,115 (for RR) and 482,367 (for NDCG) combinations.

Working only with those filtered subset of system pairs, we then ask what p_β value is observed by user β when comparing each pair of systems. The results are depicted in Figure 3(a), with a cumulative distribution of p_β plotted for each of the five effectiveness metrics. Now the values in the legend record the fraction of p_β values less than 0.5. That is, those annotated "agreement rates" reflect the total fraction of tuples for which user β also observes System A to be superior to System B.

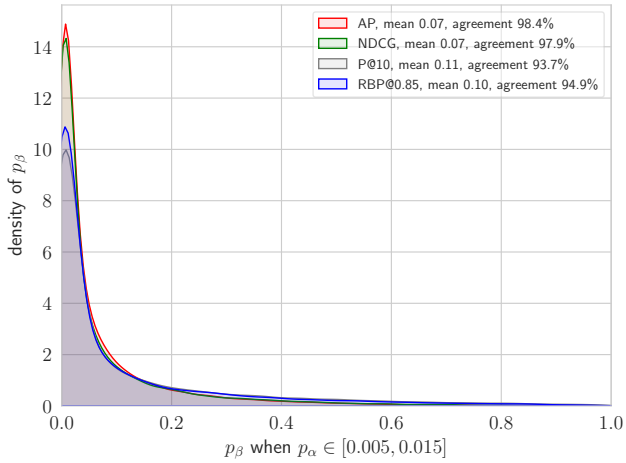
With $p_\alpha \leq 0.01$, we expect that final agreement fraction to be $\geq 99\%$, and that is indeed what occurs for four of the five metrics, as noted in the graph's legend box. On the other hand, RR only achieves a 97.7% "predictive score" in this experiment, an outcome that is not surprising given that RR is not suited to the Student t test. (That is, because the paired score differences that RR generates are unlikely to be normally distributed; thereby supplying a timely reminder in regard to statistical tests only being valid if their preconditions are satisfied).

As a further observation, note that the upper pane in Figure 3 also shows that more than 80% of the p_β values are not just less than 0.5, but are also smaller than 0.01, meaning that 80% of the time user β would also find that System A was significantly better than System B at the $p_\beta \leq 0.01$ level.

In the second part of the experiment we restrict (A, B, α, β) in a slightly different way, taking instead those combinations that yield



(a) Cumulative distributions for p_β, given that $p_\alpha \leq 0.01$



(b) Density of p_β, given that $0.005 \leq p_\alpha \leq 0.015$

Figure 3: Comparing systems (Experiment 2), A versus B outcomes over 10,000 trials each consisting of 77 randomly selected queries. In plot (b) when $p_\beta < 0.5$, System A is superior to System B; when $p_\beta > 0.5$, System B is superior to System A, with $1 - p_\beta$ plotted instead, thereby forming a single continuous scale.

$0.005 \leq p_\alpha \leq 0.015$ for each metric, that is, a band of broadly comparable significance outcomes centered around 0.01. This alternative filtering process reduces the 780,000 (A, B, α, β) combinations to between 39,263 (for NDCG) and 43,984 (for AP) combinations. The results are plotted for four of the metrics in Figure 3(b) as an inferred density distribution associated with the corresponding p_β values, calculated using the Kernel Density Estimation function in the Python seaborn package. In this plot we have used a "mirrored" horizontal scale in which the value 0.01 on the horizontal axis indicates that user β finds System A to be significantly better than System B; similarly, the point 0.99 indicates that user β observes the *reverse*, that System A is significantly *worse* than System B at the 0.01 level. That is, in this graph (and also in Figure 4(b)) we have created a single "blended p value" scale that spans the range

from 0 to 1, and similarly spans the spectrum from System A being better through to System B being better.

