9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Who Provides the Largest Megaphone?

Keywords: Ranking algorithms, search engines, online news, state media, propaganda

Extended Abstract

Introduction

The Internet has not only digitized but also democratized information access across the globe. This gradual but path-breaking move to online information propagation has resulted in search engines playing an increasingly prominent role in shaping access to human knowledge. When an Internet user enters a query, the search engine sorts through the hundreds of billions of possible webpages to determine what to show. Google dominates the search engine market, with Google Search surpassing 80% market share globally every year of the last decade. Only in Russia and China do Google competitors claim more market share, with $\approx 60\%$ of Internet users in Russia preferring Yandex (compared to 40% in favor of Google) and more than 80% of China's Internet users accessing Baidu as of 2022 [2]. Notwithstanding this long-standing regional variation in Internet search providers, there is limited research showing how these providers compare in terms of propagating state-sponsored information.

Our study fills this research gap by focusing on Russian cyberspace and examining how Google and Yandex's search algorithms rank content from Russian state-controlled media (hereon, RSM) outlets. This question is timely and of practical interest given widespread reports indicating that RSM outlets have actively engaged in promoting Kremlin propaganda in the lead-up to, and in the aftermath of, the Russian invasion of Ukraine on February 24, 2022.

Methods

We consider six RSM outlets in our study: RIA Novosti (`ria.ru`), Lenta (`lenta.ru`), Gazeta (`gazeta.ru`), TASS (`tass.ru`), RT (`rt.com`), and Izvestia (`iz.ru`). We prepare and leverage a dataset including Google Search and Yandex Search queries and results from Russia over a 90-day period, from December 1, 2021, to March 1, 2022.

The search data was collected by first identifying the top trending queries for `google.ru` by day from Google Trends. For Google Search, we sent the daily trending queries to a privately hosted instance of Searx, an open-source meta-search engine, from an IP address located in Russia, and logged all search results returned until the pagination limit was reached. In case queries were blocked by Searx, we instead leveraged the Google Search Engine Results Page (SERP) API provided by DataForSEO to populate this information. For Yandex Search, we collected search results for each of the same set of queries using the Yandex SERP API provided by SerpWow. Though the queries trending on Google are not necessarily the same queries that are trending on Yandex, this data collection method ensured that the exact same set of queries was used to collect search results across the two search engines, providing a fair comparison of the behaviors of their respective information ranking algorithms.

Results

As seen in Figure 1, we find that across all trending queries in Russia, an average of 7.74% of the articles on the first page of Google search results belonged to RSM outlets. For Yandex, this

proportion is lower at 3.73% of articles but steadily increases over the 90-day analysis window. These numbers suggest that the average Internet user in Russia was exposed to a sizable volume of RSM content using Google Search or Yandex Search. Further, they indicate that Google's algorithms ranked RSM content more favorably than Yandex's by a factor of two.

A possible explanation for the difference is that Google's search algorithms assign a higher rank to news media sources in general (RSM or otherwise), compared to Yandex. To test this theory, we defined a set of control media outlets consisting of two popular international news sources in Russia, the BBC (`bbc.com`) and Forbes (`forbes.ru`), and three independent, domestic media outlets, TV Rain (`tvrain.com`), Novaya Gazeta (`novayagazeta.ru`), and Interfax (`interfax.ru`). We then analyzed how often webpages from these news outlets appeared as results for trending queries. We found that content from the control outlets appeared more frequently on Google Search than Yandex Search by a factor of 3.45, indicating that Google's algorithms do appear to assign a higher weight to news media sources (whether implicitly or explicitly). Still, both search engines return substantively more results from RSM sources compared to the control – a fourfold increase for Google and sevenfold for Yandex.

Discussion

We recognize three limitations in our study. First, Google Trends is an imperfect representation of search queries, as previously described in the literature [1]. Notably, it does not provide search volumes and instead aggregates queries and compares their relative variations across time. We believe this limitation is acceptable here as we are most interested in the webpages that appear as search results, rather than the search volumes for queries themselves. Second, we collected Google search results using Searx and DataForSEO, and Yandex results using SerpWow. This was due to Searx and DataForSEO imposing restrictions on access to their services from IP addresses in Russia after the invasion began in February 2022. Third, our analysis does not cover all RSM outlets, instead focusing on six of the most popular, and our control group consists of only five independent media outlets. Still, we believe that given the prevalence of these outlets in Russia, the results remain representative.

This work contributes timely and important insight into the role of two of the largest search engines in sorting, and thereby, implicitly recommending, content from RSM sources amidst an ongoing armed conflict. Our results contradict the common assumption that Yandex is the primary driver of an online information environment over-represented by RSM content in Russia. Instead, while Yandex does appear to rank RSM content more highly than Google relative to the rate at which it returns results from news media sources, both search engines frequently surface content from Russia state-controlled sources and, in fact, Google returns RSM at a higher absolute rate. This finding has serious implications for our understanding of the dissemination and visibility of state-promoted information online.

References

[1] Alessandro Rovetta. "Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research". In: *Frontiers in Research Metrics and Analytics* 6 (2021), p. 670226.

[2] StatCounter. *Worldwide desktop market share of leading search engines from January 2010 to July 2022*. Accessed: 2022-12-22. 2022. URL: `https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines`.

