Judgment Pool Effects Caused by Query Variations

Alistair Moffat
Department of Computing and Information Systems
The University of Melbourne, Australia
ammoffat@unimelb.edu.au

ABSTRACT

Batch-mode retrieval evaluation relies on suitable relevance judgments being available. Here we explore the implications on pool size of adopting a "query variations" approach to collection construction. Using the resources provided as part of the UQV100 collection [Bailey et al., SIGIR 2016] and a total of five different systems, we show that pool size is as much affected by the number of query variations involved as it is by the number of contributing systems, and that systems and users are independent effects. That is, if both system and query variation are to be accommodated in retrieval experimentation, the cost of performing the required judgments compounds.

1. INTRODUCTION

Batch evaluation provides one way in which IR systems can be compared. A set of topics is provided, and matching relevance judgments for some subset of documents, based on human annotation. The systems to be measured execute queries reflecting the topics, and then the resultant rankings are converted to quantitative scores using one or more effectiveness metrics, employing the relevance judgments [3, 9]. *Pooling* is one method for determining which documents should be labeled, with the pool formed by including all documents down to some depth d in the rankings generated by a set of s contributing systems, see, as one example, Kuriyama et al. [5]. This usual form of pooling provides coverage of system variation.

Following Buckley and Walz [4], Bailey et al. [1] suggest that *query-based variation* should also be a component of batch evaluations. Bailey et al. provided crowd-workers with information-need statements (*backstories*) for 180 topics, and asked each worker what their first query would be. The queries gathered for each topic were surprisingly varied, and as a result, the rankings generated for them were also different. In followup work, Moffat et al. [8] explored the effect that those query variations have on pool size.

Bailey et al. subsequently created a further set of 100 information-need statements in connection with a different document collection, and collected partial relevance judgments [2]. Here we examine the pooling implications of the new backstories, and explore the relationship between system-based pooling and query-based pooling, evaluating a set of nearly 6,000 different queries on five different systems. We confirm the observations of Moffat et al. [8] in regard to query-based pooling; more critically, we also show that system-based variation and query-based variation are multiplicative rather than additive, and that the two factors compound. This behavior has important implications for the design of IR test environments in which query variation is also to be included.

ACM 978-1-4503-4865-2/16/12 ... $15.00
DOI: http://dx.doi.org/10.1145/3015022.3015025