We again see the expected behavior – for each of the metrics the p_β density peaks in the vicinity of 0.01, indicating that users α and β observe broadly similar outcomes when comparing System A and System B, even though they distilled the information need into different queries. Moreover, while the p_β values observed by user β have means in the range 0.07 to 0.11 (noted in the legend box), and are almost ten times larger than the corresponding $p_\alpha \approx 0.01$ values, this is not of concern. Statistical significance on the part of user α does not imply that user β should see the same level of significance, only that β is likely to observe that System A is superior when the A's mean is compared to B's mean. The levels of agreement also drop in this part of Experiment 2 – in the case of Prec@10 and RBP quite notably so – but this is also not problematic in any way. In removing the tuples (A, B, α, β) in which $p_\alpha < 0.005$ the expectations in regard to the fraction of times agreement should be observed by user β are weakened, and weakened by, as it turns out, different amounts across the suite of effectiveness metrics.

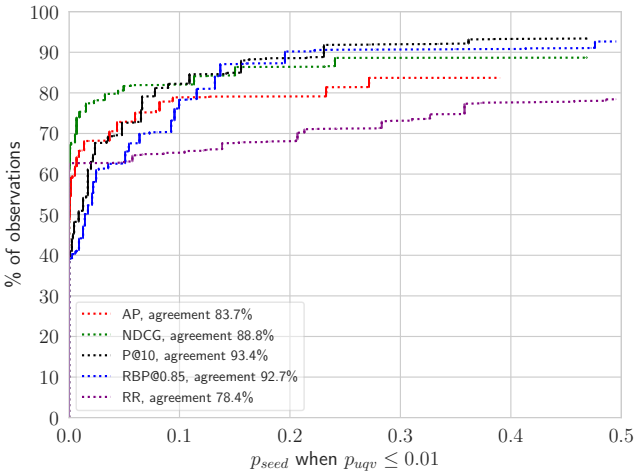
The results in Figure 3 answer the main question posed in Section 1: *if one user perceives some retrieval system (System A, say) to be better than another (System B) over a set of queries, a different user with the same set of information needs, but expressed via different queries, is also likely to perceive System A as being better.*

Experiment 3: Query Variations versus Seed Queries. Our next experiment compares the TREC seed queries against the other query variations in the UQV100 test collection. In this experiment the 77 queries for user α are a random sample from the UQV100 query set, including the TREC seed query for each topic, and we apply the same two filtering options on tuples (A, B, α, β): inclusion when $p_\alpha \le 0.01$; and inclusion when $0.005 \le p_\alpha \le 0.015$. What makes this experiment different is that user β is assumed to *always* employ the TREC-provided seed query, recalling that the 77 topics were selected because the seed query was amongst the options proposed by the crowd workers.

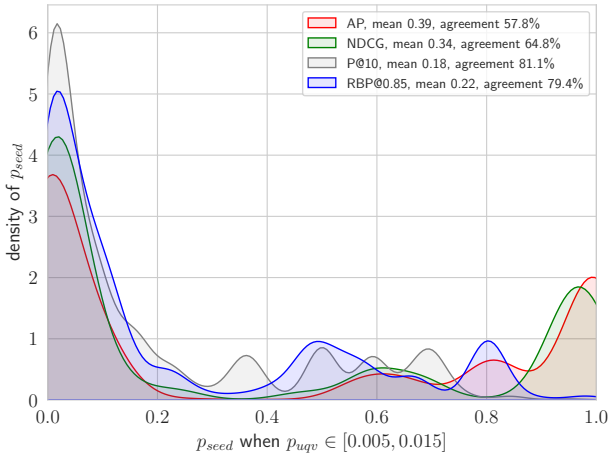
Figure 4 presents the same two views as in Figure 3, but with clear differences visible. In the upper pane in Figure 4, the agreement levels are lower than those shown in Figure 3(a). For example, even putting the RR curve to one side because of the mismatch between it and the Student t-test, if user α employs their metric of choice and detects a difference between System A and System B at the 0.01 level, that "A is better than B" relationship would not be observed by user β between around 5% and around 15% of the time.

