





# Building High-Quality Datasets for Information Retrieval Evaluation at a Reduced Cost <sup>†</sup>

David Otero \* , Daniel Valcarce , Javier Parapar  and Álvaro Barreiro 

Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicaci3ns (CITIC), Universidade da Coru3a, 15071 A Coru3a, Spain

\* Correspondence: david.otero.freijeiro@udc.es; Tel.: +34-881-01-1276

† Presented at II XoveTIC Congress, A Coru3a, Spain, 5–6 September 2019.

Published: 1 August 2019



**Abstract:** Information Retrieval is not any more exclusively about document ranking. Continuously new tasks are proposed on this and sibling fields. With this proliferation of tasks, it becomes crucial to have a cheap way of constructing test collections to evaluate the new developments. Building test collections is time and resource consuming: it requires time to obtain the documents, to define the user needs and it requires the assessors to judge a lot of documents. To reduce the latest, pooling strategies aim to decrease the assessment effort by presenting to the assessors a sample of documents in the corpus with the maximum number of relevant documents in it. In this paper, we propose the preliminary design of different techniques to easily and cheaply build high-quality test collections without the need of having participants systems.

**Keywords:** information retrieval; evaluation; datasets; cost

## 1. Introduction

In Information Retrieval, test collections are the most widespread technique to evaluate the effectiveness of new developments [1]. These collections are formed by the document set, the information needs (topics) and the human judgments [2]. They are complex to construct because the need of human work to obtain the judgments [3,4]. Datasets of general purpose like TREC (<https://trec.nist.gov>), NTCIR (<http://research.nii.ac.jp/ntcir>) and CLEF (<http://www.clef-initiative.eu>) are useful but sometimes research teams need to build their own collections within a specific task [5].

Pooling methods allow building larger datasets with less effort [6]. When using a pooling approach, only a subset—the pool—of the whole document set is assessed for relevance. The pool is built by taking the union of the top k documents retrieved by each participant system, the runs. In TREC competitions these pools are built using the runs sent by the competition participants, who execute their algorithms on the original dataset and send back their results [2]. Historically, TREC applied the most basic pooling approach (DocID) [2], but recent publications [7,8] have shown that is possible to reduce the assessor’s work without harming the quality of the obtained dataset. In particular, in TREC Common Core Track [9] NIST applied these techniques for the first time. The drawback of these techniques is that they are tied to having participant systems, condition that is not always met.

In some cases it may be necessary to obtain collections prior to the competition. Therefore, in these cases, it is not possible to use approaches where the participants are needed, such as CLEF eRisk competition (<http://erisk.irlab.org>) [10–12], where training data is released.

We propose a method to build the pool before having participant systems. Here the role of the runs will be played by different query variants and out of the box retrieval strategies. The top k documents from the runs produced by multiple combinations of query variants and retrieval strategies are used to build the pool.

## 2. Experiments

We made a series of experiments to preliminary compare the effectiveness between different pooling approaches. In particular, we want to test if the use of query variants is adequate.

### 2.1. Systems and Query Variants

We use four different retrieval models: BM25, TF-IDF, LM Jelinek-Mercer and LM Dirichlet. We want to test the effectiveness of this approach having only a few different systems.

To combine with the described models, we build a series of query variants from the original query. With the model  $\times$  query variant combinations, we can obtain larger pools, ideally having more relevant documents in them. To build a query variant we combine the original query with one of the five terms from the topic description with the highest IDF, i.e., the more specific terms.

Combining these variants along with the systems, we end obtaining a number of different runs equal to  $no. systems(4) \times no. variants(5) = 20$ .

### 2.2. Pooling Algorithms

To perform these experiments we use two pooling algorithms. The first one is the traditional pooling strategy used in TREC competitions [2], i.e., DocID. The second one, DocPoolFreq, is a simple adaptation of the former, where we order the documents by the number of times they appear in the pool and if they tie, by DocID. This is based on the intuition that if a document appears on more systems is it going to be more relevant than other that appears less in all the systems, which is part of many complex pooling algorithms [13].

## 3. Results

We performed these experiments using the TREC5 dataset (disks 2 & 4, topics 251-300). The results can be observed in Figure 1. When it comes to finding more relevant documents we can observe that the approaches that use the query variants outperform the other two. This is because when using the variants we have more systems, which results in having more relevant documents.

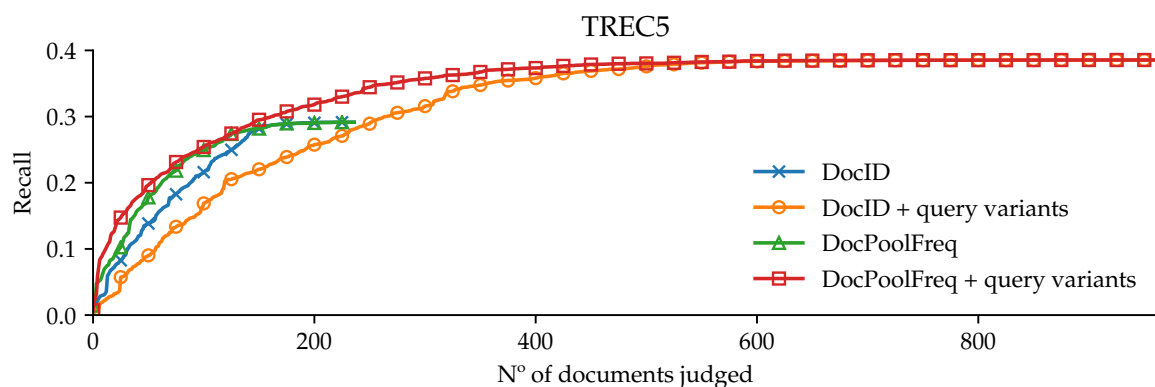


Figure 1. Comparison between pooling strategies.

We also can observe that DocPoolFreq outperforms DocID as it finds relevant documents earlier in the process. This confirms that with only four models and query variants it is possible to obtain the 40% of the relevant documents found in TREC 5 where 61 systems were used to build the pool.

## 4. Discussion

Results show that our research direction is promising. We also open a line of investigation which is to compare the quality of the datasets built with a participants-based approach with techniques that do not need the participant system, like the presented in this paper.

**Funding:** This work was supported by projects RTI2018-093336-B-C22 (MCIU & ERDF) and GPC ED431B 2019/03 (Xunta de Galicia & ERDF) and accreditation ED431G/01 (Xunta de Galicia & ERDF).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Sanderson, M. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends® Inf. Retr.* **2010**, *4*, 247–375.
2. Voorhees, E.M.; Harman, D.K. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*; The MIT Press: Cambridge, MA, USA, 2005.
3. Kanoulas, E. Building Reliable Test and Training Collections in Information Retrieval. Ph.D. Thesis, Northeastern University, Boston, MA, USA, 2009.
4. Losada, D.E.; Parapar, J.; Barreiro, A. Cost-effective Construction of Information Retrieval Test Collections. In Proceedings of the 5th Spanish Conference on Information Retrieval, Zaragoza, Spain, 26–27 June 2018; ACM: New York, NY, USA, 2018; pp. 12:1–12:2.
5. Losada, D.E.; Crestani, F. A Test Collection for Research on Depression and Language Use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, 5–8 September 2016*; Springer: Berlin, Germany, 2016; pp. 28–39.
6. Kuriyama, K.; Kando, N.; Nozue, T.; Eguchi, K. Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop. *Inf. Retr.* **2002**, *5*, 41–59.
7. Losada, D.E.; Parapar, J.; Barreiro, Á. Feeling Lucky?: Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, 4–8 April 2016; ACM: New York, NY, USA, 2016; pp. 1027–1034.
8. Losada, D.E.; Parapar, J.; Barreiro, A. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Inf. Process. Manag.* **2017**, *53*, 1005–1025.
9. Allan, J.; Harman, D.; Kanoulas, E.; Li, D.; Gysel, C.V.; Voorhees, E.M. TREC 2017 Common Core Track Overview. In Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, MD, USA, 15–17 November 2017.
10. Losada, D.E.; Crestani, F.; Parapar, J. eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. Proceedings of 18th Conference and Labs of the Evaluation Forum; Springer: Dublin, Ireland, 2017; CLEF '17, pp. 346–360.
11. Losada, D.E.; Crestani, F.; Parapar, J. Overview of eRisk: Early Risk Prediction on the Internet. Proceedings of 19th Conference and Labs of the Evaluation Forum; Springer: Avignon, France, 2018; CLEF '18, pp. 343–361.
12. Losada, D.E.; Crestani, F.; Parapar, J. Early Detection of Risks on the Internet: An Exploratory Campaign. In Proceedings of the 41st European Conference on Information Retrieval, Cologne, Germany, 14–18 April 2019; pp. 259–266.
13. Losada, D.E.; Parapar, J.; Barreiro, A. A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Inf. Fusion* **2018**, *39*, 56–71.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).