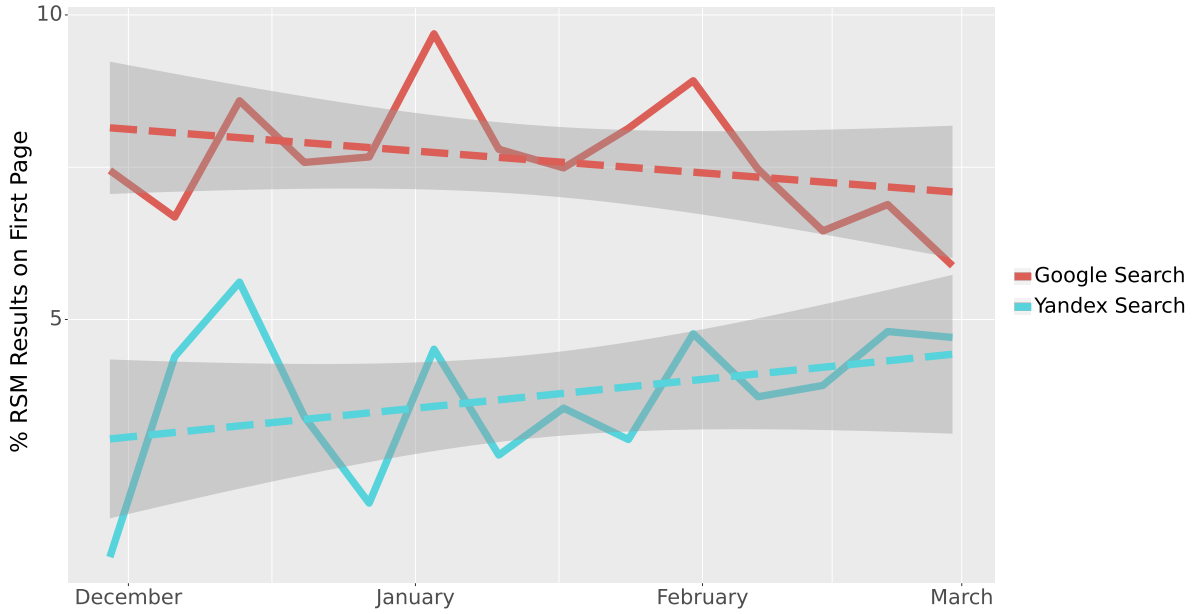

Figure 1: Percentage of Russia state-controlled news media results on the first page of search results across daily trending queries on Google Search and Yandex Search. The dotted lines represents smoothed conditional means and the gray region indicates the corresponding 95% confidence intervals.

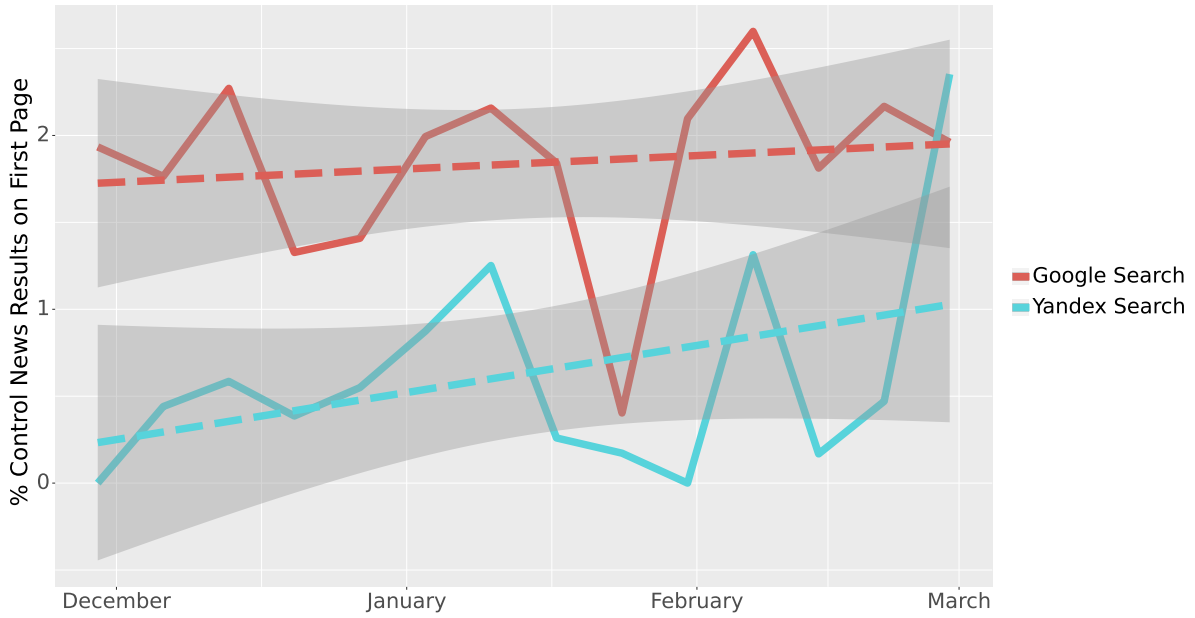

Figure 2: Percentage of Google Search and Yandex Search results from a sample of international news outlets and independent, Russian media outlets on the first page of search results across daily trending queries from Google Trends. The dotted lines represents smoothed conditional means and the gray region indicates the corresponding 95% confidence intervals.