The lower pane in Figure 4 confirms this lack of predictivity, with the plot again using the "mirrored" horizontal scale of Figure 3(b) in which values of $p_\beta > 0.5$ indicate that user β measured System A as being *inferior* to System B. The regions of moderate density for AP and NDCG at the right hand end of this lower plot are what are most startling. They indicate that it is not at all uncommon for user α to claim that System A is significantly better than System B, but for user β, who has used the TREC seed queries, to simultaneously believe – and just as strongly – that they have assembled evidence that System B is better than System A.

Experiment 4: Patterns of Agreement. Figure 5 consolidates Figures 3(b) and 4(b) into a single presentation. To make the blue box-whisker elements in this graph, each of the ≈ 40,000 (A, B, α)



(a) Cumulative distribution for p_β, given that $p_\alpha \le 0.01$.



(b) Density of p_β , given that $0.005 \le p_\alpha \le 0.015$

Figure 4: Seed versus UQV queries (Experiment 3). User α employs a randomly selected worker-generated query for each topic; user β always uses the TREC-supplied seed query for each topic. Other details are as for Figure 3.

combinations (of 780,000 as a starting point) for which $0.005 \le p_\alpha \le 0.015$ was regarded as a "reference user", and each one of those was compared to a further set of 10,000 β selections. Each point plotted in each box-whisker element is then the fraction of those 10,000 trials for which the set of β users observed the same ordering relationship between Systems A and B as did user α. There are thus ≈ 40,000 such points (the exact number varying according to the metric, as noted above) plotted in each of the five box-whisker elements. That is, the box-whisker elements show the distribution of the agreement values that are condensed into the overall averages given in the legend box of Figure 3(b). The distributions are reasonably tightly centered on the median values, shown as solid lines in the boxes, but in each case there is also a long descending tail of outliers. Each outlier represents a reference user who observed – with strong statistical significance – a System A

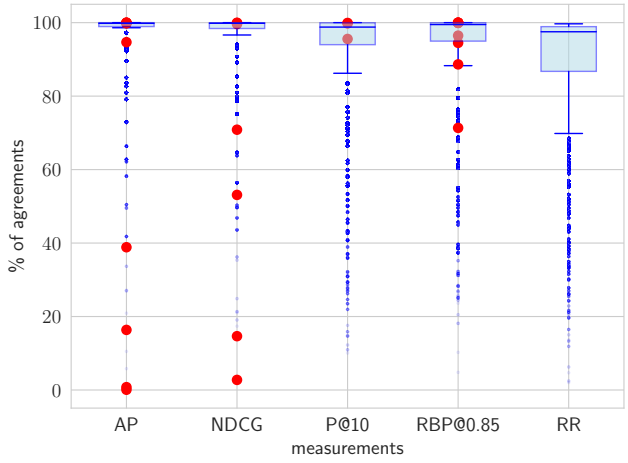


Figure 5: Distribution of agreement rates between sets of random user queries taken as α users (all blue dots, and the blue box-whisker elements) and 10,000 other query sets; and between TREC seed queries (red dots) as the α set and the same set of 10,000 β query sets. In all cases the α set led to significance between System A and System B with $0.005 \le p_\alpha \le 0.015$.

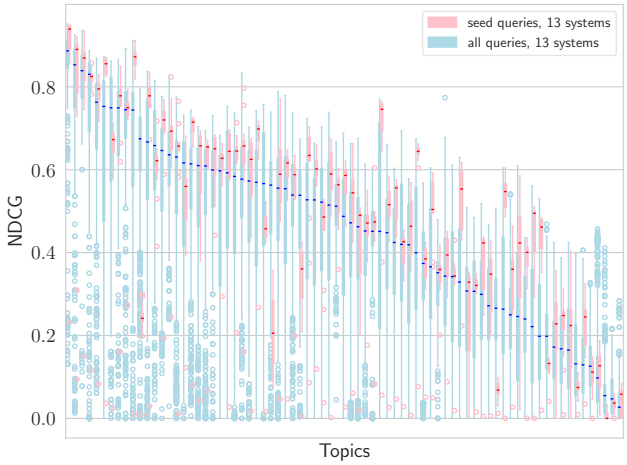


Figure 6: Absolute NDCG score distributions using all query variations per topic and 13 systems (blue), for each of 77 topics. The second distribution (pink) for each topic reflects the use of the seed queries across the 13 systems. Each pair of adjacent blue and pink bars represents one of the topics, ordered by the median (across systems) blue point.

versus system B outcome that varied markedly from the aggregate view of 10,000 other users who queried using other variations.

