Judgment Pool Effects Caused by Query Variations

Alistair Moffat, The University of Melbourne

1. Query Variations

Bailey *et al.* [1] describe the UQV100 collection, consisting of 100 topics based on ClueWeb12 queries (TREC 2013 and 2014), and a set of crowd-sourced queries for each, totaling 10,835 queries, and averaging 57.7 distinct queries per topic.

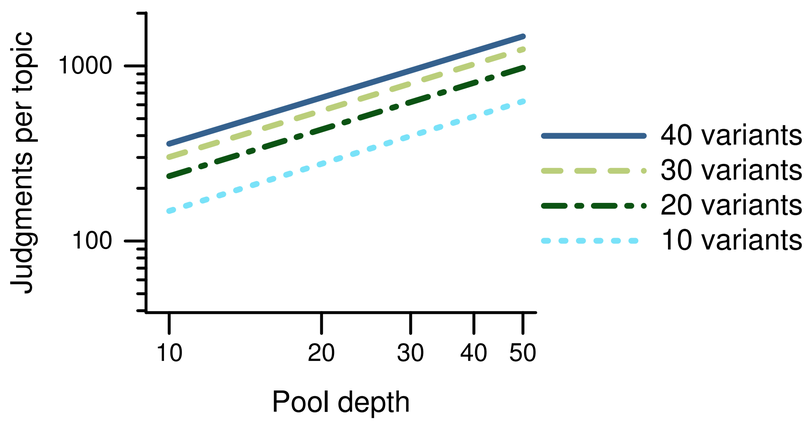
With help from others, these queries were executed on five different retrieval systems: Atire, Indri-BM, Indri-LM, Terrier PL2, and Terrier DFRFree.

Question: *If these systems/queries are pooled to depth d, how many relevance assessments are required?*

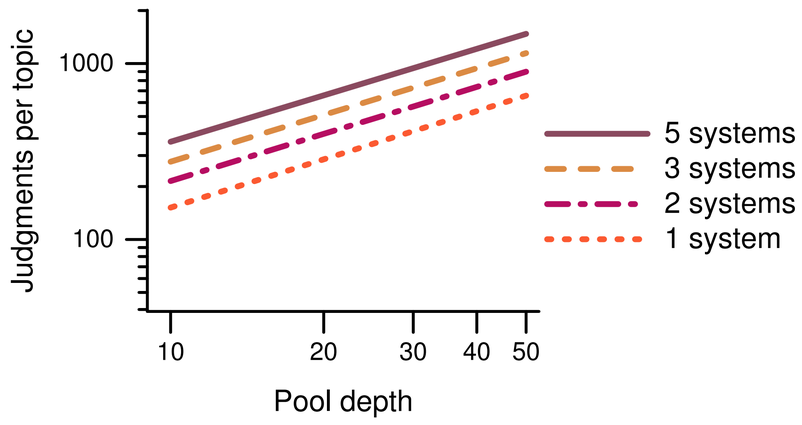
2. Collecting Judgments

If s systems, v query variations per topic, and depth d, then $J \le s \cdot v \cdot d$.

For $s = 5$ systems (average across 100 topics):



For $v = 40$ query variations (average across 100 topics):



Straight lines indicate $J \approx kd^c$, where k depends on s and v.

3. Fitted Equation

Over 100 data points given by $s \in \{1,2,3,4,5\}$, $v \in \{10,20,30,40\}$, and $d \in \{10,20,30,40,50\}$,

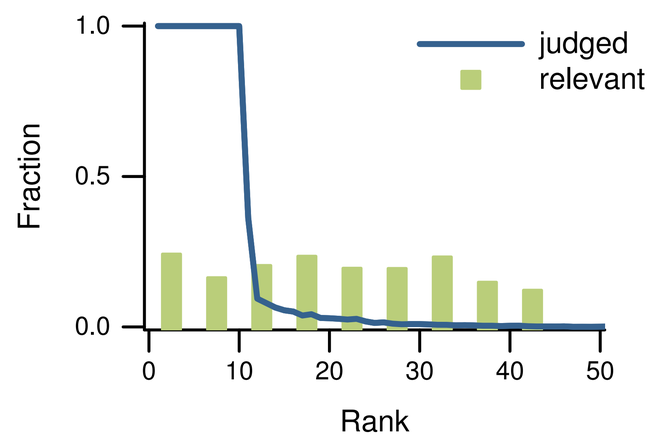
$$J = 1.85\, s^{0.52}\, v^{0.63}\, d^{0.89}$$

yields a Pearson correlation coefficient greater than 0.998.

Moffat *et al.* [3] estimate $J \approx d \cdot u^{0.7}$ for one system, where $u \approx 2v$ is the number of users; and $J \approx d \cdot s^{0.5}$ for TREC systems and unknown queries.

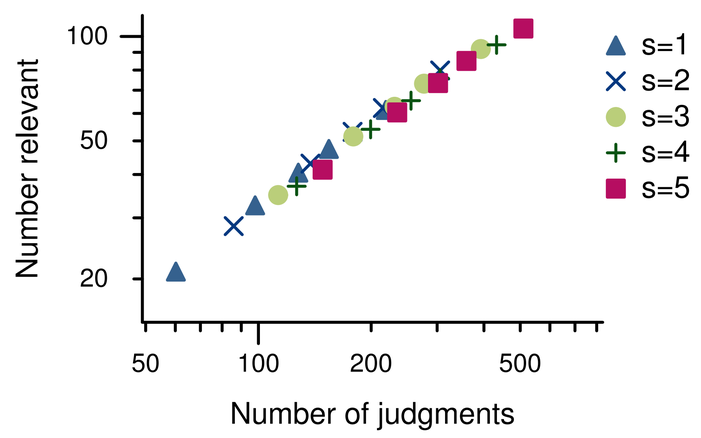
4. Locating Relevance

A total of 55,587 judgments have been collected. The graph shows the fraction of new-at-depth documents with judgments, and fraction relevant, across 28,869 query-system combinations:



Relevant documents continue to be discovered at roughly the same rate.

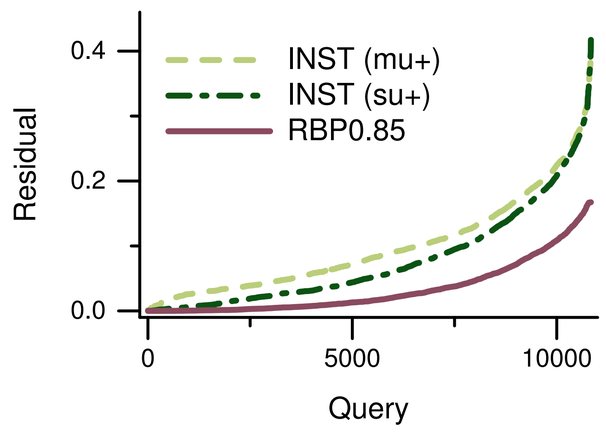
Number of *slightly useful* or better documents, with $d = 10$ and $v \in \{10,20,30,40,all\}$:



Over the range explored, systems and queries are *equally useful*.

5. Residuals

The *residual* in a weighted-precision metric score represents the influence of the unjudged documents. Even with 55,587 judgments, when the runs are scored, many of them remain with non-trivial residuals:



INST is an adaptive metric that incorporates user expectations [2]. With a different binary gain mapping (*mostly useful*+, rather than *slightly useful*+), the residuals increase, because the INST scores decrease. RBP scores also decrease, but the residuals are unaffected.

6. Acknowledgments and References

This work is part of a larger project undertaken in collaboration with Peter Bailey, Falk Scholer, and Paul Thomas [1, 2, 3]. The system runs were generated by Matt Crane, Xiaolu Lu, David Maxwell, and Andrew Trotman; thanks, guys! The UQV100 judgments were generated using resources provided by Microsoft, and are available from `dx.doi.org/10.4225/49/5726E597B8376`.

[1] Bailey, Moffat, Scholer, Thomas. UQV100: A Test Collection With Query Variability, *SIGIR 2016*.

[2] Moffat, Bailey, Scholer, Thomas. INST: An Adaptive Metric for Information Retrieval Evaluation, *ADCS 2015*.

[3] Moffat, Scholer, Thomas, Bailey. Pooled Evaluation Over Query Variations: Users are as Diverse as Systems, *CIKM 2015*.

THE UNIVERSITY OF MELBOURNE