FACULTY DESIGN MEDIA INFORMATION
Department of Information

RAT
RESULT ASSESSMENT TOOL

HAW HAMBURG

A Comparison of Source Distribution and Result Overlap in Web Search Engines

Helena Häußler, HAW Hamburg
Nurce Yagci, HAW Hamburg
Sebastian Sünkler, HAW Hamburg
Dirk Lewandowski, University of Duisburg-Essen, HAW Hamburg

Background

With Google dominating the search engine market, it often seems that there are no alternatives at all. However, there are competitors like **Bing** with their own index, operators like **DuckDuckGo** proclaiming privacy, and meta search engines like **MetaGer**. We are interested in examining **whether alternative search engines are alternatives in fact**, i.e., by providing users with different or additional results compared to Google.
In previous studies, measuring overlap of search results served to estimate the size of the web (i.e., Bharat 1998) or comparing indexes and ranking algorithms by different operators (i.e., Spink et al. 2006).

Research Questions

1. Do top results from alternative search engines differ from Google's in regard to the *number of unique sources*?
2. Do top results from alternative search engines differ from Google's in regard to *top sources*?
3. Do top results from alternative search engines differ from Google's in regard to *source concentration*, i.e., are results distributed over more or fewer sources in different search engines?

Methods

The daily Google trends for both Germany and the US were collected daily at 3 am CET from November 10th, 2021, until March 31st, 2022. After removing duplicates, our query sets consist of 1,821 queries for Germany and 2,126 queries for the US (see figure below, 1). For each query, the **top 10 results from all four search engines were collected with help of the Result Assessment Tool (RAT)** (2). This resulted in a total of 109,604 organic results for Germany and 120,711 organic results for US. The python library urllib was used to extract the domain of each search result (3). Due to data cleaning (4), the dataset is reduced to **1,672 queries and 66,880 results for Germany** and **1,865 queries and 74,600 results for the US**. We adapt Gini coefficient (Ortega et al. 2008) to measure source concentration and calculate Jaccard similarity index (González et al. 2008, Puschmann 2019) to measure the similarity between two result sets (5).

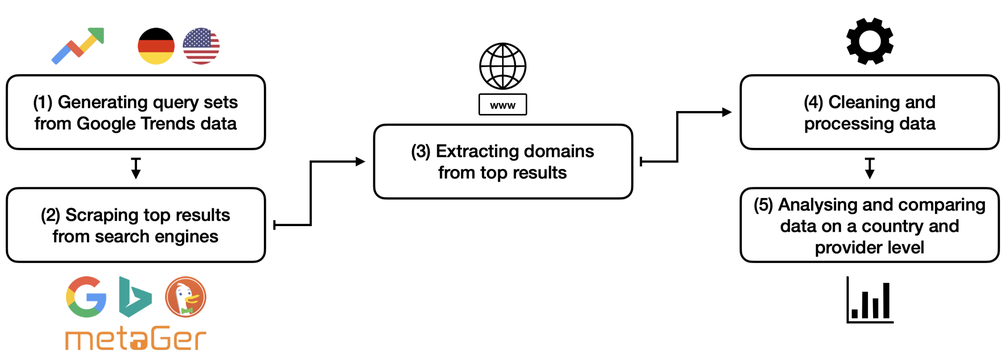

Figure 1. Methodology of the current study

Results

An evaluation of root domains of the most popular sources shows only **a small overlap between Google and the alternative search engines** (RQ1).
Overall, we found an overlap between 23% and 27% between Google and the alternatives in German results, and in the US, the overlap ranges from 24% to 26%. There is a notably higher overlap between the alternative search engines of about 36% to 66% in German results and 57% to 66% in US results (see chart below).
Across all search engines and countries, **the most popular domain was Wikipedia**, followed by sources we classified as **News services** (RQ2).
Interestingly, in German Google results, we do not find results by Instagram and YouTube in contrast to alternative search engines. On the other side, we found Facebook and TikTok results exclusively in Google's US results. It is noteworthy that Fox News is not found in Google's top 50 source, while it is present in Bing, DuckDuckGo, and Metager.
The concentration of sources and source diversity showed a tendency for **only a few root domains** to make up a large share of search results. The Gini Index values of 0.73 and 0.79 in Germany and the United States, respectively, are a clear indicator (RQ3).