The red dots in Figure 5 then reflect the same measurement, but with α always the set of seed queries, and with A and B a system pair that are differentiated by the seed queries with $0.005 \le p_\alpha \le 0.015$. For example, of 78 possible system pairs, 7 are retained when the metric is NDCG and hence lead to 7 red dots, some of which are overlaid in the figure (recall that we are examining only a relatively narrow band of p_α values in this experiment). There were no system pairs fitting this criteria for RR. An agreement rate is again computed from 10,000 randomly generated β query sets, but now we are asking, "*if two systems are significantly different according to the TREC seed queries, what fraction of user-generated query sets will obtain the same system relativity?*" The difference between the random pairings and the seed query pairings is now stark. The TREC seed queries, even though they do sometimes arise from the crowd worker elicitation process (in particular, in all of the 77 topics used in these experiments), form a query suite that when taken as a collective whole is distinctively different from the worker-supplied queries.

Experiment 5: Absolute Performance of Seed Queries. It is also interesting to consider whether the seed queries give rise to better retrieval effectiveness than do the user-generated queries. Figure 6 plots NDCG scores, and shows that in general they do. The spread of per-topic NDCG scores across the 13 systems and total (across topics) of 4,218 queries in the blue bars is very diffuse, and some of the user query variations lead to very poor performance. Only a relatively small fraction of that broad query set out-perform the seed queries (adjacent pink bars). But nor are the seed queries the best for most of the topics. We note that Culpepper et al. [12] have sought to disentangle system effects from query variation effects; whereas in this graph systems and query variants are conflated, with one blue box-whisker element per topic.

3 CONCLUSION AND FUTURE WORK

We have used the UQV100 data to explore whether query variations support consistent evaluations of relative system performance; as well as considering the more narrow question of whether the TREC-supplied seed queries are representative of worker-generated queries for the same topics. Central to these experiments is the knowledge that the seed queries – taken to be invariant in many IR experiments – are but a single way in which the underlying information needs might be represented. The UQV queries pertain to the TREC 2013 and 2014 Web track, and thus provide a case study in which the representativeness of seed queries can be tested.

Focusing on a subset of the UQV100 topics for which the seed query was also provided by one or more crowd workers, and using a total of thirteen different retrieval systems, we have measured the predictivity of statistical tests when assessing experiments making use of one query version per topic. When query variations are compared against each other via random sampling, all is well – if a paired statistical test reports for one user that two retrieval systems are different, a second user making their own selection from the queries is likely to identify the same relationship.

But when one of those two users always employs the corresponding TREC seed query, and the other issues a worker-generated query other than the seed query, the situation is more complex. Taken as a specific subset, the seed queries often give rise to different outcomes to the UQV100 query set, and a user who bases their evaluation on only the TREC-provided seed queries will, more often than can be accounted for by random fluctuations, observe reversed system relativities relative to a user who samples from the pool of query variations. Indeed, a TREC 2013 and 2014 "seed-only"

querier may well find that they have evidence in support of System *A* being statistically superior to System *B*, while at the same time a non-seed querier might have equally compelling evidence in favor of System *B*. That is, use of the TREC 2013 and 2014 seed queries alone in a challenger versus champion experiment could lead to an outcome that is at odds with the relativity observed by typical users of those same two systems.

We further note that these results are based only on the 77 topics for which one or more of the UQV crowd workers generated the TREC seed query in response to the topic's backstory. The comparison might be even more divergent for the other 23 topics, for which the crowd workers have implicitly indicated that the seed query is *not* a popular choice.

