Pooled Evaluation Over Query Variations: Users are as Diverse as Systems

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

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

Paul Thomas
CSIRO, Australia
paul.thomas@csiro.au

Peter Bailey
Microsoft, Australia
pbailey@microsoft.com

ABSTRACT

Evaluation of information retrieval systems with test collections makes use of a suite of fixed resources: a document corpus; a set of topics; and associated judgments of the relevance of each document to each topic. With large modern collections, exhaustive judging is not feasible. Therefore an approach called *pooling* is typically used where, for example, the documents to be judged can be determined by taking the union of all documents returned in the top positions of the answer lists returned by a range of systems. Conventionally, pooling uses system variations to provide diverse documents to be judged for a topic; different user queries are not considered. We explore the ramifications of user query variability on pooling, and demonstrate that conventional test collections do not cover this source of variation. The effect of user query variation on the size of the judging pool is just as strong as the effect of retrieval system variation. We conclude that user query variation should be incorporated early in test collection construction, and cannot be considered effectively post hoc.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*performance evaluation*.

Keywords

User behavior, test collections, relevance measures

1. INTRODUCTION

Batch-mode system evaluation is used extensively in information retrieval system measurement. Compared to user studies in natural or laboratory settings, batch mode experimentation has two key advantages: first, the experiments are easily repeatable; and second, once the initial relevance judgments have been formed (assuming they are reusable), carrying out experiments typically imposes little additional cost, making the methodology cheaper overall. The key outcome of a batch evaluation experiment is a numeric measurement of system effectiveness averaged over the per-topic performance, using some form of effectiveness metric which represents the aggregate relevance of an answer list. They therefore provide a basis for repeatable longitudinal system evaluations.

Measures of system quality, in this approach, are determined from "relevance judgments" which specify, for each query and document, how good that document is for that query. In principle, these judgments cover all possible query–document pairs, but in practice it quickly becomes infeasible to judge all documents in response to a query, even for relatively small collections. To manage this cost, a technique known as *pooling* is used when relevance judgments are being created [9]. In this approach, each of the systems being measured (for example, in a shared evaluation task at an event such as TREC, NTCIR, CLEF, or FIRE) contributes a ranking of d (or more) documents, in decreasing order of predicted relevance. The union of the top-d sets is then identified, and those documents are the ones that are judged for relevance. This ensures that each system has at least its top-d documents judged (and possibly more). Extensive experimentation has shown that this approach tends to lead to an unbiased set of judgments, resulting in the same system ordering as would full evaluation [6, 15].

For the sake of reusability and reliability, pooling should cover possible sources of variation – for example, pools need to cover a large range of present (and anticipated) systems since different retrieval systems act differently and will retrieve different documents. The validity of scoring systems that didn't contribute to the original pool has been explored [4, 15]. For smaller collections, it was found that non-contributing systems are fairly evaluated when test collections are re-used; however, for large collections, where pools are small relative to the full collection, a bias may be introduced towards those documents that directly match query keywords [3].

Differences in the documents that retrieval systems return in response to a query are clearly reflected in batch evaluation. However, most further sources of variation are abstracted out of the process. For example, the usability of an interface is not considered, and a retrieval system is deemed to have completed its role when a ranked list of documents is produced. The interactions between a user and the list of document summaries are also ignored. Nor is user variability considered as a factor in the evaluation process – instead, standard queries are assumed, and a standard user response to the generated rankings is assumed, via the chosen effectiveness

CIKM'15 October 19–23, 2015, Melbourne, VIC, Australia

ACM 978-1-4503-3794-6/15/10 … $15.00.
http://dx.doi.org/10.1145/2806416.2806606.

metric that is employed, as if every user that will ever use an IR system is a genetic and cultural clone of every other user. The only variable that is retained in most experimentation is the system, with the assumption that if a sufficiently representative set of queries is employed, system differences will be reliably identified.

In a recent study, we have explored the way in which users create queries from longer, multi-sentence topic descriptions [1], seeking to identify the role that the users' *a priori* expectations have on the way they react to the result listings that get generated, a project that builds on earlier investigations by Buckley and Walz [2] and Voorhees [11]. We use those variant user queries to investigate the impact of user variability on pool size; the data gathering process is described in Section 2. Our main research question, *"to what extent are query variants adequately covered by existing pooled relevance judgments?"*, is considered in Section 3. As our analysis shows, the answer is *"very little"* – query variants derived in a natural manner from a common information need description lead to markedly different sets of documents being retrieved. This result means that considerable care is required when designing batch-mode IR experiments, and that user variation is at least as important a factor in experimental design as system variability is.

