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

Nurce Yagci, Sebastian Sünkler, Helena Häußler, Dirk Lewandowski
Hamburg University of Applied Sciences, Germany

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

85TH ANNUAL MEETING OF THE ASSOCIATION FOR INFORMATION SCIENCE & TECHNOLOGY
OCTOBER 29 - NOVEMBER 1, 2022 | PITTSBURGH, PA

RAT RESULT ASSESSMENT TOOL
SEARCH STUDIES
HAW HAMBURG

AGENDA

- Introduction
- Literature Review
- Research Questions
- Method
- Results
- Discussion and Conclusion

Introduction

INTRODUCTION

- **Why use an alternative search engine, i.e., a search engine other than Google?**
  - Different results, i.e., results from different sources
  - “Second opinion” on a topic

- **Trust in search engines**
  - User trust search engines to provide relevant and useful results (European Commission, 2016; Purcell et al., 2012)
  - Users trust news found via search engines more than any other source, including traditional news outlets or social media sites (Edelman Trust Institute, 2022; Newman et al., 2021)

INTRODUCTION

- **Usage of search engines**
  - Only some users use another search engine in addition to Google (Schultheiß & Lewandowski, 2021)
  - Users predominantly select from the top results shown by the search engine (Lewandowski & Kammerer, 2021)

- **Functionality of search engines**
  - Search engines tend to prefer big websites, causing smaller websites to be hidden from users (Introna & Nissenbaum, 2000)
  - Many seem-to-be search engines are merely search portals displaying results from a partner (e.g. Ecosia, DuckDuckGo)
  - Alternative search engines can come with other benefits (e.g. DuckDuckGo's enhanced privacy policy)

Literature Review

85TH ANNUAL MEETING OF THE ASSOCIATION FOR INFORMATION SCIENCE & TECHNOLOGY, 2022
NURCE YAGCI, SEBASTIAN SÜNKLER, HELENA HÄUßLER, DIRK LEWANDOWSKI

RAT RESULT ASSESSMENT TOOL
SEARCH STUDIES
HAW HAMBURG

LITERATURE REVIEW | SEARCH RESULT OVERLAP

- **Low overlap in the 1990’s and 2000’s**
  - Overlap was initially studied to estimate the size of the web (Bharat, 1998; Ding & Marchionini, 1996)
  - Low overlap indicated that search engines had small but diverse databases
  - Only 1.1% of results overlapped between Ask Jeeves, Google, MSN Search, and Yahoo! (Spink et al., 2006)

- **Overlap increased with time**
  - 22% - 40% overlap between Google - Bing and Google - Yahoo! (Bilal and Ellis, 2011)
  - Overlap between Google - Bing slightly higher when ignoring result positions (Cardoso and Magalhães, 2011)
  - For top 50 Covid-19 results, DuckDuckGo - Yahoo! overlap by 50% and Google - Bing by 10% (Makhortykh et al., 2020)

LITERATURE REVIEW | SOURCE DIVERSITY

- **Small number of popular sources are more commonly the top result**
  - Using queries about insurance providers returns only 10 different sources in the top 1 results (Lewandowski & Sünkler, 2019)
  - Wikipedia and news websites are most popular for debated topics, e.g. climate change (Steiner et al., 2022)

- **Difference in diversity between search engines**
  - Yahoo! returns more diverse results and 10% more top-level domains than Google and LiveSearch (Thelwall, 2008)

Research Questions

RESEARCH QUESTIONS

1. **Do top results from alternative search engines differ from Google's in regard to the number of unique sources?**

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?**

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?**

Method

METHOD

(1) Generating query sets from Google Trends data

1. Google Trends collected daily between 10.11.21 - 31.03.22 ➡ 1,672 German and 1,865 US queries

METHOD

(1) Generating query sets from Google Trends data

(2) Scraping top results from search engines

1. Google Trends collected daily between 10.11.21 - 31.03.22 ➡ 1,672 German and 1,865 US queries

2. Top 10 results from Google, Bing, DuckDuckGo and MetaGer[1] collected ➡ 66,880 German and 74,600 US results

[1]*Metager: German meta-search engine*

METHOD

(1) Generating query sets from Google Trends data

(2) Scraping top results from search engines

(3) Extracting domains from top results

1. Google Trends collected daily between 10.11.21 - 31.03.22 ➡ 1,672 German and 1,865 US queries
2. Top 10 results from Google, Bing, DuckDuckGo and MetaGer[1] collected ➡ 66,880 German and 74,600 US results
3. Domains extracted using regex to match strings and patterns, e.g. coolblog.wordpress.com/blog/hello-world ➡ wordpress.com

