Is Ignorance Bliss?

Latent Knowledge and the Opportunity Cost of Internet Search

Saurabh Khanna

Stanford University

July 27, 2021

Outline

1. Problem
2. Objectives
3. Approach
   - Latent Knowledge
   - Data
   - Opportunity Cost
4. Possibilities

Problem

Searching for knowledge on the internet is biased (Selbst et al., 2019; Eubanks, 2018; Tegmark, 2017):

- 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/consume $\rightarrow$ Internet search (lacks representation)
- What we do not know/miss out on $\rightarrow$ **?**

Problem

- What we know/consume $\rightarrow$ Internet search (lacks representation)
- What we do not know/miss out on $\rightarrow$ **?**

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:

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

Approach

Latent Knowledge

Context

In the context of internet search (European Commission Report, 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

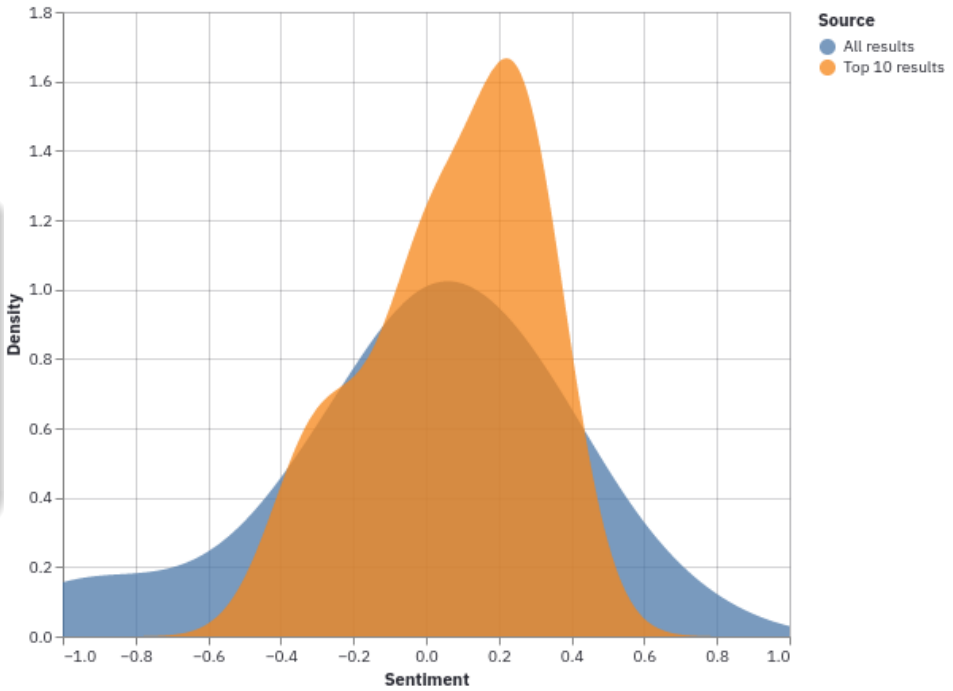
Latent Knowledge

Continuous metrics

Given N search results for a given query, Latent Knowledge is defined as the non-overlap in metric distribution of top n vs N search results ($N \geq n$).

If metric x is continuous

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



Latent Knowledge

Categorical metrics

Given N search results for a given query, Latent Knowledge is defined 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

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

Data

User facing 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

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, tolerance, 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