The Wisdom of the Rankers

A Cost-Effective Method for Building Pooled Test Collections without Participant Systems

David Otero, Javier Parapar and Álvaro Barreiro

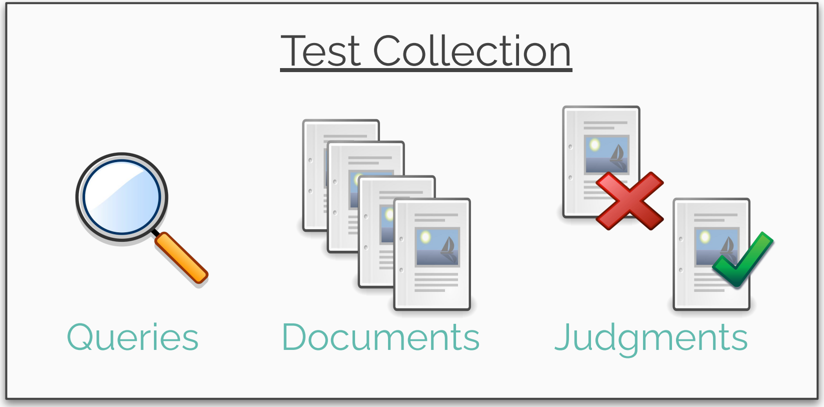
UNIVERSIDADE DA CORUÑA

citic

SAC 2021 36th ACM/SIGAPP Symposium On Applied Computing

Introduction

IR system evaluation usually relies on **test collections**



Test Collection

Queries

Documents

Judgments

- Cranfield collections: every document judged.
- TREC collections: judgments are obtained by **pooling** over top of participant results.
- Pooling provides some warranties over the **validity** of the judgments.
- However, in some situation we may not have participants available.

PROPOSAL

Proposal

Simulate participant systems by combining:

- query variants
- retrieval models

- Rankers:
  - BM25, VSM, DFR, etc: 72 different retrieval models in total.
- Query variants:
  - Manually curated.
  - Automatically generated:
    - Expanding the title query using terms ranked by IDF of the topic's *description* + *narrative*.

Simulated Runs

We generated 4 different sets of runs (using the 72 different retrieval models) that serve as the input to the pooling strategies:

- "*Title*": Use the topic's title against the 72 retrieval models.
  - *"Poliomyelitis and Post-Polio" (Topic's 302)*
- "*Title + description*": Use the topic's description as query.
  - *"Is the disease of Poliomyelitis (polio) under control in the world?"*
- "*Title + manual*": employ 8 variants per topic as the input to the retrieval.
  - *"polio incidence", "polio prevalence", "polio outbreaks", etc.*
- "*Title + automatic*": same as before, but with automatically generated variants.
  - *"Poliomyelitis and Post-Polio outbreaks", "Poliomyelitis and Post-Polio disease", etc.*

Pooling strategies

Employ intelligent pooling strategies to make the most of the assessors' work

Adjudicating methods:

- **DocID**: classical approach used in TREC workshops.
- **MTF**: dynamic method which has been demonstrated as a robust approach.
- **DocPoolFreq**: simple static proposal that ranks docs by counting how many times they appear in the pool.

EVALUATION

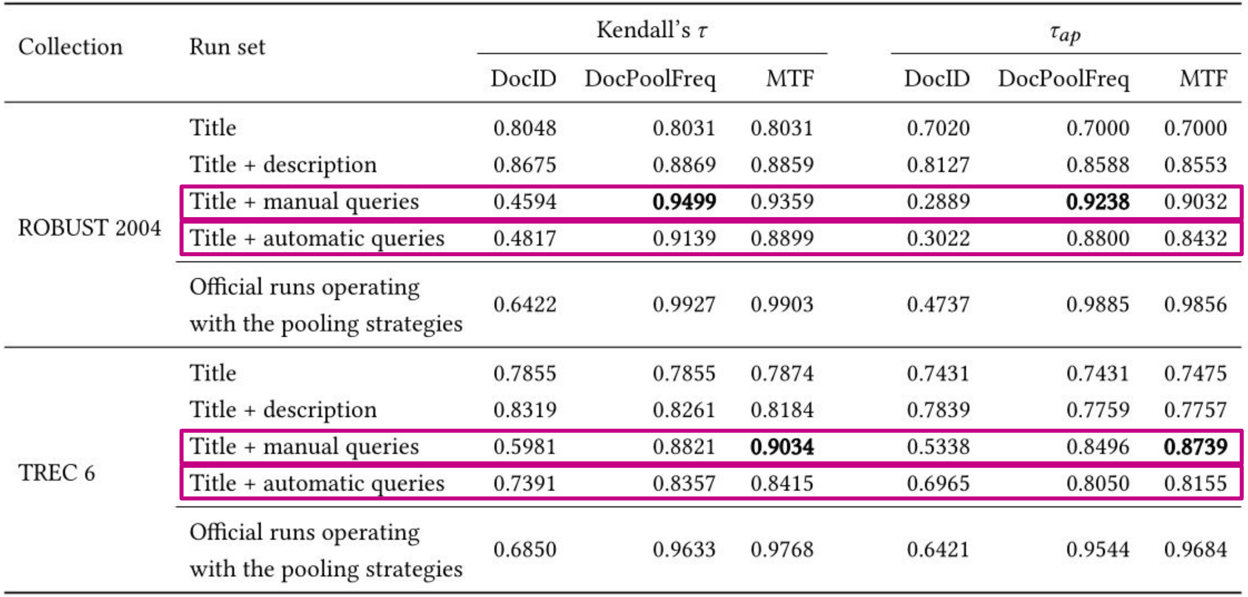
Evaluation: **Reusability**

Reusability: a collection is reusable if it fairly evaluates runs that did not contribute to the building of the collection.