[1]*Metager: German meta-search engine*

METHOD

(1) Generating query sets from Google Trends data

(2) Scraping top results from search engines

(3) Extracting domains from top results

(4) Cleaning and processing data

1. Google Trends collected daily between 10.11.21 - 31.03.22 ➡ 1,672 German and 1,865 US queries
2. Top 10 results from Google, Bing, DuckDuckGo and MetaGer[1] collected ➡ 66,880 German and 74,600 US results
3. Domains extracted using regex to match strings and patterns, e.g. coolblog.wordpress.com/blog/hello-world ➡ wordpress.com
4. Alternative domains consolidated, e.g. https://twitter.com and http://www.twitter.com/de/ ➡ twitter.com

[1]*Metager: German meta-search engine*

METHOD

(1) Generating query sets from Google Trends data

(2) Scraping top results from search engines

(3) Extracting domains from top results

(4) Cleaning and processing data

(5) Analysing and comparing data on a country and provider level

1. Google Trends collected daily between 10.11.21 - 31.03.22 ➡ 1,672 German and 1,865 US queries
2. Top 10 results from Google, Bing, DuckDuckGo and MetaGer[1] collected ➡ 66,880 German and 74,600 US results
3. Domains extracted using regex to match strings and patterns, e.g. coolblog.wordpress.com/blog/hello-world ➡ wordpress.com
4. Alternative domains consolidated, e.g. https://twitter.com and http://www.twitter.com/de/ ➡ twitter.com
5. Analysis of domains

[1]*Metager: German meta-search engine*

Results

RESULTS | DOMAIN CATEGORIES

- 8 categories generated by classifying top 50 German and US domains

Category	Example	Top 50 Germany (share)	Top 50 US (share)
Celebrities	variety.com	0.06	0.10
E-Commerce	amazon.com	0.02	0.06
Government website	bundesregierung.de	0	0
Information service	wikipedia.org	0.02	0.02
Movies & Entertainment	imdb.com	0.18	0.12
News service	cnn.com	0.54	0.30
Sports	espn.com	0.14	0.34
Social media	instagram.com	0.04	0.06

Table 1. Classes of domains and their frequency in the top 50 domains

RESULTS | DOMAIN CATEGORIES

- 8 categories generated by classifying top 50 German and US domains
- **54% of German domains are news services and 34% of US domains are sports websites**

Category	Example	Top 50 Germany (share)	Top 50 US (share)
Celebrities	variety.com	0.06	0.10
E-Commerce	amazon.com	0.02	0.06
Government website	bundesregierung.de	0	0
Information service	wikipedia.org	0.02	0.02
Movies & Entertainment	imdb.com	0.18	0.12
News service	cnn.com	**0.54**	0.30
Sports	espn.com	0.14	**0.34**
Social media	instagram.com	0.04	0.06

Table 1. Classes of domains and their frequency in the top 50 domains

RESULTS | DOMAIN VARIETY

- **In the German results, Google has the highest variety of domains in the top position**
- Google US has the highest number of domains in the top 10 results

Position	Google	Bing	DuckDuckGo	Metager
Top 1	609	389	302	372
Top 10	2,841	2,783	2,707	2,693

Table 2. Number of unique domains in top 1 and top 10 results (Germany)

Position	Google	Bing	DuckDuckGo	Metager
Top 1	508	521	493	485
Top 10	4,085	3,602	3,579	3,500

Table 3. Number of unique domains in top 1 and top 10 results (US)

RESULTS | DOMAIN VARIETY

- **In the German results, Google has the highest variety of domains in the top position**
- Google US has the highest number of domains in the top 10 results

Position	Google	Bing	DuckDuckGo	Metager
Top 1	609	389	302	372
Top 10	2,841	2,783	2,707	2,693

Table 2. Number of unique domains in top 1 and top 10 results (Germany)

Position	Google	Bing	DuckDuckGo	Metager
Top 1	508	521	493	485
Top 10	4,085	3,602	3,579	3,500

Table 3. Number of unique domains in top 1 and top 10 results (US)

RESULTS | POPULAR DOMAINS

- **Wikipedia is the most popular domain in all four search engines in both countries**

