Beaver: Efficiently Building Test Collections for Novel Tasks

David Otero
david.otero.freijeiro@udc.es
Information Retrieval Lab
University of A Coruña
A Coruña, Spain

Javier Parapar
javier.parapar@udc.es
Information Retrieval Lab
University of A Coruña
A Coruña, Spain

Álvaro Barreiro
barreiro@udc.es
Information Retrieval Lab
University of A Coruña
A Coruña, Spain

ABSTRACT

Evaluation is a mandatory task for Information Retrieval research. Under the Cranfield paradigm, this evaluation needs test collections. The creation of these is a time and resource-consuming process. At the same time, new tasks and models are continuously appearing. These tasks demand the building of new test collections. Typically, the researchers organize TREC-like competitions for building these evaluation benchmarks. This is very expensive, both for the organizers and for the participants. In this paper, we present a platform to easily and cheaply build datasets for Information Retrieval evaluation without the need of organizing expensive campaigns. In particular, we propose the simulation of participant systems and the use of pooling strategies to make the most of the assessor's work. Our platform is aimed to cover the whole process of building the test collection, from document gathering to judgment creation.

KEYWORDS

Information Retrieval, Test Collections, Pooling

1 INTRODUCTION

Information Retrieval (IR) research is deeply rooted in experimentation and evaluation [8]. Under the Cranfield paradigm, evaluation requires proper infrastructure: methodologies, metrics, and test collections. This paper focuses on building the latter. Collections are formed by documents, topics that describe the user information needs, and relevance judgments, which specify the documents that are relevant to them [10]. Typically, collections are the results of expensive evaluation campaigns such as the TREC tracks. In these forums, one of the most expensive activities is the obtention of relevance judgments, which requires much time and human effort. This is a handicap to teams that aim to build a new dataset of a specific domain or to many new tasks that need to provide training data before the challenge celebration [7]. In these cases, the construction of judgments can not depend on the results of competition participants.

In this paper, we present a platform to ease the construction of test collections without the need to organize evaluation campaigns and thus facilitating the research in IR. We joined in a single platform the process of obtaining the source documents, producing the relevance judgments, and exporting the collection.

2 MOTIVATION

For illustrating our platform, we will use a novel task example: CLEF eRisk[1]. This is a workshop organized each year with the aim of evaluating methodologies for the early detection of risks on the Internet [5]. These risks are especially related to mental diseases like self-harm, anorexia, and depression.

In previous years, eRisk organizers freed test collections formed by texts written by users of Reddit[2]. Those datasets were used by the competition participants to train their models to be evaluated in the test splits. In past editions, the ground truth for datasets (training and test) was built by manually searching relevant posts, that talked about the correspondent topics, to be judged by assessors, resulting in a prolonged and laborious process.

Our platform proposes to ease the process of building the collection by simulating participant systems results and using pooling strategies that make the most of the assessor's work. We will guide this article throughout the process of building a test collection about self-harm. The reader can build his own test collection with this platform, which is live on the following link[3]. Two user roles are defined in the platform. To create new *experiments* and export the collections you can log in as *admin@admin.com* (pass: *admin*). To judge documents from an experiment you can log in as *assessor@assessor.com* (pass: *assessor*).

3 SELF-HARM EXAMPLE

According to the World Health Organization (WHO), self-harm is '*an act with non-fatal outcome, in which an individual deliberately initiates a non-habitual behavior that, without intervention from others, will cause self-harm, or deliberately ingests a substance in excess of the prescribed or generally recognized therapeutic dosage, and which is aimed at realizing changes which the subject desired via the actual or expected physical consequences*'. Inside the self-harm disease, there are various classifications according to the means a person uses to inflict harm on himself (ICD10 X71-X83[4]). We will use these various types of self-harm to guide the document gathering process.

