Is Ignorance Bliss?

Latent Knowledge and the Opportunity Cost of Internet Search

Saurabh Khanna

Stanford University

July 28, 2021

Outline

Problem

Knowledge search on the internet is biased (Nissenbaum & Introna 2000, Vaughan & Thelwall 2004, Eubanks 2018, Selbst et al. 2019):

- Targets ~~representation~~ consumption
- Trained on ~~diverse~~ mainstream data
- Trained by ~~diverse~~ homogeneous teams
- ~~Public~~ Proprietary access
- Illusion of fairness
- Speed and scale doesn't help here

Problem

- What we know $\rightarrow$ Internet search (lacks representation)
- What we do not know $\rightarrow$ **?**

Problem

- What we know $\rightarrow$ Internet search (lacks representation)
- What we do not know $\rightarrow$ **?**

- An old problem

  I know that I know nothing.
  – Apology of Socrates, Plato (399 BC)

- But we have the data now to *approximate* an answer.

Research Questions

In the context of internet search:

1. Latent Knowledge: How much do we not know?
2. Opportunity Cost: What are the implications of knowing what we know as opposed to what we do not?

Approach

In the context of internet search (European Commission v Google, 2017):

- First page of search results receives 95% of all clicks
- First result on page 2 receives only 1% of all clicks
- 91% of pages get no organic search traffic

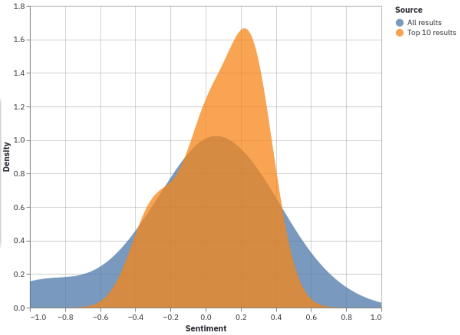
Latent Knowledge

Case 1: Continuous metrics

Given N search results for a given query, define latent knowledge as the non-overlap in metric distribution of top n vs N search results $(N \geq n)$.

If metric x is continuous

$$K_{latent} = 1 - \left[\frac{1}{2} \int |f_N(x) - f_n(x)| \; dx \right]$$



$K_{latent} \in [0, 1]$

Latent Knowledge

Case 2: Categorical metrics

Given N search results for a given query, define latent knowledge as the scaled difference in proportions across all categories for top n vs N search results $(N \geq n)$.

If metric is categorical with C categories

$$K_{latent} = \frac{1}{C-1} \sum_{c=1}^{C} \left[\frac{max\left(\frac{N_c}{N} - \frac{n_c}{n}, 0\right)}{1 - \frac{n}{N}} \right]^{1/2}$$

Binary case

$$K_{latent} = \left[\frac{max\left(\frac{N_b}{N} - \frac{n_b}{n}, 0\right)}{1 - \frac{n}{N}} \right]^{1/2}$$

$K_{latent} \in [0, 1]$

Data

User facing open-source platform (Sonder):

- Fetch web search results based on user query
- For every result, can calculate: Text metrics (sentiment, subjectivity, readability, novelty, etc.), Geo-location, Green web hosting

Logs:

- Daily top trends for web and news
- Run each trend through the platform: 47 countries (6 continents) $\times$ 40 trends/country $\times$ 100 results/trend = 188,000 web + news search results every day
- In the pipeline – Yandex (Russia), Baidu (China), Naver (South Korea)

Demo

S🎈nder

Opportunity Cost

OC = Return(Latent Results) - Return(Top Results)

Experimental Design

- Treatment: K_{latent} (i.e. results reordered to maximize exposure to latent knowledge)
- Control: Default ordering
- Outcomes: User-level metrics like polarization, access to misinformation, general click through behavior during and after treatment

Possibilities

- Descriptive Studies
  - Polarization
    - search sentiment variance by rank
  - Press freedom
    - sentiment difference between web and news trends by country
  - Compare search platforms
  - Trends in carbon cost
- Latent Knowledge
  - Variation by metric, by search platform, by search rank, by trend rank, across countries, over time

References

- Cardoso, R. (2017). *Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service.* European Commission.
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.
- Nissenbaum, L. D. I. H., & Introna, L. D. (2000). Shaping the web: Why the politics of search engines matters. *The Information Society, 16*(3), 169-185.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59-68.
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information processing & management, 40*(4), 693-707.