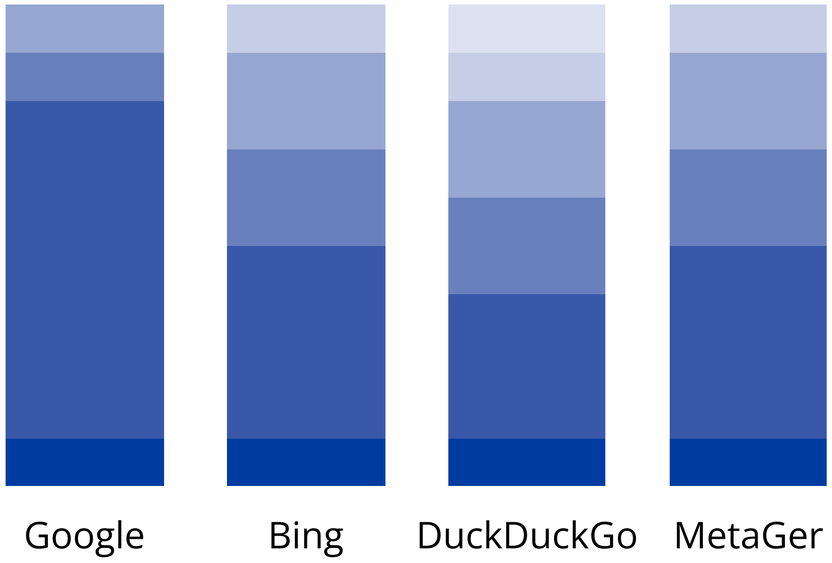

Fig 1. Top 10 domains for by category (Germany)

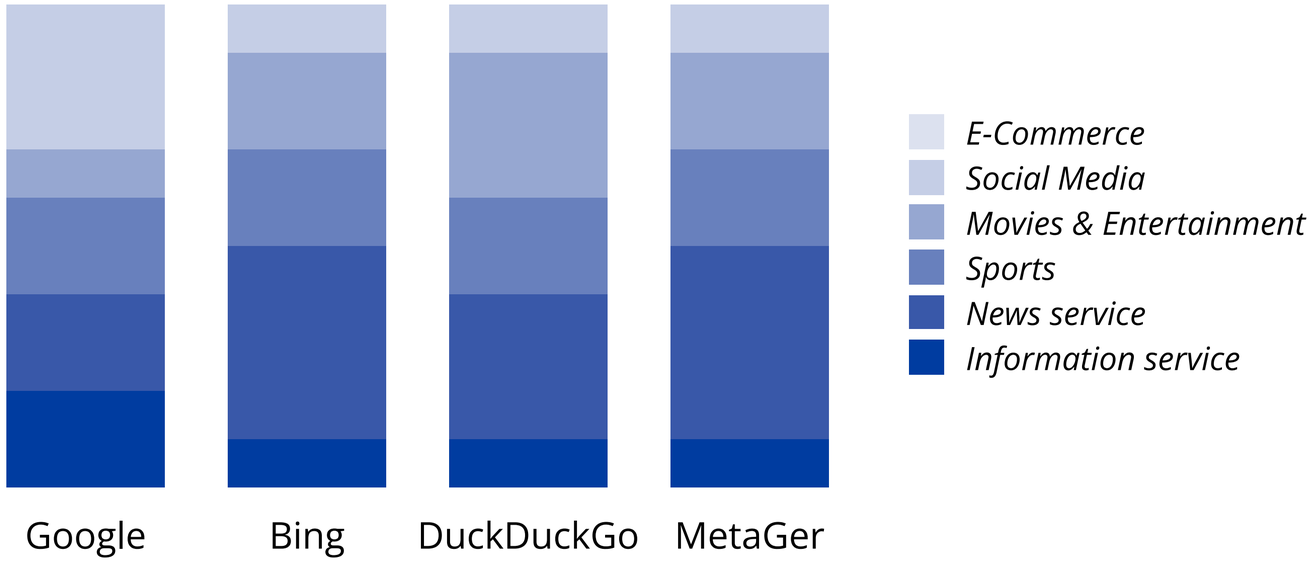

Fig 2. Top 10 domains for by category (US)

RESULTS | POPULAR DOMAINS

- **Wikipedia is the most popular domain in all four search engines in both countries**
- News services make up at least 20% of the top 10 results for all search engines