2. TOPICS AND QUERIES

User Variability To analyze the impact of user variability, we constructed English-language information need statements, and asked participants to indicate what search query they would enter into a search engine to fulfill the stated information need. Starting with a selection of 180 TREC topics drawn from the 2002 Question Answering Track [12], the 2003 Robust Track [13], and the 2004 Terabyte Track [5], we created statements that represent scenario-based motivations for the corresponding TREC queries [1]. For example, the TREC topic R03.314 asks about the benefits of eating marine vegetation and is crystallized by the "title" query "*marine vegetation"*. This topic led us to create an information need statement: "*You recently heard a commercial about the health benefits of eating algae, seaweed and kelp. This made you interested in finding out about the positive uses of marine vegetation, both as a source of food, and as a potentially useful drug*".

Crowd workers were then asked to read a subset of the statements in randomized order, and, for each information need, to indicate what their first query would be. An average of 44 responses were collected for each of 180 information needs, with a very broad range of queries generated. For example, the 47 responses to Topic R03.314 included eight instances of "*positive uses of marine vegetation*", a phrase extracted directly from the information need statement; but there were also 37 other queries constructed, 35 of which occurred just once, including "*marine vegetation as food or drugs*" and "*edible seaweeds*". Just one person issued the query "*marine vegetation*" that matched the TREC title query for the topic.

Our goal in this part of the project was to measure the extent to which the different queries that were submitted for the same information need retrieve documents that are not fetched by the canonical TREC title query. To do this, each query variant was executed using two standard retrieval mechanisms (Indri[1] implementations of Okapi BM25 [10] and the Sequential Dependency Model [7] ranking functions) to generate answer rankings. The rankings for each topic were then compared in two different ways, seeking to quantify the extent to which the rankings overlapped. For this investigation on pooling, we omit topics from the Question Answering track, primarily because the judging process was oriented towards identification of facts rather than ad hoc relevance labels.

[1] `www.lemurproject.org/indri`

Pool Size In the first approach, the size of the pool of documents that would need to be judged if all variant user queries were to be evaluated fairly by a depth-d effectiveness metric is computed, regardless of whether or not the TREC topic query is present in the set. For example, if $d = 10$, the set of 38 distinct queries generated for R03.314 places anywhere between 10 and 380 documents in to the judgment pool. In fact, over the 38 queries (from 47 crowd workers), a pool of 153 documents arises at $d = 10$, and 295 at $d = 20$. Full details of pool size as a function of d appear in Section 3.

Ranked List Similarity The second approach takes the canonical TREC topic-only query as a reference point, and computes the average divergence from that document ranking. To give appropriate weighting to the more highly ranked documents, and to allow for the fact that the pairs of lists are not necessarily permutations of each other (which is required to compute Kendall's tau), we employ the *Rank-Biased Overlap* computation (RBO) described by Webber et al. [14]. Rank-Biased Overlap quantifies the expected overlap observed by a probabilistic user who compares the two lists, advancing from depth d' to depth $d'+1$ with probability p. When $p = 0.95$, the value used in our experiments, the expected depth reached is $d' = 20$, but the measure is top weighted, and the greatest single contribution (with weight 0.05) arises at depth $d' = 1$, depending entirely on whether the two lists have the same first element. Hence, a high value of RBO indicates that the two lists are similar, at least to a user with the level of persistence p; conversely, low values of RBO indicate dissimilar or non-overlapping orderings.

Note there is no particular reason to suppose that the TREC topic-only query is more "correct" for the topic than is any of the (other) crowd-generated variants, and we use it as the reference point purely because it represents a default starting position. Any query could have been used as an anchor point for each topic; or all pairwise RBO scores could have been averaged. Our use of the TREC topic-only queries is primarily in deference to more than two decades of TREC-sponsored batch-mode retrieval experimentation; but also because these are the queries that have the most complete coverage in the TREC judgments created in the past.