2. POOL GROWTH IN UQV100

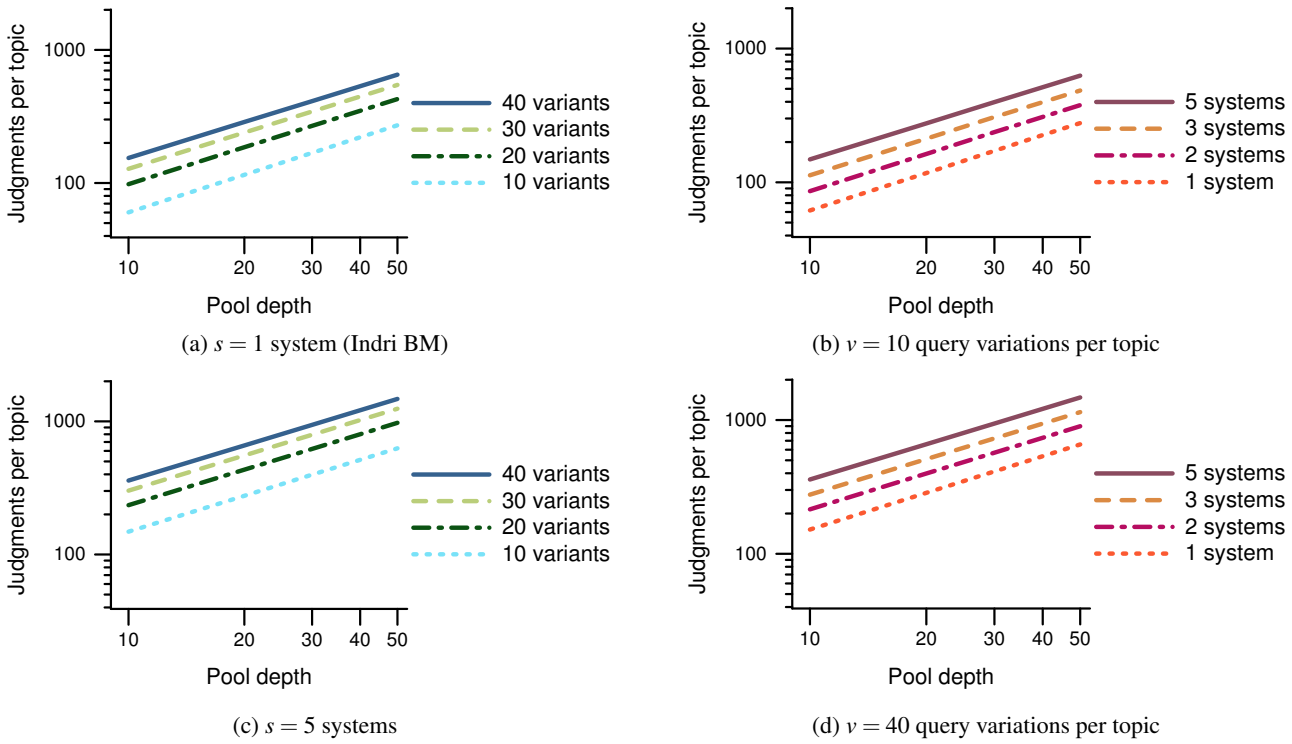
Topics and Queries The UQV100 collection [2] consists of 100 information-need statements derived from TREC 2013 and TREC 2014 ClueWeb topics; a set of approximately 100 user-generated queries for each of those topics; and (currently) 55,587 relevance judgments[1]. Of the total set of 10,835 queries supplied by the crowd workers, there are 5,765 distinct topic-query combinations – that is, on average, each distinct query occurs (only) 1.9 times.

Systems A total of five different retrieval systems have each executed all of those queries, in order to feed runs into this evaluation. The five systems are: Indri[2], using a BM25-based computation; Indri, using a language model-based evaluation; Terrier[3] [7], using a PL2 computation; Terrier, using a DFRFree computation; and Atire[4], using a quantized-impact computation. In total, 28,869 distinct system-topic-query combinations were generated (not a multiple of 5,765 because an ongoing data curation process meant that slightly different query sets were run by each system).

Judgments Partial relevance judgments for the 100 topics are provided as part of the UQV100 data resource. The current set of 55,587 judgments is the result of several rounds of activity, starting with a depth-10 pool based on the two Indri runs, and then extended to ensure coverage of the top-10 of the runs for the other three systems as they became available, and further augmented by evaluation of some "popular-but-not-judged" documents. Each document was judged by three crowd workers against a five-point scale, listed below; and a median label selected [2].

Pool Growth Three quantities affect the per-topic size J of a rank-based query-variations judgment pool: s, the number of systems that contributed runs; v, the number of query variations incorporated per topic; and d, the pool depth selected. In a worst-case sense, all that can be said is $J \le s \cdot v \cdot d$ (as a fourth factor, there are t distinct topics involved; all results here are averages over a set of $t = 100$ topics in the UQV100 collection). To measure the empirical relationship between s, v, d, and J, pools were formed based on a range of combinations. Figure 1 show the results of these experiments. The

[1] http://dx.doi.org/10.4225/49/5726E597B8376
[2] http://www.lemurproject.org/indri/
[3] http://terrier.org
[4] http://atire.org



(a) $s = 1$ system (Indri BM)