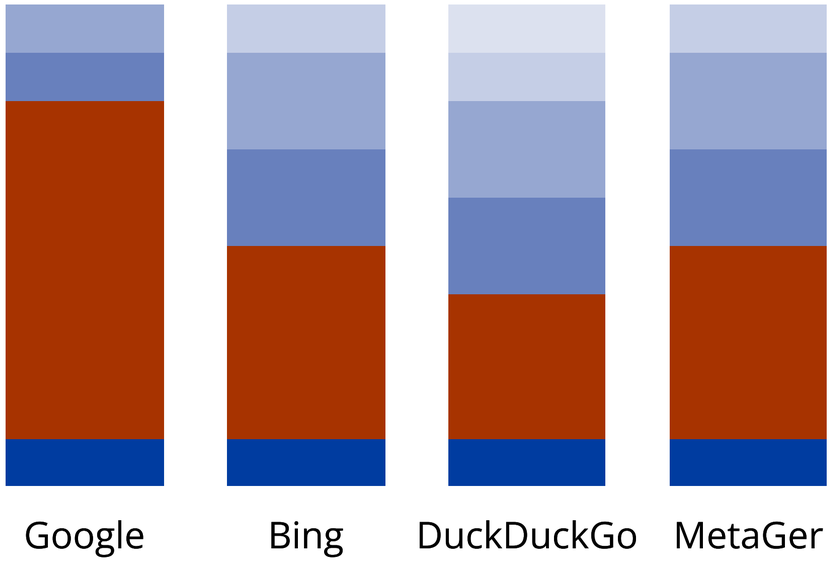

Fig 1. Top 10 domains for by category (Germany)

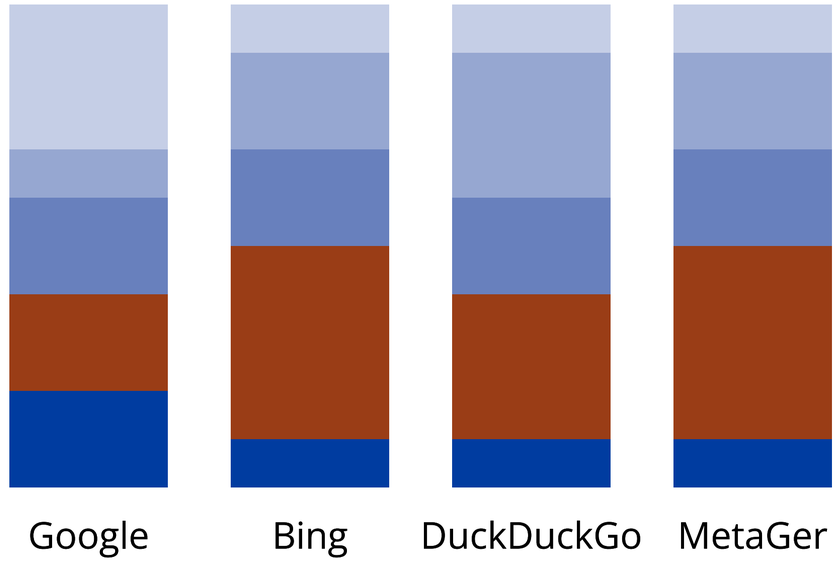

Fig 2. Top 10 domains for by category (US)

RESULTS | POPULAR DOMAINS

- **Wikipedia is the most popular domain in all four search engines in both countries**
- News services make up at least 20% of the top 10 results for all search engines
- 7 of the top 10 results in the German Google results are news services

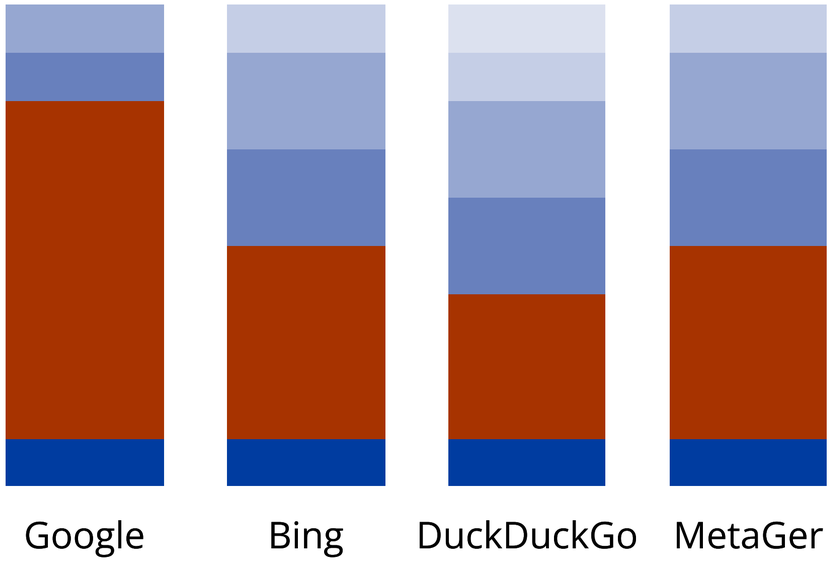

Fig 1. Top 10 domains for by category (Germany)