One possible cause of the divergence noted in Experiment 3 arises from the way the backstories were constructed. It is possible that topic drift was introduced when the UQV100 backstories were authored from the TREC seed queries, and hence that the crowd-supplied queries were responses to subtlety different information needs. To try and isolate this possible effect and measure whether it had occurred would be complex, requiring that the initiator of the seed query had also concurrently captured their search intention via a information need commentary. Such analysis was not available to us. Another possible limitation of our experiments is the time that elapsed between the collection of the seed queries and the corresponding corpora, and when the UQV queries and judgments were solicited. Again, we can speculate that it may have affected the integrity of our experiments, but are unable to control for or eliminate it as a factor.

Even with those possible limitations, our findings provide further support to the encouragement already articulated by Bailey et al. [2, 3] for researchers and practitioners alike to make use of multiple queries per topic; and to ensure that the judgments being used in experiments are equitably suited to the evaluation of all of the experimental systems. While we again acknowledge that we have explored only a single collection, this one test already provides a counter-example to the assumption that systems can be reliably evaluated on TREC-specified seed queries alone. If and when other collections have similar query variations and suitable relevance judgments created, it will become possible to consider the prevalence of the problem we have documented here. But one such instance – the case study of this paper – is sufficient to establish that the problem can and does arise.

In terms of future work, we note that what we have done here can be thought of as "query bootstrapping", while holding the set of topics constant. From that point of view it complements the previous work by Zobel and Rashidi [26] that explores "corpus bootstrapping", and by Rashidi et al. [19] that explores "judgment bootstrapping". That observation then opens the possibility of combined bootstrapping modes, in which more that one of these elements is varied, so as to further investigate experimental predictivity.

Acknowledgment. We thank the referees for their constructive comments. This work was supported under the Australian Research Council's Discovery Projects funding scheme (project DP190101113).

REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMass at TREC 2004: Novelty and HARD. In *Proc. TREC*, 2004. NIST Special Publication 500-261.

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

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

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

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

[6] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–78. MIT Press, 2005.

[7] C. Buckley and J. Walz. The TREC-8 query track. In *Proc. TREC*, 1999. NIST Special Publication 500-246.

[8] J. Chen, Y. Liu, J. Mao, F. Zhang, T. Sakai, W. Ma, M. Zhang, and S. Ma. Incorporating query reformulating behavior into web search evaluation. In *Proc. CIKM*, pages 171–180, 2021.

[9] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In *Proc. SIGIR*, pages 234–241, 2010.

[10] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. TREC 2013 web track overview. In *Proc. TREC*, 2013. NIST Special Publication 500-302.

[11] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2014 web track overview. In *Proc. TREC*, NIST Special Publication 500-308, 2014.

[12] J. S. Culpepper, G. Faggioli, N. Ferro, and O. Kurland. Topic difficulty: Collection and query formulation effects. *ACM Trans. Inf. Sys.*, 40(Article 19):1–36, 2021.

[13] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *Proc. SIGIR*, pages 480–487, 2005.

[14] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

[15] M. P. Kato, T. Sakai, and K. Tanaka. When do people use query suggestion? A query suggestion log analysis. *Inf. Retr.*, 16:725–746, 2013.

[16] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2, 2008.

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

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

[19] L. Rashidi, J. Zobel, and A. Moffat. The impact of judgment variability on the consistency of offline effectiveness measures. *ACM Trans. Inf. Sys.*, 42(1):19:1–19:31, 2023.

[20] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC*, pages 109–126, 1995. NIST Special Publication 500-225.

[21] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. Trnd. Inf. Retr.*, 4(4):247–375, 2010.

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

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

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

[25] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *SIGIR Forum*, 51(2):268–276, 2017.

[26] J. Zobel and L. Rashidi. Corpus bootstrapping for assessment of the properties of effectiveness measures. In *Proc. CIKM*, pages 1933–1952, 2020.

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