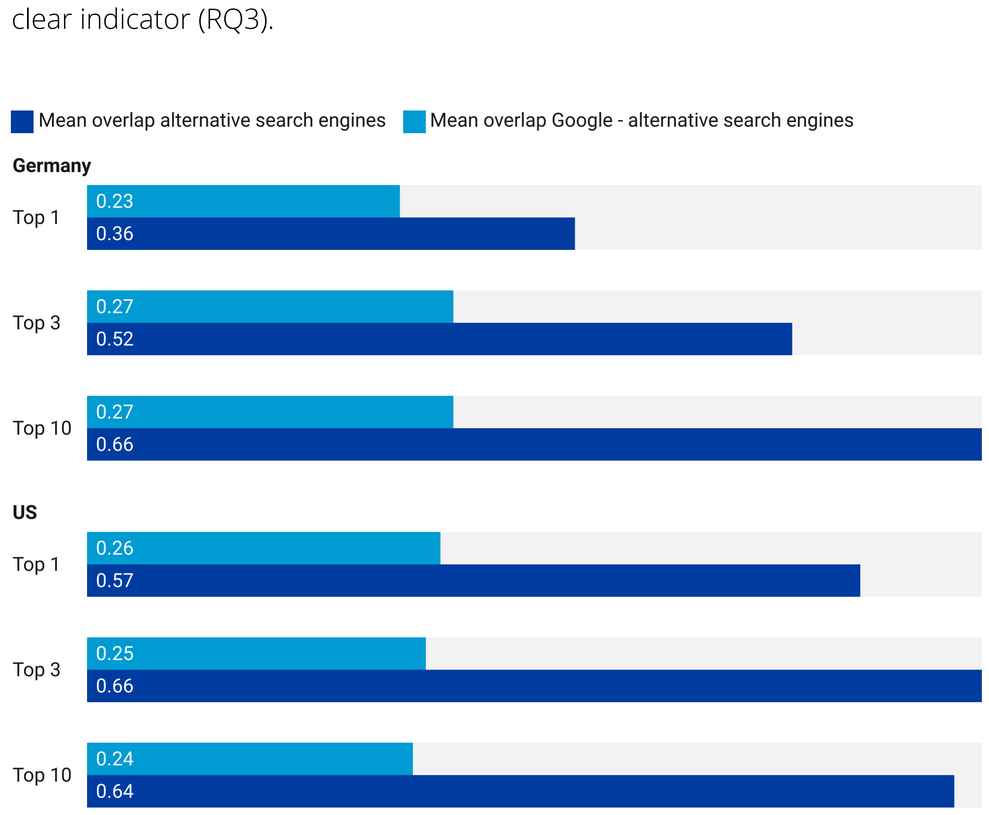

Figure 2. Overlaps between the search engines

Discussion / Conclusion

Interestingly, in the US results, YouTube was in the top 10 most popular domains in all search engines but Google. The same was the case for the top 50 domains in German results. This is unexpected because YouTube is a subsidiary of Google. However, this finding may be explained by the fact that we only collected organic search results, and Google might be using universal search results to display YouTube results.
Our results show that **using another or more than one search engine leads to seeing more diverse search results**, allowing users to inform themselves more comprehensively. It should be noted that within each search engine's results, the concentration of sources shows that only a few top sources dominate the results, meaning whichever search engine a user chooses to use will shape what sources the information they get to see comes from.

THIS WORK IS A CONTRIBUTION TO THE 85TH ANNUAL MEETING FOR INFORMATION SCIENCE & TECHNOLOGY:
Yagci, N., Sünkler, S., Häußler, H., & Lewandowski, D. (2022). A Comparison of Source Distribution and Result Overlap in Web Search Engines. Proceedings of the Association for Information Science and Technology, 59(1), 346–357.
https://doi.org/10.1002/pra2.758

REFERENCES:
Bharat, K. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems, 30(1–7)*, 379–388.
González, C. G., Bonventi, W., & Rodrigues, A. L. V. (2008). Density of Closed Balls in Real-Valued and Autometrized Boolean Spaces for Clustering Applications. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics):* Vol. 5249 LNAI (pp. 8–22). Springer Verlag.
Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the Inequality of Contributions to Wikipedia. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 304–304.
Puschmann, C. (2019). Beyond the Bubble: Assessing the Diversity of Political Search Results. *Digital Journalism*, *7(6)*, 824–843.
Spink, A., Jansen, B. J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. *Information Processing and Management*, *42(5),* 1379–1391.

Research data is available at: https://osf.io/nt3wv/
This work is funded by the German Research Foundation (DFG – Deutsche Forschungsgemeinschaft; Grant No. 460676551).

CONTACT
Nurce.yagci@haw-hamburg.de

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences