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

Alistair Moffat, Falk Scholer, Paul Thomas, Peter Bailey

THE UNIVERSITY OF MELBOURNE   RMIT UNIVERSITY   CSIRO   Microsoft

1. Batch Evaluation

Test collection-based IR evaluation provides low-cost and repeatable evaluations.

To create the required *judgments*, a technique known as *pooling* is employed, to ensure that variation across systems is catered for.

Contributing systems are free to run their own query for that topic. But many use the standard "topic title" as their query.

Research Question: Given an information need, how big are the variations caused by *user* differences? Does pooling based on system variations cater adequately for user variations?

2. Information Needs

A set of 180 TREC topics was extracted:

- Question Answering Track, 2002, 70 topics, 1824–1893;
- Robust Track, 2003, 60 topics selected from 303–610;
- Terabyte Track, 2004, 50 topics, 701–750

A *backstory* was written for each topic to motivate the information need.

For example, Topic R03.314, "marine vegetation":
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.

3. Crowd-Sourced User Variations

Crowd workers were asked to read the backstory, and then answer "what would your first query be?" An average of 44 responses were collected for each of the 180 topics.

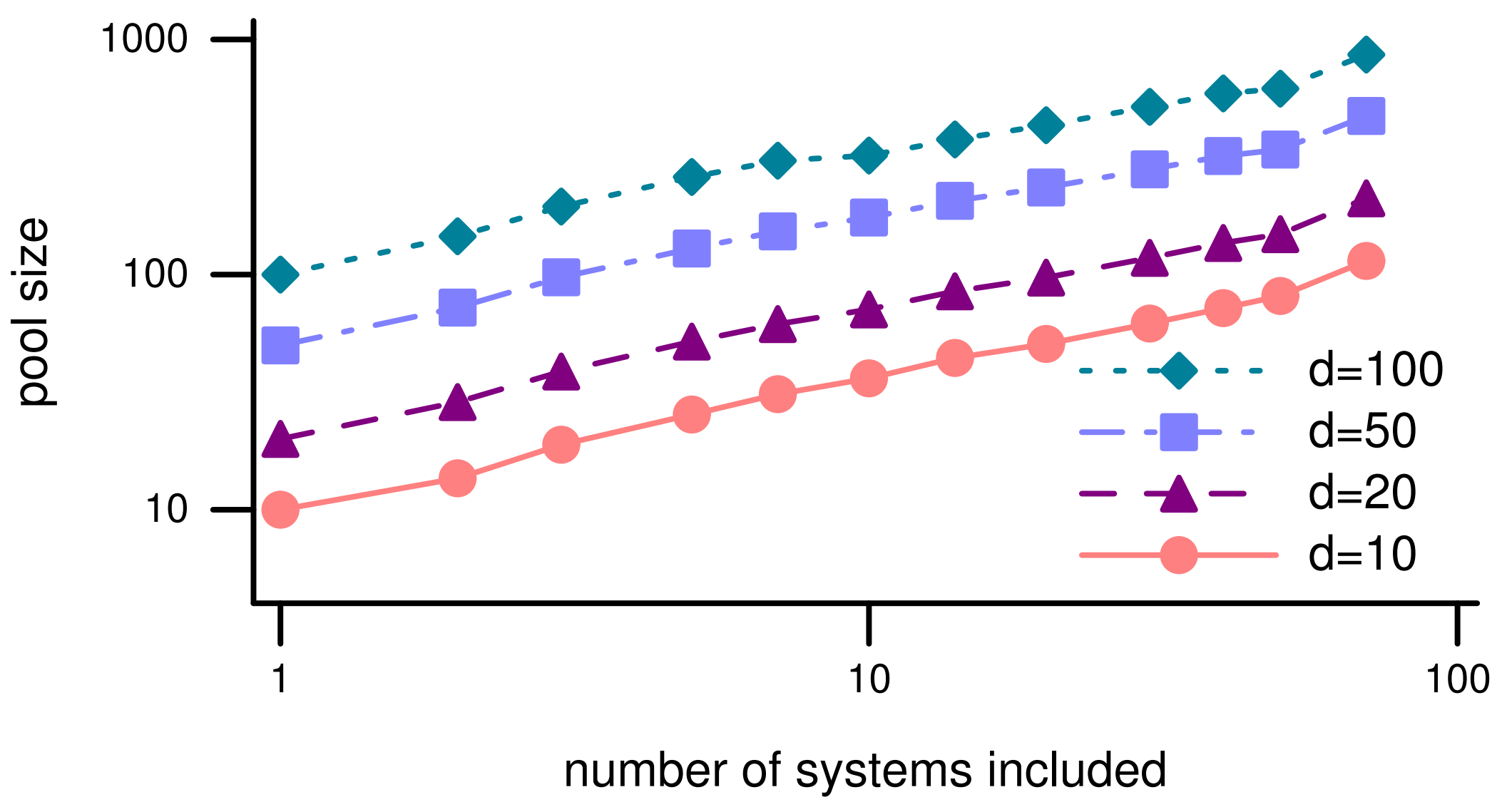
Very few first queries matched the corresponding TREC topic.

algae health benefits, algae seaweed kelp nutrition medicine, benefits of eating algae seaweed and kelp, benefits of marine vegetables, benefits to eating algae seaweed and kelp, different application of marine vegetation, edible seaweeds, finding out about the positive uses of marine vegetation, health benefits, health benefits of algae seaweed and kelp, health benefits of marine vegetation, health benefits of seaweed algae kelp food supply medical benefits, is sea veggies really good for you, **marine vegetation**, marine vegetation algae seaweed kelp, marine vegetation as food or drugs, marine vegetation benefits, marine vegetation food and a drug use, marine vegetation food and drugs, marine vegetation good for health, marine vegetation health benefits, marine vegetation positive effects, marine vegetation positive uses, marine vegetation uses, positive uses of marine vegetation (×8), positive uses of marine vegetation as source of food, positive uses of marine vegetation both as a source of food and as a potentially useful drug (×2), research into health benefits of algae seaweed and kelp, the positive uses of marine vegetation, the uses of marine vegetation in food and drugs, uses of algae seaweed and kelp, uses of marine vegetation (×2), what are some good uses of maritime vegetation, what are the benefits of eating algae seaweed and kelp, what are the health benefits of eating algae seaweed and kelp, what are the health benefits of seaweed, what are the positive uses of marine vegetation as a food and a medical treatment, what is the benefit of eating algae seaweed and kelp.

