Web search engine biases.

M2R internship proposal: 1 person, 5 months.

Supervisors :
- Philippe Mulhem, LIG
- Lydie du Bousquet, LIG
- Lorraine Goeuriot, LIG

This work aims to formalize and study the bias among several web search engines on the web. Commercial search results may be biased by some elements (for instance the ads in Google Search as with the ``Basecamp example'' [1]). Such bias may avec have potential impact on citizens, as is has been shown for their vote in [2], and also about the relevance of the documents retrieved [3]
When algorithms are open, it is possible to study extensively the possible biases, but commercial web search engines are black boxes, from which only inputs and outputs are known [4, 5]. The limitations of the works that focus on the stability of web search results, like [5, 6] are that it is not possible to provide a ground truth against which we may study actual biases. An interesting work [3] tried to solve some part of this problem by focusing on one specific topic, US elections, and by characterizing some results of Google search using a classification of web sites [10].
The internship proposed here will then take inspiration from [3] to propose a model that studies biases of web search engines. It will be applied on one topic, (to be defined) from which some referential will have to be defined. Such referential could be other search engines that are supposingly not personalized, like Qwant or other tools that characterize web pages like Curlie [7] or Best of the Web [8] for general purpose topics, or Health on the Net [9] for medical topics.

The expected work comprises three steps:
1. A state of the art about biases in web search, and inter web engines biases;
2. A proposal of a model to evaluate inter search engine biases: the protocol of bias evaluation will be defined, as well as accurate measures to quantify the bias. One novelty of the proposal is that we will integrate the fact that several search engines are used;
3. An experiment that provides preliminary results on a specific topic (to be defined) for several web engines, namely: Google Search, Bing and Qwant.

Bibliographic references:
[1] https://www.cnbc.com/2019/09/04/google-paid-search-ads-shakedown-basecamp-ceo-says.html
[2] Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences, 112(33) :E4512–E4521,2015.
[3] Cameron Lai and Markus Muczak-Roesch, You can's see what you can't see: Experimental evidence for how much relevant information may be missed due to Google Web search personalization, 11th Conference on Social Informatics, Doha, Qatar, Novembre 2019, To be published.
[4] Juhi Kulshrestha and Motahhare Eslami and Johnnatan Messias and Muhammad Bilal Zafar and Saptarshi Ghosh and Krishna P. Gummadi, and Karrie Karahalios.

Search bias quantification: investigating political bias in social media and web search. nformation Retrieval Journal, 2019, vol 22, N. 1, pp. 1573-7659
[5] Lydie du Bousquet, Philippe Mulhem, Sara Lakah: Quelques pas vers l'Honnêteté et l'Explicabilité de moteurs de recherche sur le Web. CORIA 2019.
[6] Hannák A., Sapiezynski P., Khaki A. M., Lazer D., Mislove A., Wilson C. (2017). Measuring personalization of web search. CoRR, vol. abs/1706.05011
[7] Curlie: https://curlie.org/
[8] Best of the Web (BOTW) : https://botw.org/
[9] Health on Net Search: https://www.hon.ch/HONcode/Search/search.html
[10] http://balancestudy.org/whitelist-classifiable.html