3. USERS VERSUS SYSTEMS

Pool Size Figure 1 shows how total pool size grows as a function of the number of systems contributing to the pool for each topic (top), and of the number of users contributing to the pool for each topic (bottom), in both cases averaged across 60 topics selected from the TREC 2003 Robust Track (see Bailey et al. [1] for details of the topics). The "systems" are the 78 runs submitted by the TREC participants, taken in alphabetical order of run name, and are identical for each topic. In the lower graph, the horizontal axis represents queries as generated by "users" (crowd workers) based on the information need statements; they are ordered by workers' unique CrowdFlower ID. Repeat queries are included in the count, even though they do not increase the pool size. Note that the pool of users varies across topics, because most of the workers only processed a subset of the 60 selected topics.

The trend in Figure 1 is clear: for each of the pool depths d, the growth in total pool size follows the same trends as it does for systems. Both options give rise (very broadly) to straight lines on log-log graphs, as plotted, but with slightly different exponents: in the case of systems, the four lines are approximated by $v \approx dn^{0.5}$, where v is the volume of judgments, d is the pool depth, and n is the number of systems; in the case of users, the growth rate is higher, with $v \approx dn^{0.7}$. That is, for these users, and these systems, the pools that arise from a given number of users are larger than the pools that arise from the corresponding number of systems.

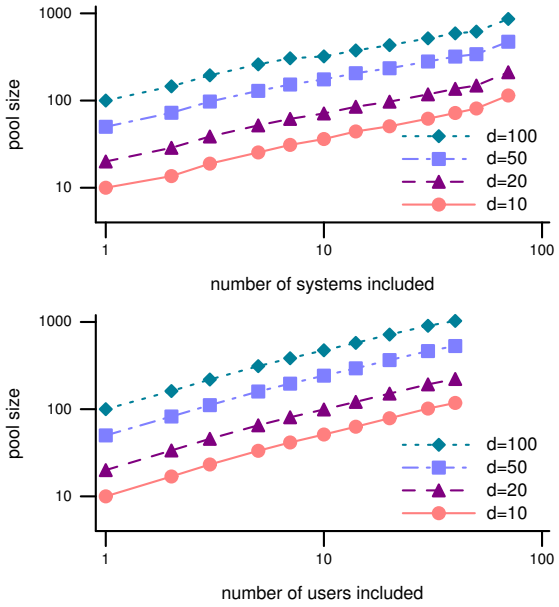

Figure 1: Pool size (average documents per topic) as a function of systems (top) and users (bottom), using 60 topics from the TREC 2003 Robust Track, and a range of pooling depths d. The "systems" are TREC runs, ordered (and included incrementally in the pool) alphabetically; the "users" are crowd workers, ordered (and included incrementally in the pool) by their CrowdFlower ID. The set of workers who created queries differs from topic to topic; in total, 92 workers contributed an average of 28.7 queries each.

Ranked List Similarity Figure 2 shows the RBO scores for system-versus-system and user-versus-user comparisons for the TREC 2003 Robust Track and the TREC 2004 Terabyte Track. In this experiment the "systems" are TREC contributed runs that self-identified as having made use of title-only queries. In the case of the 2003 Robust Track, there were 8 runs used, across 60 topics and hence $60 \times 8 \times 7/2 = 1{,}680$ RBO scores computed. In the case of the 2004 Terabyte Track, the eight highest scoring title-only runs as identified by the Track organizers [5, Figure 6] were used, across 50 topics, and hence 1,400 RBO scores were generated. To compare "users", the sets of user queries for each topic were evaluated using the BM25 and SDM computations implemented in Indri, and then those runs compared against the TREC title-only run for that topic using the same computation. In the 2003 Robust Track, this yields 2,649 RBO scores; for the 2004 Terabyte Track, there are 2,222 RBO scores generated. Figure 2 plots the resulting RBO distributions. In both cases a one-sided Mann-Whitney U test indicates that the "varying user" RBO scores (with persistence parameter 0.95) were significantly lower than the "varying system" scores, at $p < 0.001$, further supporting our contention that user variability is at least as high as system variability.

Strategic Pooling Moffat et al. [8] suggest that if a limited judgments budget is available, documents in the pool could be ordered according to how much they reduce the total amount of imprecision in the measurement system as a whole. They do this in the context of weighted-precision effectiveness metrics – Rank-Biased Precision (RBP) in particular – by summing the "votes" for each document, and if J judgments can be carried out, identifying the J documents with the largest sums of weighted votes. The selection can be done on a topic-by-topic basis, or on a global all-topics basis.

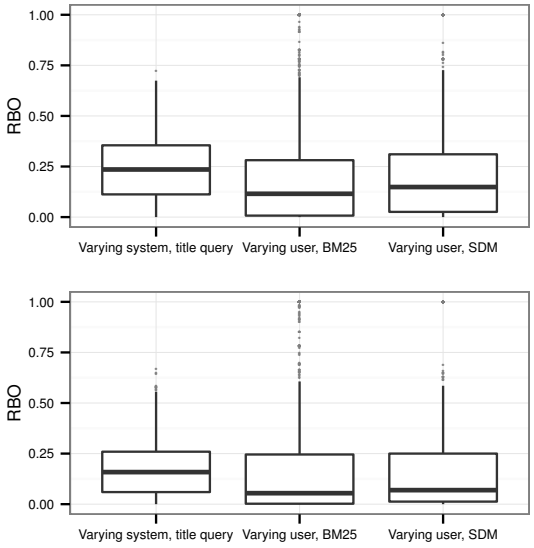

Figure 2: RBO comparisons between ranked lists for the TREC 2003 Robust Track (top), and the TREC 2004 Terabyte Track (bottom).

Figure 3 shows how the 1,438 TREC relevance judgments for topic R03.314 would sit with this approach. Each of the user query variations was evaluated using Indri's Okapi BM25 computation, and a run of 200 documents created. Each document in each run was then assigned a weight depending on its position in the run, with the document at rank r given a weight of $(1-p)p^{r-1}$, with persistence parameter $p = 0.95$. The total weight of each document was the sum of the weight accorded that document in the 47 runs generated from the query variations, including repeats, or in the 78 system variations. To create the document ordering shown in the horizontal axes in the figure, the pool was then ordered by decreasing total weight, and plotted against minimum rank position, that is, the shallowest depth in any of the corresponding runs at which that document appeared. The correlation that shows is a consequence of the minimum RBP weight associated with each rank. A gray square is plotted for each document. The TREC relevance judgments are then overlaid – red crosses for the 1,394 documents judged as non-relevant, and black plus signs for the 44 documents judged relevant, for the first 1,000 documents in each of the two pools.

Tables 1 and 2 detail the relationship between the system-generated pools for the selected topics, the user-generated pools (using the Indri Okapi similarity rankings), and the existing TREC relevance judgments. In both data sets the user-generated pools have many more unjudged documents than the system-generated pools, further evidence that user variations give rise to quite different sets of documents being retrieved. In the rows labeled "combined", all system runs and all Okapi-based user runs are combined in a single pool. The fact that the combined pools are close in size to the sum of the two separate pools highlights the relatively disjoint nature of the respective sets of documents.

4. CONCLUSION

Our findings can be summarized simply: diversity in the pools of documents for judging arising from user variation is at least as substantial as that from system variation. We have also demonstrated that pools grow in size at a comparable rate in each case. Even when using the same system, different user queries typically produce very

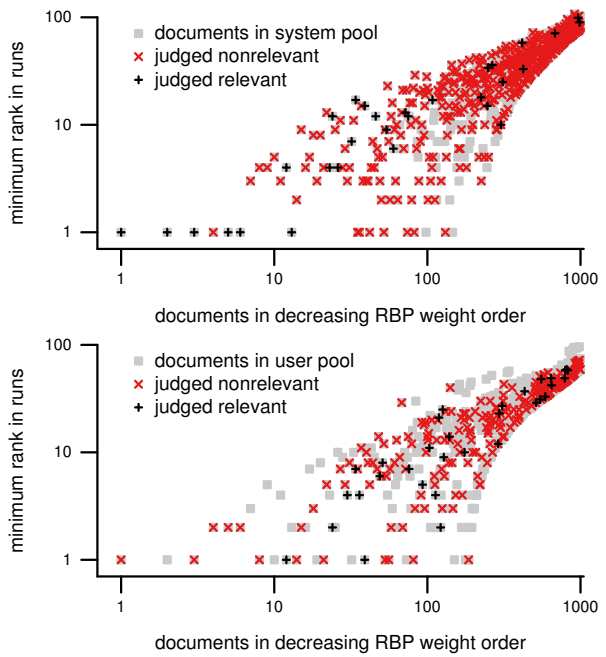

Figure 3: Depth of judged and unjudged documents for TREC 2003 Robust Track topic R03.314, pooled by systems (top), and users (bottom), covering the first 1,000 documents ordered by aggregate RBP weight. The horizontal axis orders documents by their summed RBP-like assessment weights when evaluated using BM25; the vertical axis shows the minimum rank of that document in any of the runs. The $d = 200$ pool for the 47 query variations for this query contains 3,292 documents; the $d = 200$ pool for the 78 system variations contains 2,131 documents.

Source	Depth	Size	Rele.	Irre.	Unjd.
TREC systems (78 runs)	$d = 20$	216.0	11.3%	74.0%	14.7%
	$d = 100$	883.1	5.5%	61.5%	32.9%
Users (44.2 avg.)	$d = 20$	240.5	10.3%	45.8%	43.9%
	$d = 100$	1110.9	4.5%	32.3%	63.3%
Combined (122.2 avg.)	$d = 20$	394.7	8.5%	56.3%	35.2%
	$d = 100$	1701.2	3.6%	39.2%	57.2%

Table 1: Fraction of pool covered by existing judgments, averaged over 60 TREC 2003 Robust Track topics evaluated using BM25. The judgments for these topics were created in 1998, 1999, 2000, and 2003, with the 2003 runs not re-judged against the earlier topics, see Voorhees [13].

different ranked lists. Finally, we have shown that incorporating user variation cannot be a post-hoc exercise, after a collection has been created in a conventional manner using variation solely from systems. There are substantially larger sets of unjudged documents from new "user runs" than existing "system runs", and these can occur high in rankings. Examining combined pools, we found that approximately one third of documents are unjudged at a pool depth of 20, and more than half are unjudged at depth 100.

We conclude that user variation should be designed in when creating test collections, by collecting different query responses to specific information needs, and that this would lead to more representative sets of documents being judged. In future work, we

Source	Depth	Size	Rele.	Irre.	Unjd.
TREC systems (70 runs)	$d = 20$	445.7	22.4%	63.0%	14.6%
	$d = 100$	1912.9	11.6%	49.0%	39.4%
Users (44.4 avg.)	$d = 20$	236.7	25.4%	28.6%	46.0%
	$d = 100$	974.7	14.3%	22.0%	63.7%
Combined (114.4 avg.)	$d = 20$	607.7	19.3%	51.6%	29.1%
	$d = 100$	2527.8	9.1%	38.3%	52.6%

Table 2: Fraction of available pool covered by existing judgments, averaged over 50 TREC 2004 Terabyte Track topics evaluated using the BM25 retrieval mechanism. The original judging for these topics covered a subset of the 70 runs submitted to the Track, and to a depth of $d = 85$ [5].

plan to investigate building just such a test collection. The collection will allow us to examine various aspects of relevance evaluation, when using both a diversity of systems and a diversity of users.

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (projects DP110101934 and DP140102655). We thank Xiaolu Lu, who generated all of the user runs. This work was approved by the CSIRO Human Research Ethics Committee, ref. 077/14.

References

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

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

[3] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *J. Inf. Ret.*, 10(6):491–508, 2007.

[4] S. Büttcher, C. L. Clarke, P. C. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. SIGIR*, pages 63–70, 2007.

[5] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 terabyte track. In *Proc. TREC*, 2004.

[6] D. Harman. The TREC test collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 2, pages 21–52. MIT Press, 2005.

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

[8] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. SIGIR*, pages 375–382, 2007.

[9] K. Sparck Jones and C. J. van Rijsbergen. Report on the need for and the provision of an "ideal" information retrieval test collection. Technical report, Computer Laboratory, University of Cambridge, 1975. British Library Research and Development Report No. 5266.

[10] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. Part 1. *Inf. Proc. Man.*, 36(6):779–808, 2000.

[11] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Proc. Man.*, 36(5):697–716, 2000.

[12] E. M. Voorhees. Overview of the TREC 2002 question answering track. In *Proc. TREC*, 2002.

[13] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proc. TREC*, 2003.

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

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