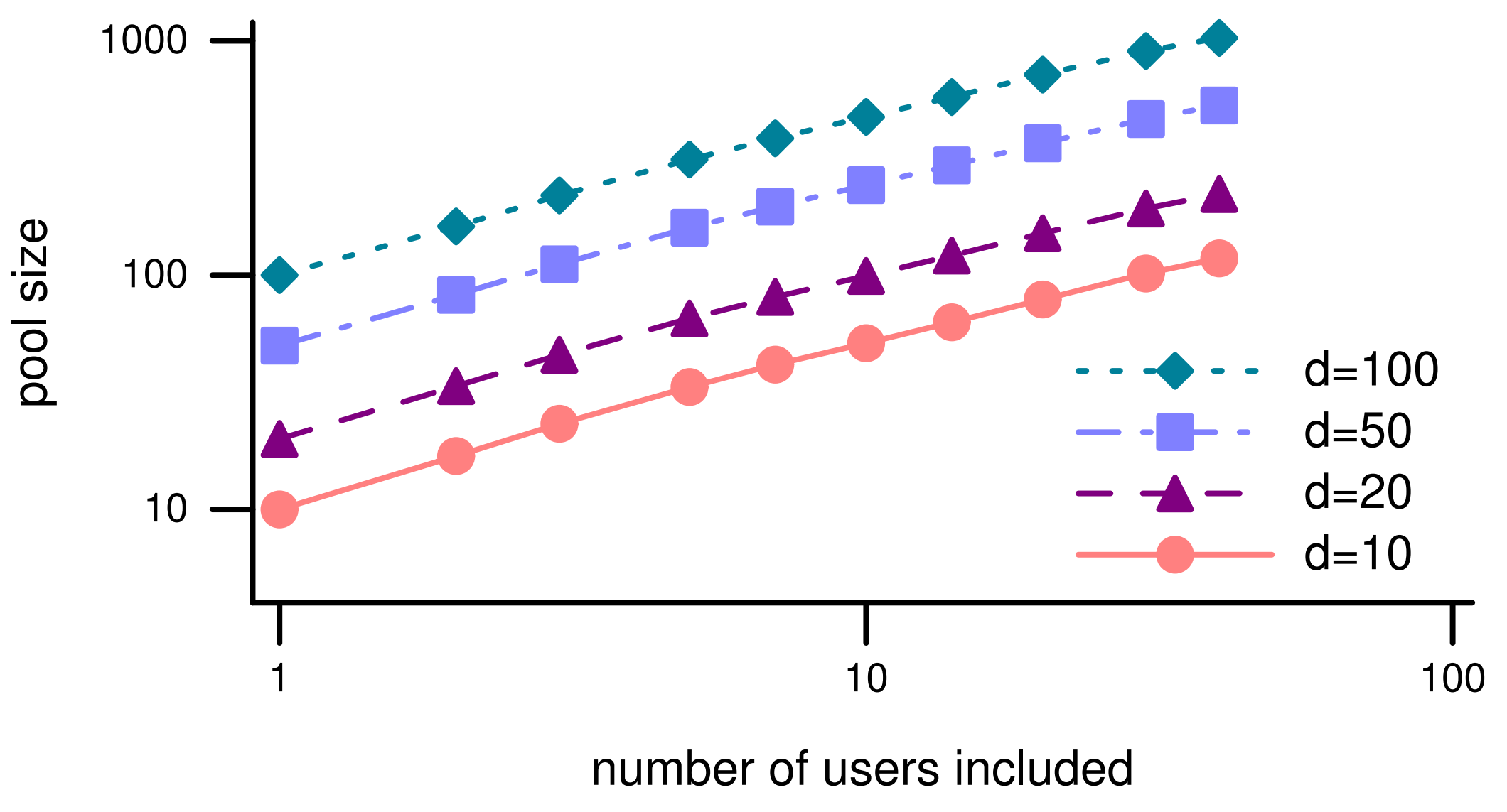
The 7,969 user queries that remained after data cleansing were executed using Indri Okapi BM25 and SDM to create *runs*. Those runs were then compared with the runs submitted by the original TREC contributors.

4. Growth in Pool Size

Average per-topic pool size v as *systems* are added in run-name order, and as *users* are added in CF-ID order, R03, with four different pooling depths d:



$v \approx dn^{0.5}$



$v \approx dn^{0.7}$

5. Overlapping Pools

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 $d = 85$.

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

6. Summary

User-based variations are a significant factor, and need to be designed in at the time a test collection is being constructed.

Starting with information need statements (aka TREC topics) is a sound approach, but current pools do not cover query diversity.

7. Reference

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. SIGIR*, 2015, dx.doi.org/10.1145/2766462.2767728.

Funding

This work was supported by the Australian Research Council's *Discovery Projects* scheme (DP110101743 and DP140102655).