(b) $v = 10$ query variations per topic

(c) $s = 5$ systems

(d) $v = 40$ query variations per topic

Figure 1: Growth in pool size, computed as average number of documents per topic, averaged over 100 different topics. In the left column, query variations are altered within each graph; in the right column, the number of included systems is varied within each graph. The system ordering from $s = 1$ to $s = 5$ is Indri-BM, Indri-LM, Terrier-DFR, Terrier-PLC, Atire. All axes are logarithmic.

two left-hand graphs show pool sizes for two values of s, with the four plotted lines showing different values of v; the right-hand pair show pool sizes for two values of v, with lines showing different values of s. Both axes in each of the four graphs are logarithmic, and the parallel straight lines that appear for each combination indicate in each case that $J \approx kd^c$, where k in turn depends on s and v.

Fitting a Curve Similar relationships are evident if the data is plotted with each of s and v on the horizontal axis. Overall, if $s \cdot v$ runs are pooled to depth d, then the anticipated number J of judgments to be performed per topic is empirically approximated by

$$J = 1.85\, s^{0.52}\, v^{0.63}\, d^{0.89}\,. \qquad (1)$$

That is, the volume J of judgments required will double if the number of systems increases by a factor of 4, or if the number of query variations per topic increases by a factor of 3.2, or if the pool depth increases by a factor of 2.2. For example, for $s = 5$, $v = 40$, and $d = 20$, a pool of $J = 648$ judgments per topic is predicted by Equation 1; the actual average value plotted in Figure 1 is 659, close to that figure. More generally, the Pearson correlation coefficient between the measured pool sizes and those predicted by Equation 1 over the 100 data points given by $s \in \{1,2,3,4,5\}$, $v \in \{10,20,30,40\}$, and $d \in \{10,20,30,40,50\}$ was greater than 0.998. Despite the fact that (to date) only five systems have been included, Equation 1 is also in broad agreement with the results given by Moffat et al. [8], who note the approximate relationships $J \approx d \cdot u^{0.7}$ and $J \approx d \cdot s^{0.5}$ based on two separate evaluations, one in which the number of TREC contributing systems is varied over a much wider range than is used here, and one in which the number of users u is varied (and noting that $v \approx 0.5 \cdot u$). Other collections that include query variations, such as CLEF eHealth data (2015 and 2016), could also be analyzed in the same way.

3. LOCATING RELEVANCE

Distribution of Outcomes Of the 55,587 graded judgments associated with the current UQV100 collection, 43,471 (78.2%) of the labels reflect documents deemed to be *not useful*; 7,918 (14.2%) of the assigned labels were *slightly useful*; 3,046 (5.5%) were *mostly useful*; 1,058 (1.9%) were *very useful*; and 94 (0.2%) of the documents judged were assessed as being *essential*. That is, 11,482 documents, or 20.7% of the ones judged, are relevant at the *slightly useful* level or better. If the pool is limited to a strict depth of $d = 10$ (the UQV100 judgments include some documents beyond that depth), a total of 50,980 documents are covered. Figure 2 plots the per-topic distribution of those judgments, and shows that (a) there is considerable variation in numbers of documents in the per-topic pools; and (b) that while the average fraction of *slightly useful* or greater documents is a little over 20% of those judged, there is considerable variation in that regard too.

Pooling in Practice Figure 3 shows the distribution of relevant documents across the system-query combinations described earlier. The top line shows, for the set of documents at each rank k across the runs that contributed to the judgment pool, the fraction of documents for which judgments exist. Rank-based pooling to depth d will thus result in a horizontal line at 1.0 out to d. At ranks $k > d$, documents in one run may be judged because of their appearance at depth $k' \le d$ in another run, hence the gradual reduction in the fraction of documents judged as the ranks k extend beyond d. The other two lines show the fraction of documents at that depth (counting repeated documents once for each run that contains them at that given depth) that were judged *not useful*, and *slightly useful* or better. The two lower curves sum to the top curve.