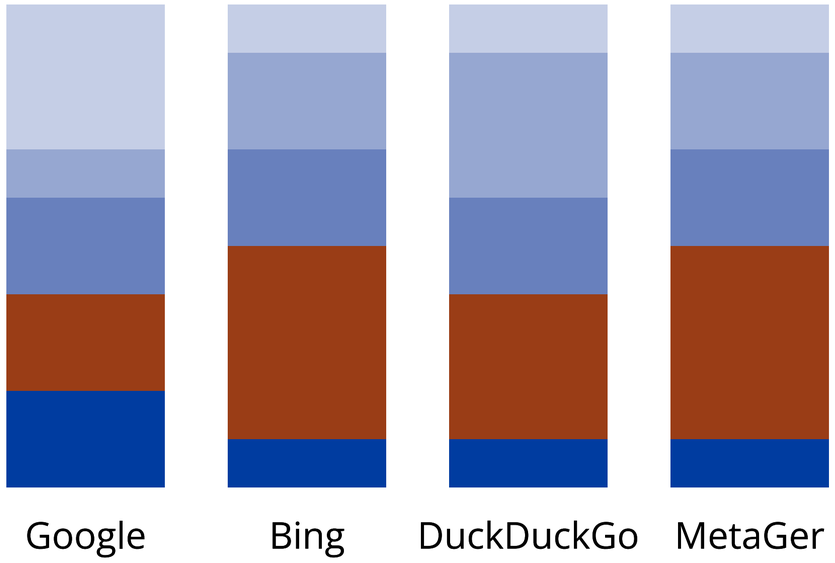

Fig 2. Top 10 domains for by category (US)

RESULTS | EXCLUSIVE DOMAINS

- **Two types of exlusive domains**
  - Domains that only appear in Google results (Google only)
  - Domains that do not appear in Google but do appear in the other three search engines (all but Google)

	Google only	All but Google
Germany	20	13
US	17	13

Table 3. Exclusive domains in the top 50 domains

RESULTS | EXCLUSIVE DOMAINS

- **Two types of exlusive domains**
  - Domains that only appear in Google results (Google only)
  - Domains that do not appear in Google but do appear in the other three search engines (all but Google)

- **YouTube, a subsidiary of Google, is an exclusive domain (all but Google) in the German results**

	Google only	All but Google
Germany	20	13
US	17	13

Table 3. Exclusive domains in the top 50 domains

RESULTS | EXCLUSIVE DOMAINS

- **Two types of exlusive domains**
  - Domains that only appear in Google results (Google only)
  - Domains that do not appear in Google but do appear in the other three search engines (all but Google)

- **YouTube, a subsidiary of Google, is an exclusive domain (all but Google) in the German results**
- **Fox News is another exclusive domain (all but Google; US), along with 5 other news services like CNN, Huffpost, and Newsweek**

	Google only	All but Google
Germany	20	13
US	17	13

Table 3. Exclusive domains in the top 50 domains

RESULTS | DOMAIN DISTRIBUTION

- **The Gini coefficient**
  - A measure of statistical dispersion used to measure income or wealth inequality (Gini, 1936)
  - A single number ranging from 0 to 1, where 0 represents perfect equality, and 1 maximum inequality

