Exploiting Pooling Methods for Building Datasets for Novel Tasks

David Otero[0000−0003−1139−0449]

Information Retrieval Lab
Department of Computer Science
University of A Coruña, Spain
david.otero.freijeiro@udc.es

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 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. The quality of these collections is also crucial, as the value of any evaluation depends on it. In this article, we propose the design of a system for building test collections easily and cheaply by implementing state-of-the-art pooling strategies and simulating competition participants with different retrieval models and query variants. We aim to achieve flexibility in terms of adding new retrieval models and pooling strategies to the system. We want the platform also to be useful to evaluate the obtained collections.

Keywords: Information retrieval · Test collections · Pooling.

1 Introduction

In Information Retrieval, under the Cranfield paradigm, test collections are the most widely used method for evaluating the effectiveness of new systems [15]. These test collections consist of a set of documents, the information needs (topics), and the relevance judgments indicating which documents are relevant to those topics [15]. Collections play a vital role in the process of providing measures to compare the effectiveness of different retrieval models and techniques [14]. However, they are complex and expensive to construct [4, 12]. Some collections of general purpose, such as the ones developed in TREC[1], NTCIR[2]

[1] https://trec.nist.gov

[2] http://research.nii.ac.jp/ntcir

and CLEF[3], are very useful resources for the evaluation of established tasks, but sometimes research teams need to build their own test collection within a specific domain [6].

When building new collections, it is essential to consider their quality. This aspect is crucial, as they are going to be used to evaluate new developments, and the value of this evaluation depends on it. One common problem is to have biased relevance judgments that unfairly rank some models, as Buckley et al. did observe in TREC AQUAINT 2005 Task [2], or to produce non-discriminative results among systems [13]. Because of this, it is important to have a way of evaluating the collections built.

Nowadays, with the huge growth in the number of novel tasks, it would be convenient to have a cheap way of building the evaluation datasets. When creating an evaluation collection, the most straightforward approach to obtain the relevance judgments is to judge the documents as they are retrieved from the data source. This is a very expensive process because it requires a lot of time from the assessors, as typically they judge many documents that end up not being relevant. This process can be alleviated by using pooling techniques.

Pooling is a well-known approach to extract a sample of documents from the entire document set [15]. Using this technique we avoid judging the entire corpus. When using pooling methods we want to obtain the most complete and unbiased set of relevant documents judged [2]. In community evaluation workshops like TREC, pooling is commonly done over the systems sent by the participants, who run their algorithms on the original dataset and send back their results. [15].

In this article, we present the design of a platform to build test collections. With this platform, we aim to tackle three problems: first, to have an easy and cheap way of building the datasets by reducing the assessor's work; second, to build the most complete and the most unbiased collections that are effective to measure and compare the effectiveness of different systems; finally, we focus also on evaluation as we want the platform to be useful to compare different combinations of retrieval models and pooling strategies to reduce the most of the assessor's work and to evaluate the quality of the obtained collections.

2 Background

System evaluation has been a cornerstone in advance of IR. Building test collections for evaluation is expensive, as it requires the work of human assessors to produce relevance judgments. Pooling strategies aim to reduce this cost, as they allow to build test collections much larger than with *complete* judgments [5]. Pooling allows researches to assume completeness over the judgments with a reasonable degree of certainty. The assessor's work is more profitable when they mark a document as relevant. The documents that are not in the pool are considered being non-relevant. On the other hand, for getting true complete judgments assessors would judge the relevance of every document in the collection. If there

[3] http://www.clef-initiative.eu

are many information needs (queries), they would have to assess the relevance the whole set of documents with respect to every query.

In pooled collections, only a subset –the pool– of the entire corpus is judged. For each topic, the pool of documents is generally constructed by taking the union of the top k –pool depth– documents retrieved by each participant systems, called runs. When we have enough relevant documents in the pool, we can assume that the rest of the documents are non-relevant. These obtained pools are then assessed for relevance.

When we apply pooling strategies, we want to obtain unbiased pools. Unbiased means that the sample of relevant documents obtained does not favour any of model, avoiding to unfairly rank some models over others. Another crucial factor is that when an assessor is judging the obtained pool of documents, the process in which documents are presented can introduce some type of bias to the collection [1]

Historically in TREC, assessors have judged the entire pool following an arbitrary strategy, i.e., by DocID, but a lot of work and research has been done in creating pooling algorithms that impose an order of evaluation intending to reduce the assessment effort without harming the quality of the collection. In particular, in TREC Common Core Track 2017 [1], NIST applied for the first time a pooling algorithm based on Bayesian Bandits [10, 11] which has been demonstrated as an effective and unbiased pooling algorithm which improves the state-of-the-art models.