- Ranking correlation between the official ranking of runs and the ranking obtained with our qrels: **none of the ranked runs (the official ones) participated in the building process!**

Results: **reusability**

We obtain **strong correlations** when simulating the participant systems and employing a well-performing pooling strategy



Collection	Run set	Kendall's τ			τ_{ap}		
		DocID	DocPoolFreq	MTF	DocID	DocPoolFreq	MTF
ROBUST 2004	Title	0.8048	0.8031	0.8031	0.7020	0.7000	0.7000
	Title + description	0.8675	0.8869	0.8859	0.8127	0.8588	0.8553
	Title + manual queries	0.4594	**0.9499**	0.9359	0.2889	**0.9238**	0.9032
	Title + automatic queries	0.4817	0.9139	0.8899	0.3022	0.8800	0.8432
	Official runs operating with the pooling strategies	0.6422	0.9927	0.9903	0.4737	0.9885	0.9856
TREC 6	Title	0.7855	0.7855	0.7874	0.7431	0.7431	0.7475
	Title + description	0.8319	0.8261	0.8184	0.7839	0.7759	0.7757
	Title + manual queries	0.5981	0.8821	**0.9034**	0.5338	0.8496	**0.8739**
	Title + automatic queries	0.7391	0.8357	0.8415	0.6965	0.8050	0.8155
	Official runs operating with the pooling strategies	0.6850	0.9633	0.9768	0.6421	0.9544	0.9684

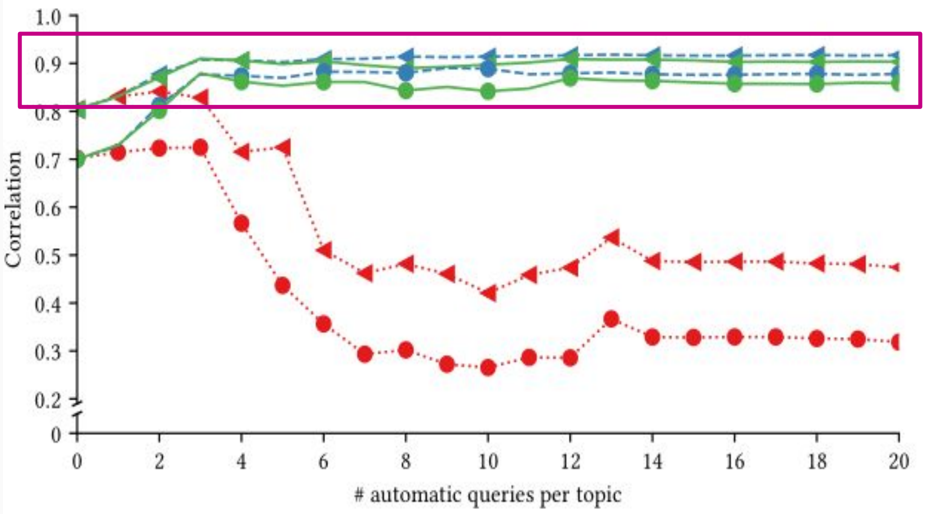
UNIVERSIDADE DA CORUÑA

Evaluation: **Reusability of automatic variants**

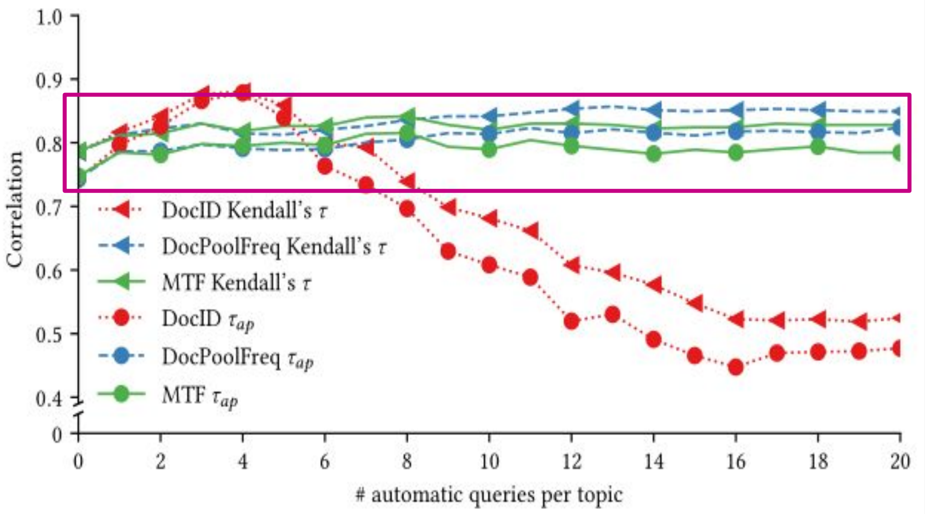
Reusability: a collection is reusable if it fairly evaluates runs that did not contributed to the building of the collection.

- Same evaluation as before, but trying **various number of variants per topic.**

Results: **automatic variants**



Robust04

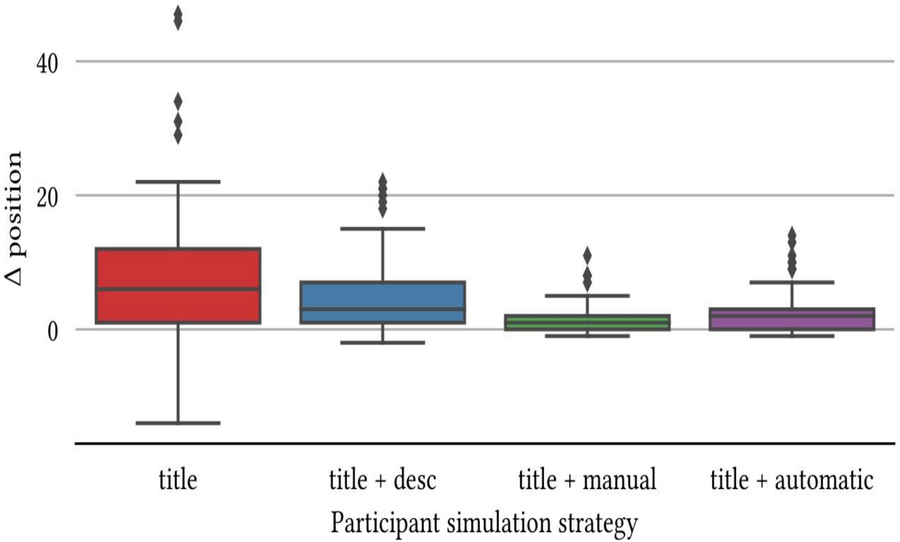


TREC6

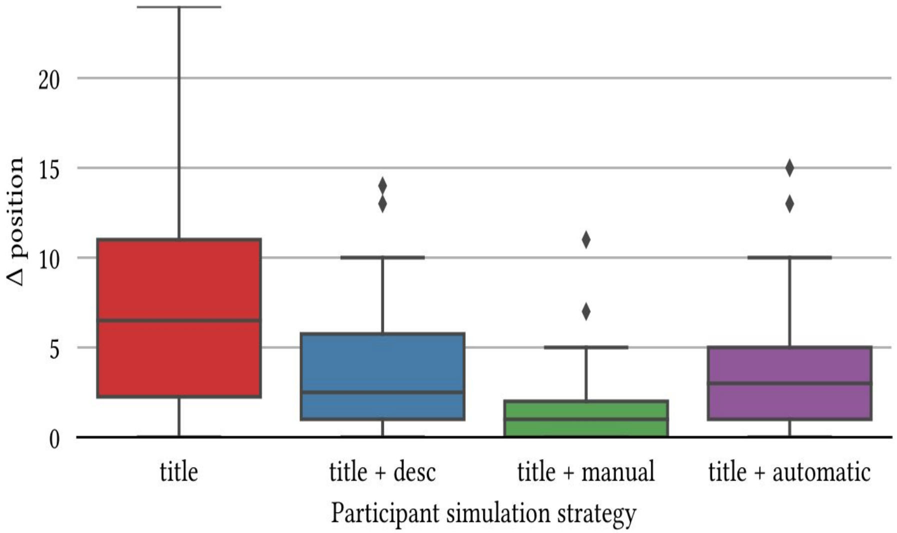
Evaluation: **Fairness**

- **Fairness**: a collection is fair if it fairly evaluates run that contributed to the building of the collection.
    - We compute the change in the ranking position of each official TREC run when ranking it with and without including it in the simulated runs sets.

Results: **fairness**



Robust04



TREC6

Thank You!

David Otero, Javier Parapar and Álvaro Barreiro

@davidoterof | @jparapar | @AlvaroBarreiroG

Conclusions

- Developed a new methodology to obtain relevance assessments where gathering participant results is not possible.
  - We got the best results with the manual variants.
  - We showed that automatically generated query variants are also a good alternative
- Proposed a new static pooling method that performs similarly to MTF.

Future Work

- Study better approaches of variants generation.
- Employ other types of rankers: relevance models, neural models.
- Study other pooling strategies: Bayesian Bandits, Hedge.

@davidoterof

@jparapar

@AlvaroBarreiroG

¡Gracias!

Thank you!

Photography by Jose Luis Cernadas Iglesias

UNIVERSIDADE DA CORUÑA