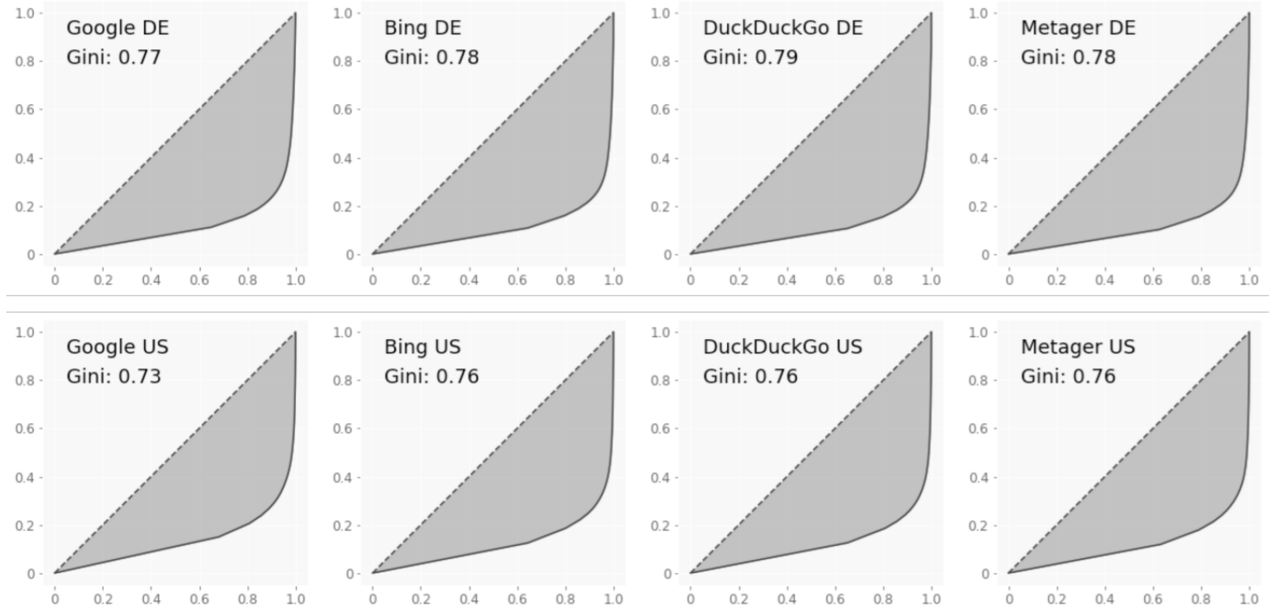

Fig 3. Source distribution of domains

RESULTS | DOMAIN DISTRIBUTION

- **The Gini coefficient**
  - A measure of statistical dispersion used to measure income or wealth inequality (Gini, 1936)
  - A single number ranging from 0 to 1, where 0 represents perfect equality, and 1 maximum inequality

- **All search engines have a Gini coefficient between 0.73 (Google, US) and 0.79 (DuckDuckGo, Germany)**

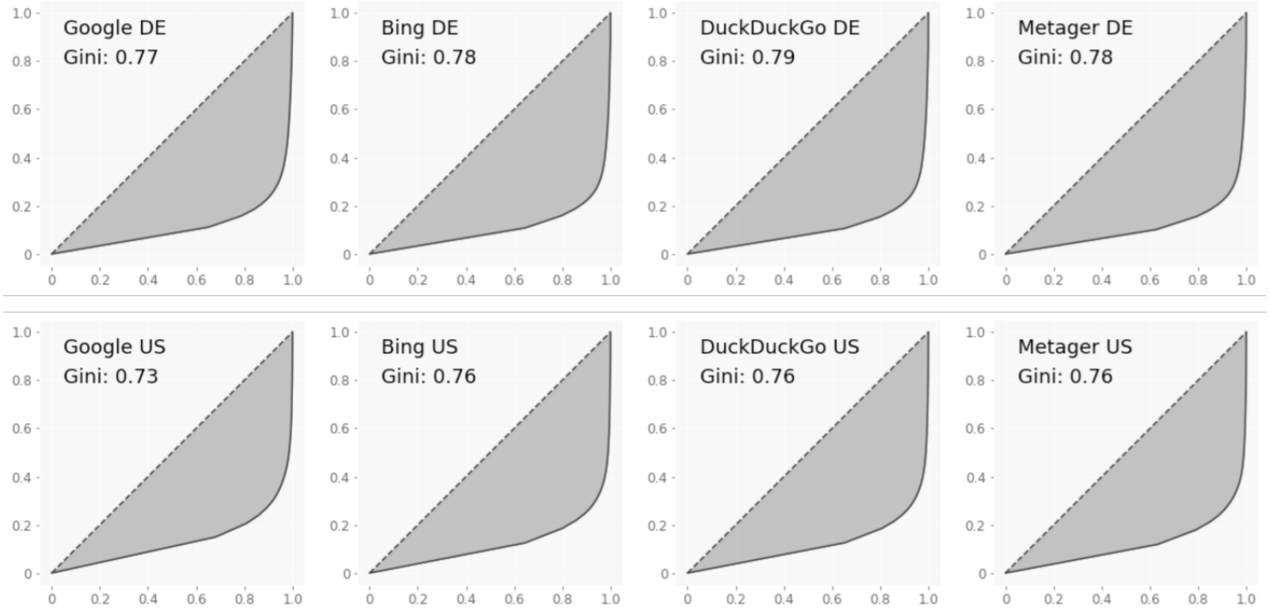

Fig 3. Source distribution of domains

RESULTS | DOMAIN OVERLAP

- **The highest overlap in the top 10 results is between Bing and Metager in the Germany (70%)**
- **Google and Bing overlap by 25% (US) and 28% (Germany)**

