Effective Collection Construction for Information Retrieval Evaluation and Optimization

Dan Li
University of Amsterdam
d.li@uva.nl

Abstract

The availability of test collections in Cranfield paradigm has significantly benefited the development of models, methods and tools in information retrieval. Such test collections typically consist of a set of topics, a document collection and a set of relevance assessments. Constructing these test collections requires effort of various perspectives such as topic selection, document selection, relevance assessment, and relevance label aggregation etc. The work in the thesis provides a fundamental way of constructing and utilizing test collections in information retrieval in an effective, efficient and reliable manner. To that end, we have focused on four aspects.

We first study the document selection issue when building test collections. We devise an active sampling method for efficient large-scale evaluation [Li and Kanoulas, 2017]. Different from past sampling-based approaches, we account for the fact that some systems are of higher quality than others, and we design the sampling distribution to over-sample documents from these systems. At the same time, the estimated evaluation measures are unbiased, and assessments can be used to evaluate new, novel systems without introducing any systematic error.

Then a natural further step is determining when to stop the document selection and assessment procedure. This is an important but understudied problem in the construction of test collections. We consider both the gain of identifying relevant documents and the cost of assessing documents as the optimization goals. We handle the problem under the continuous active learning framework by jointly training a ranking model to rank documents, and estimating the total number of relevant documents in the collection using a"greedy" sampling method [Li and Kanoulas, 2020].

The next stage of constructing a test collection is assessing relevance. We study how to denoise relevance assessments by aggregating from multiple crowd annotation sources to obtain high-quality relevance assessments. This helps to boost the quality of relevance assessments acquired in a crowdsourcing manner. We assume a Gaussian process prior on query-document pairs to model their correlation. The proposed model shows good performance in terms of interring true relevance labels. Besides, it allows predicting relevance labels for new tasks that has no crowd annotations, which is a new functionality of CrowdGP. Ablation studies demonstrate that the effectiveness is attributed to the modelling of task correlation based on the axillary information of tasks and the prior relevance information of documents to queries.

After a test collection is constructed, it can be used to either evaluate retrieval systems or train a ranking model. We propose to use it to optimize the configuration of retrieval systems. We use Bayesian optimization approach to model the effect of a δ-step in the configuration space to the effectiveness of the retrieval system, by suggesting to use different similarity functions (covariance functions) for continuous and categorical values, and examine their ability to effectively and efficiently guide the search in the configuration space [Li and Kanoulas, 2018].

Beyond the algorithmic and empirical contributions, work done as part of this thesis also contributed to the research community as the CLEF Technology Assisted Reviews in Empirical Medicine Tracks in 2017, 2018, and 2019 [Kanoulas et al., 2017, 2018, 2019].

Awarded by: University of Amsterdam, Amsterdam, The Netherlands.
Supervised by: Evangelos Kanoulas.
Available at: https://dare.uva.nl/search?identifier=3438a2b6-9271-4f2c-add5-3c811cc48d42.

Selected Publications

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. CLEF 2018 technologically assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. CLEF 2019 technology assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

Dan Li and Evangelos Kanoulas. Active sampling for large-scale information retrieval evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 49–58, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3133015.

Dan Li and Evangelos Kanoulas. Bayesian optimization for optimizing retrieval systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 360–368, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159665.

Dan Li and Evangelos Kanoulas. When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents. *ACM Trans. Inf. Syst.*, 38(4), September 2020. ISSN 1046-8188. doi: 10.1145/3411755.