An important conclusion from Figure 3 is that systems continue

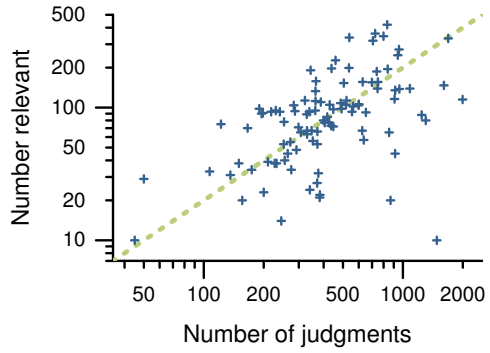


Figure 2: Distribution of relevance judgments over 100 UQV topics, and number of judgments of *slightly useful* or better, based on a set of 50,980 relevance judgments formed by pooling $s = 5$ systems and all query variations to a pooling depth of $d = 10$. The dotted line represents 20% of judged documents being relevant.

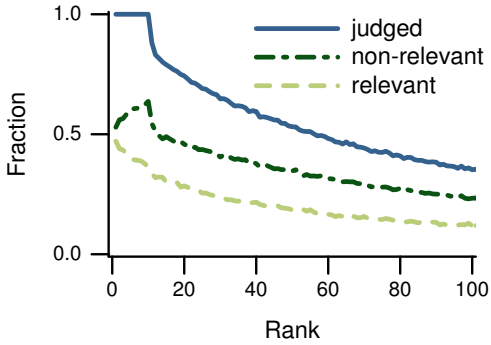


Figure 3: Coverage and relevance rates (*slightly useful* or better) for the 55,587 UQV100 judgments, including documents added to the pool at earlier ranks and repeats if the same document is presented more than once at the same rank. Five systems' runs provide a total of 28,869 query-system combinations over 100 topics.

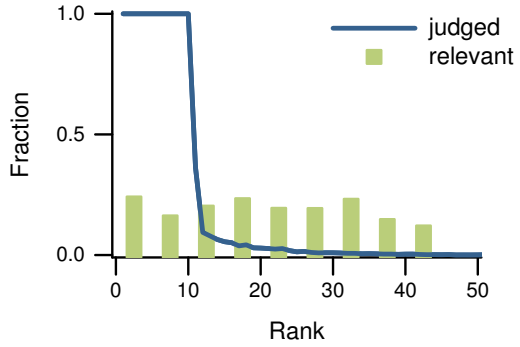


Figure 4: Fraction of hitherto unseen documents that have judgments, and (in groups of five ranks) the fraction of those judgments that are *slightly useful* or better, counting only the distinct documents presented at each rank level. Other details are as for Figure 3.

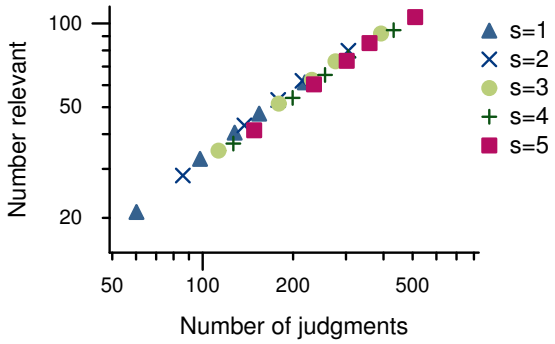


Figure 5: Average number of relevant documents identified for each UQV topic as a function of the number of judgments carried out, for the same combinations of v and s listed in Table 1 (with $d = 10$ in all cases). For each value of s the five different values of v are all plotted with the same symbol.

to return relevant documents at quite deep ranks. For example, at depth $k = 100$ there are around 35% of the 28,869 documents across the runs that have judgments, and 12% of those documents (or 34% of the documents that have been judged) are at least *slightly useful*. Another 65% of those 28,869 documents have not been judged.

