Federated Search with Searsia[*]

Djoerd Hiemstra and Dolf Trieschnigg
Searsia, The Netherlands
{djoerd,dolf}@searsia.com

1 INTRODUCTION

Web search is the single most important application of the web. Without search, the web would be virtually useless. We love Google (and Yahoo, and Bing), but there are some things that worry us about search engines. This paper discusses three of those worries:

The Search Engine Manipulation Effect. First, search engines influence people. In 2014, Robert Epstein and Ronald Robertson [1] of the American Institute for Behavioral Research and Technology in the US conducted the following experiment: They asked people what candidate they would support in the Australian election, Tony Abbott or Julia Gillard? Then they let them search information about the candidates. However, the researchers did not tell the people that they manipulated the search results in favour of Abbott or Gillard. In all cases, people changed their minds towards the manipulated search engine's results much and much more often than you would expect if people change their minds at random. The study shows that people have a great trust in search engines. If results are willingly manipulated, search engines could easily swing an election. Epstein and Robertson call this the *Search Engine Manipulation Effect.*

Privacy. The second thing that worries us: Search engines receive information from their users that is really, really private. People search things about their health, about their sexual preferences, about their financial situation. Most search engines log queries, and they hold on to that information, indefinitely. They also log people's clicks. Search engines might sell this information, and they use it to target people with advertisements. Sadly, there is an incentive for companies to use this information to exploit people's weaknesses: The most vulnerable people in our society, people with bad health or credit scores, are in practice targeted the most [2].

Crawling, crawling, crawling. The third thing that worries us: Search engines need a lot of resources. They constantly crawl the web, downloading pages to check if things changed. Less popular pages, like personal blogs might be downloaded more by web robots that crawl the site, than by people that read the blog posts. But constantly downloading a blog to detect changes is crazy. Blogs often have their own search field, which instantly shows updates when something changes. If only, somehow, we can redirect queries from search engines directly to a blog's search field?

2 FEDERATED SEARCH

Searsia's approach to federated search can ease some of these worries. A Searsia search engine is a federation of search engines, like the European Union (EU) is a federation of countries. Within the EU, each country has its own government and its own laws. Within Searsia, each search engine indexes whatever it likes and returns whatever it likes.

Sites like Wikipedia, Amazon, Stack Overflow and YouTube provide their own search interface. Obviously, they do not need to crawl their site, because they own the data. They instantly know when something changed. Searsia comes with a configuration mechanism that makes it easy to add search engines to the federation. When a Searsia engine receives a query, it forwards the query to a selection of the search engines in the federation. It will try to select those search engines that most likely satisfy the user's need, for instance forwarding a query to YouTube, if it believes the user would like a video. This way, Searsia displays live results from YouTube, instead of results that were crawled a while ago. Each Searsia engine is a search engine too, pretty much like any other search engine, so we might make a federation of those search engines again.

Searsia is open source, and available at: http://searsia.org. Searsia is used amongst others for site search of the University of Twente (https://utwente.nl/search) and the Sheet Music search engine Dr. Sheet Music (https://drsheetmusic.com).

3 CONCLUSION

Let's go back to our three worries: Manipulation, Privacy, and Costs of crawling. What do you gain by installing Searsia? First, Searsia delegates queries to multiple search engines that it does not control. Even if one or two of those manipulate or censor their results, there will be other engines in the federation that show more objective results. Second, all queries go via Searsia, so engines in the federation can no longer track individual users. Of course, now Searsia will get all your queries, but no worries, anyone can run a Searsia server. Third, Searsia does not crawl the web, ever. It will learn over time from each engine in the federation what its contents are. This makes it cheap to set up a Searsia engine, and easy to maintain. You can run Searsia on a cheap server in your own network.

ACKNOWLEDGMENTS

We are grateful to the Vietsch Foundation for supporting our work. We also thank NLNet Foundation and the University of Twente.

REFERENCES

[1] Robert Epstein and Ronald Robertson. 2014. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc. Nat. Acad. Sci. U. S. A.* 112, 33 (2014).

[2] Cathy O'Neil. 2016. *Weapons of Math Destruction: How big data increases inequality and threatens democracy.* Crown, New York.

[*] Presented at the New Year Event of the Dutch Internet Society (ISOC) on 12 Jan. 2017

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

DIR, November 2017, Hilversum, The Netherlands.

© 2017 Copyright held by the owner/author(s).