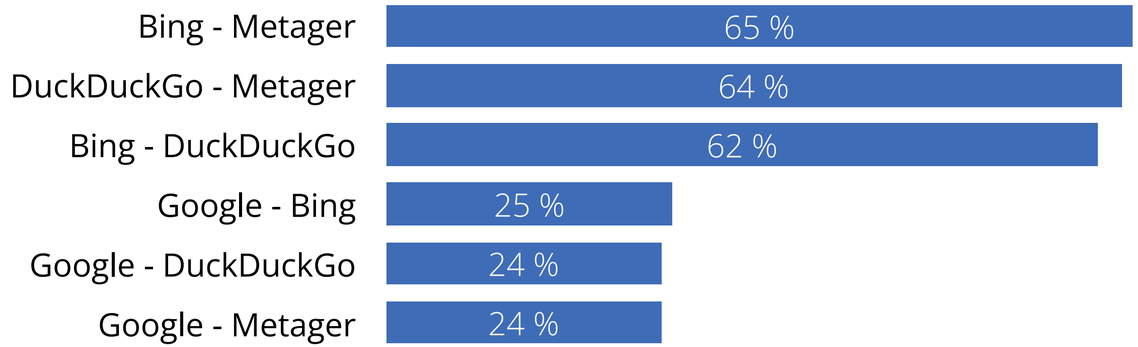

Fig 4. Domain overlap in top 10 results (US)

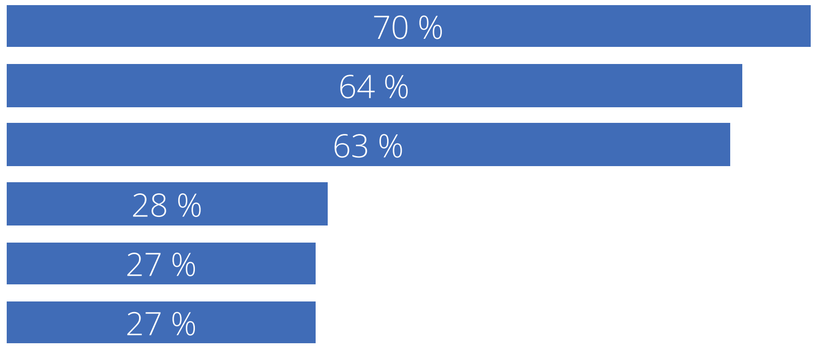

Fig 5. Domain overlap in top 10 results (Germany)

RESULTS | DOMAIN OVERLAP

- **The highest overlap in the top 10 results is between Bing and Metager in the Germany (70%)**
- **Google and Bing overlap by 25% (US) and 28% (Germany)**
- **Effect more pronounced in top 1 result (Germany):**
  - Highest overlap between Bing and Metager (78%)
  - Lowest overlap between Google and DuckDuckGo (10%)

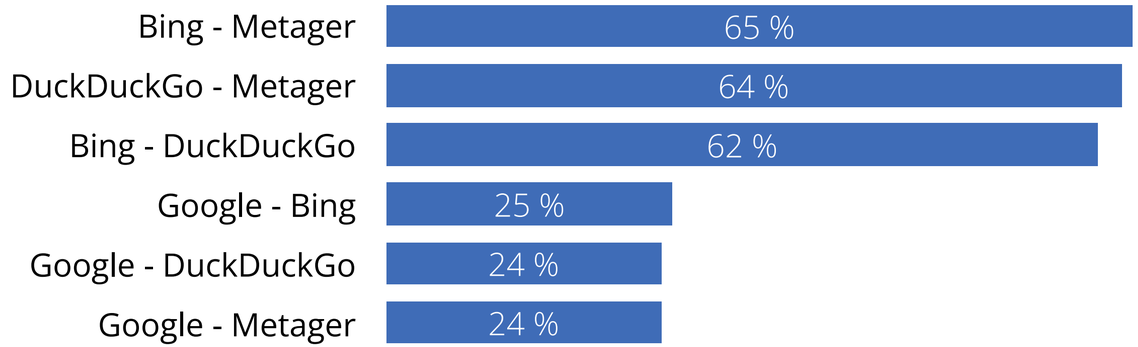

Fig 4. Domain overlap in top 10 results (US)

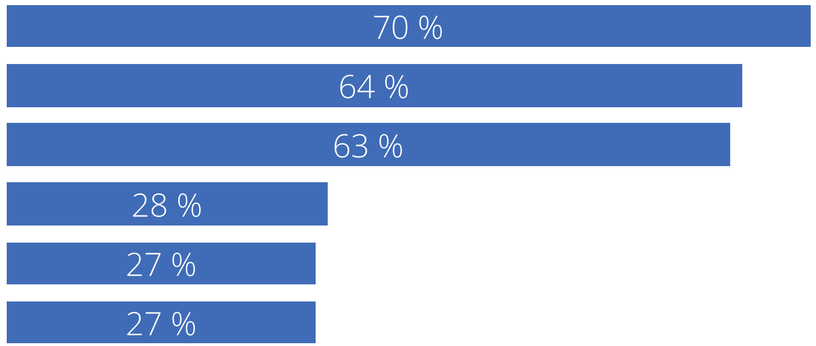

Fig 5. Domain overlap in top 10 results (Germany)