New Relevant Documents at Depth Figure 4 shows the set of documents at each rank k that do not appear in any run at any depth $k' < k$, that is, the documents that make their first appearance in the pool as it is extended from depth $k-1$ to depth k. The solid line shows the fraction of that set of documents that were judged; again, that rate is 1.0 for ranks $k \le d$, where d is the pool depth. If the judgment pool was formed by strict adherence to a depth-based cutoff at rank d, then the fraction would be 0.0 for ranks $k > d$; as can be seen, in the UQV100 collection a small number of documents beyond the pool cutoff were also selected and judged.

The bars in Figure 4 show the fraction of those new-to-pool and newly-judged documents that were deemed *slightly useful* or better, averaged across bands of five ranks for each bar. For example, the last visible bar covers ranks 41 to 45; of the 41 new-to-pool documents in that band that were judged, there were 5 that were judged at least partially relevant, or 12% of that subset. Another 18,427 documents were new-to-pool across those five ranks, but were *not* selected for judgment. In the next zone, from ranks 46 to 50, a further 16 documents were judged (of a total of 18,197 documents that had their first occurrence within that band of ranks), but none were relevant, hence the missing bar.

The best way of summarizing Figures 3 and 4 is that relevant documents continue to arise at relatively deep ranks, and that the UQV100 judgments should not be assumed to have identified all of the relevant documents for each topic. Similar outcomes have been identified for other collections [6, 10].

Total Relevance Volume Figure 5 shows the average number of judgments required per topic and the number of *slightly useful* or better documents identified, as the pool is expanded by adding further systems and/or adding further query variations, in two independent dimensions. Five different values of v are plotted in each case, but not individually identified. That the points line up in a single band for both sources of variation indicates that there is no quality distinction to be made between pools built from system variation and pools built from query variation – both generate additional judgments, and those judgments give rise to similar numbers of relevant documents.

Fidelity of Measurement Bailey et al. [1] also introduce an effectiveness metric INST, parameterized by a quantity T, the expected volume of relevance sought by the user. INST models users as pro-

Systems	Query variations				
	$v = 10$	$v = 20$	$v = 30$	$v = 40$	*all*
$s = 1$	0.53	0.43	0.38	0.35	0.30
$s = 2$	0.45	0.36	0.32	0.28	0.23
$s = 3$	0.41	0.33	0.28	0.25	0.21
$s = 4$	0.39	0.31	0.27	0.24	0.19
$s = 5$	0.32	0.23	0.18	0.15	0.10

Table 1: Average INST residual for Atire over 10,835 query variations using pools constructed from different combinations of v and s. The Atire run contributed to the pool only in the $s = 5$ row. All evaluations in the table are carried out using a pool depth of $d = 10$, and with all *useful* grades converted to a gain of 1. The largest judgment set ($s = 5$, $v = all$) covers 50,980 document-topic pairs.

ceeding further down the ranked list if they encounter less relevance; meaning that their propensity to continue examining documents increases if the documents seen so far do not result in increasing satisfaction relative to their initial target T. That is, INST is a weighted-precision metric that is both goal-sensitive and adaptive.

As is the case for other weighted-precision metrics, a *residual* can be computed to estimate the volume of assessment "weight" attached to unjudged documents. To compute INST scores and residuals, a mapping is required from relevance grades (in the UQV100 collection: *not useful*, *slightly useful*, *mostly useful*, *very useful*, and *essential*) to numeric *gain* values that sum towards T. Moreover, because INST is adaptive, that mapping also affects the residuals. Residuals are minimized when gain is maximized, and so for the purpose of generating lower bounds on residuals, we assume an optimistic gain mapping *not useful* $\rightarrow 0$, and *slightly useful* and better $\rightarrow 1.0$. Table 1 shows average INST residuals under this scenario, across the 10,835 queries (5,765 distinct) when processed via the Atire system, based on judgment pools of differing numbers of query variations v, and on differing numbers of systems s. When $s < 5$, the Atire system does not contribute to the pool; when $s = 5$ it does. Each topic's queries were processed with a single T value, being the mean of the T estimates supplied by the crowd workers who generated the queries. Over the full set of 10,835 queries the mean of the T parameters used was 4.70, corresponding (approximately) to a rank-biased precision parameter $0.8 \le p \le 0.9$.