The first step is to obtain the documents from a document source. In this example, our platform uses the Reddit API to download the texts published by users of this social network. Figure 1 shows the architecture and main components of the system. The flexibility of our architecture will allow us to introduce further sources of documents in the future. To obtain the documents, the user may specify different query variants to be used for retrieving posts from Reddit. In this example, it makes sense to use several query variants related to the different classes of self-harm explained above. In particular: 'drown myself', 'cut myself', 'punch myself', 'shot myself', 'burn myself' are good examples. The system will use those

[1] https://erisk.irlab.org

[2] https://www.reddit.com

[3] https://beaver.irlab.org

[4] https://www.icd10data.com/ICD10CM/Codes/V00-Y99/X71-X83

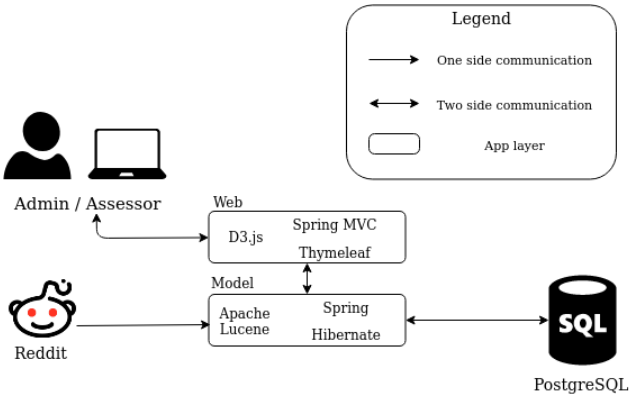

Figure 1: System architecture.

queries to download the whole history of users with posts matching them (in the case of eRisk, the retrieval unit is a user history).

After downloading the documents, our platform creates different document rankings simulating participant systems. The rankings are created by combining the introduced query variants with different retrieval models, such as BM25, Language Models, TF-IDF, or other ones implemented in our platform. In our example, we simulate 20 runs by combining the 5 aforementioned self-harm query variants with 4 ranking models.

Then the admin selects the pooling strategy over the simulated participants' results. This choice will decide the order for presenting the documents to the assessor. Currently, there are two pooling methods implemented: traditional DocID [10] and MoveToFront (MTF) [3], although more strategies are being implemented, such as Hedge [2] and Bayesian Bandits [6]. The CORE Track 2017 used the last one for creating the judgments [1], being the first time that TREC decided to replace traditional DocID. These strategies aim to reduce the assessor's time and effort in the creation of the relevance judgments without harming their quality. In particular, the Max Mean method from [6] was recently demonstrated as the best one in terms of bias [4]. However, the reusability of judgments constructed with these approaches is still an open research issue [9] that we hope this platform will help to investigate.

When the pooling phase starts, the assessor may begin to judge the relevance of the Reddit users. To this aim, he will see all the posts written by each user, divided into various pages. On every page, there are available two buttons to judge the relevance of the user. These buttons are presented in every page because the assessor may not need to see all the posts to establish if a user is relevant or not. For each post, the assessor is presented with its content, with the publication date and with a link referring to the original post in Reddit. The platform does not show any additional data (apart from the user id) about the user to avoid introducing any bias in the assessor's decision. Additionally, the assessor has the option to specify a query to search through the user's publication history to speed up the judging process. Finally, when the assessor completes the work, the test collection can be exported by the administrator along with the judgments.

4 CONCLUSIONS AND FUTURE WORK

Test collections are vital for IR evaluation, but obtaining the relevance judgments is an expensive task. In this article, we have presented a platform to easily and cheaply build test collections by lessening the need for organizing an evaluation campaign. The use of intelligent pooling strategies that heavily reduce the assessor's work makes this process a cheaper task. This system is very suitable to be used by research teams that want to build a collection within a specific domain because they do not need to previously organize a competition to obtain the runs of the participant systems. We plan to use the system as a testbed for evaluating pooling effects in the datasets' quality.

ACKNOWLEDGMENTS

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

REFERENCES

[1] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M Voorhees. 2017. TREC 2017 Common Core Track Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, Ellen M Voorhees and Angela Ellis (Eds.), Vol. Special Pu. National Institute of Standards and Technology NIST.

[2] Javed A. Aslam, Virgiliu Pavlu, and Robert Savell. 2003. A Unified Model for Metasearch, Pooling, and System Evaluation. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 484–491.

[3] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM, New York, NY, USA, 282–289.

[4] Aldo Lipani, David E. Losada, Guido Zuccon, and Mihai Lupu. 2019. Fixed-Cost Pooling Strategies. (2019). to appear in TKDE.

[5] David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019 Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 340–357.

[6] David E. Losada, Javier Parapar, and Alvaro Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing and Management* (2017).

[7] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE '19)*. Association for Computing Machinery, New York, NY, USA, 14–17.

[8] Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems (CLEF '01)*. Springer-Verlag, London, UK, UK, 355–370.

[9] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 407–416.

[10] Ellen M Voorhees and Donna K Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.