UNIVERSITY OF AMSTERDAM

Search Engines that Learn from Their Users
A.G. Schuth

Summary

More than half the world's population uses web search engines, resulting in over half a billion queries every single day. For many people, web search engines such as Baidu, Bing, Google, and Yandex are among the first resources they go to when a question arises. Moreover, for many search engines have become the most trusted route to information, more so even than traditional media such as newspapers, news websites or news channels on television. What web search engines present people with greatly influences what they believe to be true and consequently it influences their thoughts, opinions, decisions, and the actions they take. With this in mind two things are important, from an information retrieval research perspective. First, it is important to understand how well search engines (rankers) perform and secondly this knowledge should be used to improve them. This thesis is about these two topics: *evaluation of search engines* and *learning search engines*.

In the first part of this thesis we investigate how user interactions with search engines can be used to evaluate search engines. In particular, we introduce a new online evaluation paradigm called *multileaving* that extends upon interleaving. With multileaving, many rankers can be compared at once by combining document lists from these rankers into a single result list and attributing user interactions with this list to the rankers. Then we investigate the relation between A/B testing and interleaved comparison methods. Both studies lead to much higher sensitivity of the evaluation methods, meaning that fewer user interactions are required to arrive at reliable conclusions. This has the important implication that fewer users need to be exposed to the results from possibly inferior search engines.

In the second part of this thesis we turn to online learning to rank. We learn from the evaluation methods introduced and extended upon in the first part. We learn the parameters of base rankers based on user interactions. Then we use the multileaving methods as feedback in our learning method, leading to much faster convergence than existing methods. Again, the important implication is that fewer users need to be exposed to possibly inferior search engines as they adapt more quickly to changes in user preferences.

The last part of this thesis is of a different nature than the earlier two parts. As opposed to the earlier chapters, we no longer study algorithms. Progress in information retrieval research has always been driven by a combination of algorithms, shared resources, and evaluation. In the last part we focus on the latter two. We introduce a new shared resource and a new evaluation paradigm. Firstly, we propose Lerot. Lerot is an online evaluation framework that allows us to simulate users interacting with a search engine. Our implementation has been released as open source software and is currently being used by researchers around the world. Secondly we introduce OpenSearch, a new evaluation paradigm involving real users of real search engines. We describe an implementation of this paradigm that has already been widely adopted by the research community through challenges at CLEF and TREC.