Independent Study Proposal - Spring 2023

Sam Pasco
January 9th, 2023
ASC faculty supervisor: Matthew Brook O'Donnell, Ph.D.

IS title:

Creating an Unbiased Search Directory for Localized Transactional Searches

Statement of Purpose:

As the amount of online content continues to grow, it becomes increasingly important to have effective information regimes (navigation tools) to help us find and access the most relevant media and information. However, current search engines (SE) like Google often prioritize results based on factors such as advertising, website speed, and popularity, which can lead to biased results. The goal of this independent study is to understand and quantify the nature and sources of bias in transactional search results in Google and other SE. From this analysis I aim to propose and develop a pilot system of organizing transactional search results in a less biased, more neutral way, centered on a comprehensive local directory that can help consumers better find products.

My goal is to develop a directory-based, meta search tool where all results for a given product/business are visible and sortable, regardless of user location, ad spend, or site engagement - like a modern, digital version of the Yellow Pages. To achieve this, I will utilize web scraping to gather a large dataset of web pages and implement Python code to label and sort the websites within certain categories and provide the user with more relevant links.

This research is related to my previous coursework in Surveillance Capitalism (COMM 4630) and Computational Text Analysis for Communication Research (COMM 3130), as they involved using programming and text analysis techniques to analyze and process data to better prevent large technology companies from exploiting people through technology. It also relates to my work in Communication Behavior (COMM 125) as we explored filter bubbles and echo chambers.

Motivating Problem: Over break, I was searching for a pair of track spikes. Despite knowing

several running-specific stores nearby that could potentially sell the spikes, none were displayed when I searched. Google primarily displayed Nike ads and general athletic stores that didn't actually have the product I was looking for.

Google claims their mission is "to organize the world's information and make it universally accessible and useful" - but when it comes to transactional searches with the aim to find products from local businesses, they fail, as their results are not universally accessible as they tailored based on user location, search history and other factors not always actually related to the searched product or business.

Syllabus:

**Note: Meetings with advisor will happen approximately weekly, time TBD*

Textbooks and Readings

- Manning, C.D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge University Press.
    - Available online: https://nlp.stanford.edu/IR-book/
- Mitchell, R. (2018). *Web Scraping with Python* O'Reilly Media, Inc. (2nd Ed) ○ Available online: https://edu.anarcho copy.org/Programming%20Languages/Python/Web%20Scraping%20with%20Python,%202nd%20Edition.pdf
- Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

- Articles and relevant blog posts/online sources are listed through out the schedule below.

Assessment

- Annotated Bibliography (10%)
    - Create an annotated bibliography of 8-10 key readings on transactional search, search neutrality and bias.
- SERP Assignment (10%)
    - Collect and analyze the search engine results pages form a set of products from a range of search engines. Write a report summarizing findings.
- Web Scraping Assignment (10%)
    - Utilize web-scraping tools and techniques to build a search results dataset based on a selection of 5-10 products with known local availability across a selection of search engines.
- Search directory project (45%)
    - Develop and test Python code to implement a meta-search and unbiased ranking

algorithm for transactional searches for a small number of known locally available products.
    - Build a proof-of-concept web search application to demonstrate this algorithm.
- Project presentation (25%)

Tentative Schedule

(subject to change based on initial weeks of literature search and

planning) Week 1 - Introduction to search engine results and bias

Readings

- Blog posts and online articles
    - Ranking Results – How Google Search Works
    - Our Approach – How Google Search Works
    - 2022 Google Algorithm Recap and Looking Forward to 2023 - Vizion Interactive
    - Optimize your site for search engines (for beginners) - Search Console Help ○ SERP 101: All About Search Engine Results Pages
    - Google's RankBrain: A Complete Guide to the Google RankBrain Algorithm ○ Do Unbiased Search Engines Exist? Top Alternative Search Engines to Consider - Neeva
- Manning IR textbook Ch. 19 'Web search basics'
- Scholarly articles
    - Are Search Engines Biased? - https://www.sciencedirect.com/science/article/abs/pii/S1567422322000163 ○ SEARCH NEUTRALITY AS AN ANTITRUST PRINCIPLE - https://heinonline.org/HOL/Page?handle=hein.journals/gmlr19&id=1213&collection=journals&index=

Week 2 - Current Systems: Yahoo & Information Retrival

Readings

- Blog posts and online articles
    - Yahoo killing off Yahoo after 20 years of hierarchical organization | Ars Technica ○ The Yahoo Directory -- Once The Internet's Most Important Search Engine -- Is To Close
- Manning IR textbook Ch. 20 'Web crawling and indexes'
- Scholarly articles

