

New Algorithms and Methodologies for Building Information Retrieval Collections

David Otero

Universidade da Coruña

Spain

david.otero.freijeiro@udc.es

Abstract

Information retrieval systems play a crucial role in addressing users' information needs by aiding their exploration of vast collections of information. This thesis is framed in a critical information retrieval research aspect: evaluation. In particular, we propose new approaches for creating annotated test collections. Such collections are essential for evaluating retrieval systems' effectiveness in controlled experiments. Reflecting real-world conditions accurately in these test collections is pivotal for progress in the field.

We aim to introduce innovative techniques for efficiently assembling reliable test collections, facilitating broader research and development in information retrieval. The thesis first proposes a new method for building new pooled test collections without requiring costly evaluation campaigns [Otero et al., 2021b]. This approach simplifies and economizes the process of building new benchmarks. Then, we introduce a novel adjudication method for determining which pooled documents warrant human judgment, aiming to reduce the need for extensive expert assessments [Otero et al., 2023a]. This method is both cost-effective and efficient. Additionally, the thesis presents a fresh perspective on evaluating adjudication methods, emphasizing statistical significance, an aspect often overlooked in previous document adjudication research [Otero et al., 2023b]. As a demonstration of the methods explored in this thesis, we applied them to develop a new test collection whose construction process we describe here as an example of the use of reduced-budget methods [Otero et al., 2021a, 2020].

In summary, this thesis integrates established information retrieval knowledge with new methodologies to create annotated collections that are both cost-effective and reliable. This fusion is crucial for advancing the development of more effective retrieval systems.

Awarded by: Universidade da Coruña, A Coruña, Spain **on** 5 April 2024.

Supervised by: Javier Parapar, Álvaro Barreiro.

Available at: <http://hdl.handle.net/2183/36378>.

Selected Publications

David Otero, Javier Parapar, and Álvaro Barreiro. Beaver: Efficiently building test collections for novel tasks. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe*, CIRCLE '20. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2621/CIRCLE_20_23.pdf.

David Otero, Patricia Martin-Rodilla, and Javier Parapar. Building cultural heritage reference collections from social media through pooling strategies: The case of 2020's tensions over race and heritage. *Journal on Computing and Cultural Heritage*, 15(1):1–13, 2021a. ISSN 1556-4673. doi: 10.1145/3477604.

David Otero, Javier Parapar, and Álvaro Barreiro. The wisdom of the rankers: A cost-effective method for building pooled test collections without participant systems. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, pages 672–680, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450381048. doi: 10.1145/3412841.3441947.

David Otero, Javier Parapar, and Álvaro Barreiro. Relevance feedback for building pooled test collections. *Journal of Information Science*, 2023a. doi: 10.1177/01655515231171085.

David Otero, Javier Parapar, and Nicola Ferro. How discriminative are your qrels? how to study the statistical significance of document adjudication methods. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, New York, NY, USA, 2023b. Association for Computing Machinery. doi: 10.1145/3583780.3614916.