If the Atire run is not included in the pool (creating a "leave one run out" experiment), then the smallest average residual is 0.19, a value that is of the same magnitude as the INST scores being generated. Even when Atire is included in the pool (the row marked $s = 5$), and the full set of UQV100 judgments is used (55,587 document-topic pairs, including some judgments beyond $d = 10$), the residual remains non-trivial, at 0.08. These high values imply that further refinement is required before the UQV100 judgments can be used with confidence to calculate INST scores for new systems.

If different relevance mappings are used, the situation worsens. Figure 6 shows the spread of residual values for the Atire runs when *slightly useful* or better is regarded as being relevant (gain 1.0), and also when a more restrictive binarization is used, with only *mostly useful* and better given a non-zero gain. In both curves it is topics with high values of T (the largest in UQV100 is $T = 8.2$), corresponding to deeper expected inspection of the ranked list by the user, that generate the high residuals at the right. Figure 6 also shows the residuals associated with rank-biased precision with $p = 0.85$. RBP provides more accurate evaluations via consistently smaller residuals, but without the virtues of being goal sensitive or adaptive. If a higher value of p were used, the residuals would increase.

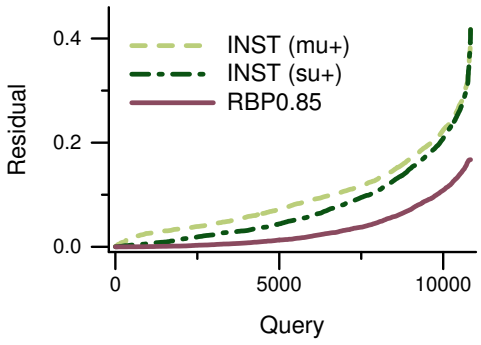


Figure 6: INST and RBP residuals, sorted by value, for the Atire runs and for 10,835 query variations (5,765 distinct) over 100 topics, scored using the full set of 55,587 UQV100 relevance judgments. The line "INST (su+)" uses binarized judgments with *slightly useful* and better mapped to a gain of 1.0; the line "INST (mu+)" has *mostly useful* and better mapped to a gain of 1.0.

4. CONCLUSION

We have examined some of the characteristics of the relevance judgments generated for the UQV100 collection, and demonstrated that the variations arising from users via their queries are just as important as the variations arising from system differences, in terms of both determining the size of the pool, and also in terms of locating relevant documents. This combination means that when query-based variation is included, pools become substantially larger than previously, and the cost of accurate experimentation is compounded.

Acknowledgment This work is part of a project undertaken in collaboration with Peter Bailey, Falk Scholer, and Paul Thomas. Matt Crane, Xiaolu Lu, David Maxwell, and Andrew Trotman assisted greatly, providing the system runs that were analyzed. The UQV100 judgments were generated using resources provided by Microsoft.

References

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

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016. Public data: `http://dx.doi.org/10.4225/49/5726E597B8376`.

[3] 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 53–75. MIT Press, 2005.

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

[5] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a large-scale test collection: An analysis of the search results from the first NTCIR workshop. *Inf. Retr.*, 5(1):41–59, 2002.

[6] X. Lu, A. Moffat, and J. S. Culpepper. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.*, 19(4):416–445, 2016.

[7] C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis. From puppy to maturity: Experiences in developing Terrier. *Proc. Wrkshp. Open Source IR*, pages 60–63, 2012.

[8] 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.

[9] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. & Trends in IR*, 4(4):247–375, 2010.

[10] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pages 307–314, 1998.