Learning to Search for Datasets

de Rijke, Maarten

DOI
10.1145/3184558.3191604

Publication date
2018

Document Version
Final published version

Published in
The Web Conference 2018

License
CC BY

Citation for published version (APA):
Learning to Search for Datasets

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

CCS CONCEPTS
• Information systems ➔ Content analysis and feature selection;
Learning to rank; Specialized information retrieval;

KEYWORDS
Dataset search

ACM Reference Format:

Over the years, search engines have developed to return a broad range of retrievable items, from documents to answers, people, locations, and products. Research datasets are increasingly being turned in retrievable items too. This raises a number of interesting challenges. Starting from the user end (“What do users want from datasets?”) to increasing the retrievability of datasets (“What kind of contextual information is available to enrich datasets so as to make the more easily retrieval?”) to optimizing rankers for datasets in the absence of large volumes of interaction data (“How can we train learning to rank datasets algorithms in weakly supervised ways?”).

There are interesting recent developments concerning each of these three areas. For instance, there are a number of recent studies on understanding dataset retrieval practices [3, 7]. We are also getting a better handle on contextual information for dataset search [5, 6]. And advances in supervised and weakly supervised learning to rank [1, 2] and in training neural networks using logged bandit feedback [4] hold great promise for dataset search. In the talk I will survey recent progress in these three areas and identify important open problems.

Acknowledgments
This research was supported by Ahold Delhaize, Amsterdam Data Science, the Bloomberg Research Grant program, the Criteo Faculty Research Award program, Elsevier, the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Google Faculty Research Awards program, the Microsoft Research Ph.D. program, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs CI-14-25, 652.002.001, 612.001.551, 652.001.003, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES