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 into 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 retrievable?”) 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 nr 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