3 Proposal

In this paper, we present the design of a system for experimenting with the creation of test collections. The main goal of the platform is to address the problem of building test collections for novel tasks at affordable cost.

The main contribution of this platform is that, instead of building the pools with a runs-based approach, we build these systems by combining different query variants and retrieval models. This free us of the need to wait for the participants results. This is very convenient, for example, in competitions where the organizers have to release training data to the participants.

The functionality of our system can be seen from two perspectives: one from the system manager, whose function is to define the user information needs, that are manually created but in a future this process can be automated, and select the retrieval models and the pooling strategies; the second one from an assessor, whose work is to judge the relevance of the documents presented to him.

In Figure 1, we can see an overview of the workflow of the platform: the two roles of the system, the system manager –the competition organizer– and the assessor, along with their tasks.

First of all, the platform allows the manager to create different jobs, each one to produce a different collection with its corresponding information needs and relevance assessments. Different types of collections can be built: for example, a multi-topic dataset in which the manager defines one information need per topic;

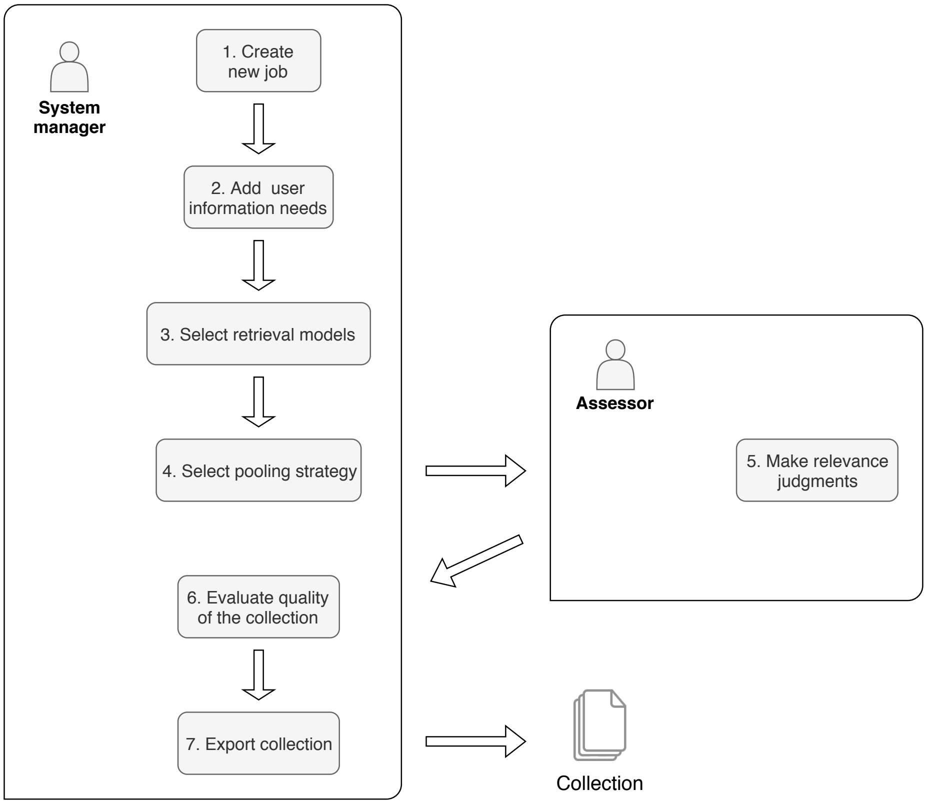
System manager
1. Create new job
2. Add user information needs
3. Select retrieval models
4. Select pooling strategy
6. Evaluate quality of the collection
7. Export collection
Assessor
5. Make relevance judgments
Collection

Fig. 1. Platform's workflow overview.

another example is a classification style dataset, in which the manager defines the criteria for the positive cases of each class.

There are two options to obtain the set of documents: the system can use an off-line static collection or can use an API to retrieve documents from an external data source. At this initial stage, we have developed the components to consume documents from the Reddit API. We aim to expand the platform to more data sources soon. We also aim to make the platform flexible to allow the manager to choose among both offline data and different APIs freely.