RESULTS | SUMMARY

- **Improvements to previous research**
  - Comparing the currently popular search engines (Google, Bing, DuckDuckGo) and a meta search engine (Metager)
  - Comparing results from two different countries/languages (Germany, US)
  - Using a large number of queries (Germany: 1,672 queries and 66,880 results; US: 1,865 queries and 74,600 results)

- **24 to 28% of overlap between Google and Bing; up to 70% of overlap between Bing and Metager (RQ1)**
- **Wikipedia and news services are the most popular website categories (RQ2)**
- **Few popular domains make up a large share of search results leading to low diversity (RQ3)**
  - Google Germany has the highest diversity in the top result (609 sources compared to 302-389)
  - The number of "Google only" exclusive domains is higher than the "not on Google" domains

Discussion and Conclusion

DISCUSSION AND CONCLUSION

- **Google Trends data as queries**
  - High number of diverse queries
  - Queries are mostly about news, celebrities, and sports, which leads to many news sources in the search results

- **Website categories**
  - Categories (News Service, Movies and Entertainment, etc.) are very broad
  - More refined categories, e.g. grouping sourced by trustworthiness, might explain the exclusive domains

DISCUSSION AND CONCLUSION

- **Using another or more than one search engine leads to seeing more diverse search results, allowing users to inform themselves more comprehensively**

- **Within each search engine's results, the concentration of sources shows that only a few top sources dominate the results**

- **The decision of which search engine to use, shapes what sources the user gets to see**

Thank you for your attention!

Nurce Yagci, Sebastian Sünkler, Helena Häußler, Dirk Lewandowski
Hamburg University of Applied Sciences, Germany

www.searchstudies.org
dirk.lewandowski@haw-hamburg.de

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

RAT RESULT ASSESSMENT TOOL
SEARCH STUDIES
HAW HAMBURG

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. https://doi.org/10.1016/S0169-7552(98)00127-5
- Bilal, D., & Ellis, R. (2011). Evaluating Leading Web Search Engines on Children's Queries. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6764 LNCS (Issue PART 4, pp. 549–558). https://doi.org/10.1007/978-3-642-21619- 0_67
- Cardoso, B., & Magalhães, J. (2011). Google, bing and a new perspective on ranking similarity. Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11, 1933–1936. https://doi.org/10.1145/2063576.2063858
- Ding, W., & Marchionini, G. (1996). A comparative study of web search service performance. *Proceedings of the ASIS Annual Meeting*, *33*, 136–142. https://eric.ed.gov/?id=EJ557172
- Edelman Trust Institute. (2022). *Edelman Trust Barometer 2022 - Global Report*. https://www.edelman.com/sites/g/files/aatuss191/files/2022-01/2022 Edelman Trust Barometer FINAL_Jan25.pdf
- European Commission. (2016). *Special Eurobarometer 447 – Online Platforms*. European Commission. https://doi.org/10.2759/937517
- Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, *208*(1), 73–79.
- Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, *16*(3), 169–185. https://doi.org/10.1080/01972240050133634
- Lewandowski, D., & Kammerer, Y. (2021). Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. Behaviour & Information Technology, 40(14), 1485–1515. https://doi.org/10.1080/0144929X.2020.1761450
- Lewandowski, D., & Sünkler, S. (2019). What does Google recommend when you want to compare insurance offerings? *Aslib Journal of Information Management*, *71*(3), 310–324. https://doi.org/10.1108/AJIM-07-2018-0172
- Makhortykh, M., Urman, A., & Ulloa, R. (2020). How search engines disseminate information about COVID-19 and why they should do better. Harvard Kennedy School Misinformation Review, 1(May), 1–12. https://doi.org/10.37016/mr-2020-017
- Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C., & Kleis Nielsen, R. (2021). *The Reuters Institute Digital News Report 2021*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf
- Purcell, K., Brenner, J., & Rainie, L. (2012). *Search Engine Use 2012*. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf
- Schultheiß, S., & Lewandowski, D. (2021). A representative online survey among German search engine users with a focus on questions regarding search engine optimization (SEO): a study within the SEO Effect project - *Working Paper 2*. https://osf.io/wzhxs
- 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. https://doi.org/10.1016/j.ipm.2005.11.001
- Steiner, M., Magin, M., Stark, B., & Geiß, S. (2022). Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society*, *25*(2), 217–241. https://doi.org/10.1080/1369118X.2020.1776367
- Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, *59*(11), 1702–1710. https://doi.org/10.1002/asi.20834