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:
Maarten de Rijke. 2018. Learning to Search for Datasets. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3184558.3191604

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

[1] Zhuyun Dai, Yubin Kim, and Jamie Callan. 2017. Learning to rank resources. In *SIGIR*. 837–840.
[2] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *SIGIR*. 65–74.
[3] Kathleen Gregory, Helena Cousijn, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2018. Understanding data retrieval practices: A social informatics perspective. *arXiv preprint arXiv:1801.04971* (2018).
[4] Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. 2018. Deep Learning with Logged Bandit Feedback. In *ICLR 2018*.
[5] Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017*, Jordi Cabot, Roberto De Virgilio, and Riccardo Torlone (Eds.). Springer International Publishing, 429–436.
[6] Dagmar Kern and Brigitte Mathiak. 2015. Are there any differences in data set retrieval compared to well-known literature retrieval?. In *Research and Advanced Technology for Digital Libraries*, Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla (Eds.). Springer International Publishing, Cham, 197–208.
[7] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The trials and tribulations of working with structured data: A study on information seeking behaviour. In *CHI*. ACM, 1277–1289.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191604