In TREC-like competitions, each participant sends the results of one or more systems. These results –the runs– are used to build the pool with the top k documents from each system. We propose to build the pool before having runs for participant systems. Here the role of the runs will be played by different query variants and retrieval strategies that the manager can choose to be associated to the job. The top k documents from the runs produced by multiple combinations of query variant and retrieval strategies are used to build the pool.

Our system will allow the manager to select among different state-of-the-art pooling strategies to present the documents to the assessors, such as MTF

[3] and Multi-armed bandits [10, 11]. The function of the assessors is to judge the documents that are presented to them to build the set of judgments of the dataset. Finally, with the documents retrieved, the topics file and the judgments made by the assessors, the platform allows exporting the final collection.

This platform is designed in such a way that is easy to implement and add new retrieval algorithms as well as new pooling strategies. The platform will also be used to analyse the obtained collections. The system will allow the analysis of the different desired properties for a fair evaluation of systems. We want to analyse the combinations of different simulated participants and different pooling strategies in terms of relevant document found at a given budget and the quality of those judgments. The main goal is to reduce the needed time to build the collections drastically. This is achieved by reducing the time that assessor wastes judging non-relevant documents and by allowing faster retrieval of the documents.

3.1 Pilot Task: CLEF eRisk

CLEF eRisk[4] is an initiative organized with the objective of evaluating the effectiveness of methodologies and metrics for the early detection of risks on the Internet, especially those related to health, such as depression, anorexia or self-inflicted harm. For this purpose, collections of texts written by users on social networks are released annually. The lab is mainly oriented to assist advisors who perform diagnoses on users of social networks, as well as to evaluate the effectiveness of different models when building new collections.

Previous tasks have focused on the detection of depression (2017[5]) [7], as well as the detection of anorexia and depression (2018[6]) [8]. The task of 2019 is about anorexia, depression and self-inflicted harm [9]. This lab will serve as pilot task for our systems. We plan to use the platform to build the collections that will be used in the competition in 2020.

4 Conclusions and Future Work

Building cheap and good test collections is crucial for evaluation. We have seen that obtaining the human judgments of these collections is a time and resource consuming task. There are another risks associated with this task: we may end up building collections that have some bias [2] or with incomplete judgments.

In this paper, we have presented the design of an approach whose aim is to tackle those aspects. Our main goal was to have a cheap way of building these datasets by making the most of the assessor's work. We had to leverage that objective with the build of high-quality collections that are *complete* in terms of judgments and, at the same time, unbiased. It was also essential for us to design a flexible platform to include new models and pooling strategies.

[4] http://erisk.irlab.org

[5] https://early.irlab.org/2017

[6] https://early.irlab.org/2018

This work opens an interesting line of future research, which is to compare the quality and usefulness of collections built from participants runs with collections built with other techniques like our approach.

Acknowledgments. 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). I also would like to thank Daniel Valcarce, Javier Parapar and Álvaro Barreiro for their advise on this work.

References

1. 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, Maryland, USA, November 15-17, 2017. vol. Special Pu. NIST (2017)
2. Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.: Bias and the limits of pooling for large collections. Information Retrieval (2007)
3. Cormack, G.V., Palmer, C.R., Clarke, C.L.A.: Efficient construction of large test collections. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 282–289. SIGIR '98, ACM, New York, NY, USA (1998)
4. Kanoulas, E.: Building Reliable Test and Training Collections in Information Retrieval. Ph.D. thesis, Boston, MA, USA (2009)
5. 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. **5**(1), 41–59 (Jan 2002)
6. Losada, D.E., Crestani, F.: A Test Collection for Research on Depression and Language Use. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 28–39. Springer (2016)
7. Losada, D.E., Crestani, F., Parapar, J.: CLEF 2017 eRisk overview: Early Risk prediction on the internet: Experimental foundations. In: CEUR Workshop Proceedings (2017)
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). In: CEUR Workshop Proceedings (2018)
9. 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. pp. 259–266. ECIR '19, Springer, Cologne, Germany (2019)
10. 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. pp. 1027–1034. SAC '16, ACM, New York, NY, USA (2016)
11. Losada, D.E., Parapar, J., Barreiro, A.: Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. Information Processing and Management (2017)
12. 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. pp. 12:1–12:2. CERI '18, ACM, New York, NY, USA (2018)

13. Lu, X., Moffat, A., Culpepper, J.S.: The Effect of Pooling and Evaluation Depth on IR Metrics. Inf. Retr. **19**(4), 416–445 (Aug 2016)
14. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends® in Information Retrieval (2010)
15. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing). The MIT Press (2005)