- ○ Hierarchical Navigation: An Exploration of Yahoo! Directories
- ○ "PROPOSED REMEDIES FOR SEARCH BIAS" (Section 1)- https://lauren.vortex.com/google-ammori-pelican.pdf

Week 3 – More on product search and bis

Readings

- Manning IR textbook Ch. 21 'Link Analysis'
- Scholarly articles
  - Bonart, M., Samokhina, A., Helsenberg, G. and Schaer, P. (2019). An investigation of biases in web search engine query suggestions. *Online Information Review*.
  - Humphreys, A., Issac, M.H. Wang, R.J. (2021). Construal matching in online search: Applying text analysis to illuminate the consumer decision journey. *Journal of Marketing Research* 58 (6), 1101-1119
  - Maillé, P., Maudet, G., Simon, M. & Tuffin, B. (2022) Are Search Engines Biased? Detecting and Reducing Bias using Meta Search Engines. *Electronic Commerce Research and Applications*
  - Rao, N., Bansai, C., Mukherjee, S. & Maddila, C. (2020). Product insights: Analyzing product intents in web search. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2189-2192
  - Schultheiß, S. and Lewandowski, D. (2021). How users' knowledge of advertisement influences their viewing and selection behavior in search engines. *Journal for the Association for Information Science and Technology* 72 (3), 285-301.
  - Schultheiß, S. and Lewandowski, D. (2021). Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference. *Journal of Information Science*
  - Schultz, C.D. (2020), Informational, transactional, and navigational need of information: Relevance of search intention in search engine advertising. *Information Retrieval Journal* 23 (2), 117-135
  - Welty, C., Aroyo, L., Korn, F., McCarthy, S. M., & Zhao, S. (2021, October). Rapid instance-level knowledge acquisition for google maps from class-level common sense. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 9, pp. 143-154

Assignment 1: Write an annotated bibliography of 8-10 key readings and identify prominent themes

Week 4 - Exploring Web Scrapers

Readings

- Mitchell *Web Scrapping* textbook

Tasks

- Experiment with various web scrapers on various search engines, ex:
    - Scrapers
        - Browse AI
            - Python-based: requests + BeautifulSoup, Selenium and Playwright
        - Instant Data Scraper
    - Search Engines
        - Google, Google Maps, Bing, DuckDuckGo, Neeva, Valentin.app
        - 11 Privacy-Focused, Alternative Search Engines to Google

Assignment 2: Using incognito windows and various search engines, do various comparisons of their SERPS (search engine results pages)

Week 5 - Business Directories

Tasks

- Explore current websites that have business directories
- See if a library has a yellow pages/telephone directory book

Reading

- Yellow pages - Wikipedia
- https://www.loc.gov/resource/usteledirec.usteledirec04066/?sp=75&r=-0.051,0.082,1.103,0.552,0

Assignment 3: Determine 5-10 product related search queries (this will be used as the sample basis for the search engine tool) and begin scraping

Weeks 6 through 10 - Develop the directory based search tool

Assignment 4: Create (meta) search engine focused on localized transactional searches ●

Aim: To help consumers find physical items that may be physically closer than they realize and would otherwise never find due to Google's advertising bias; to correct the bias and level the playing the field for small businesses with quality products that are currently hidden

● Inspiration: A modern, digital and unbiased yellow pages search tool for local businesses and websites

Week 6 – Meta-search

Tasks

- Select search engines to use and parse each of the SERP to extract consistent and comparable results
- Collect pilot dataset of results

Week 7 – Result comparison and ranking comparison

Tasks

- Analyze overlap and differences in results from pilot dataset
- Consider ranking algorithm approaches and classify results

Week 8 – Identification of bias and missing results

Tasks

- For sample products in pilot dataset apply measures of bias
- Explore what is missing: Are certain items that should be retrieved not in dataset?

Week 9 – Collecting larger dataset

Tasks

- Use developed scripts to collect and organize a larger dataset of target products

Weeks 10 & 11 – Web interface for search and Sentiment Analysis

Tasks

- Develop simple proof-of-concept web-interface for code
- Look into incorporating sentiment analysis, keyword analysis, and topic modeling into the search tool

    - Linguistic Features · spaCy Usage Documentation
    - Topic Modelling in Python with spaCy and Gensim | by Tarek Ghanoum | Towards Data Science
    - Keyword analysis

Week 12: Testing and refining

- Contact search engines companies
    - Visit DuckDuckGo Headquarters (only a 20 minute drive from campus)
    - Email Neeva
- Continue to test the search engine prototype with a variety of search queries and gather feedback. Use this feedback to refine and improve the tool.
    - *Optional: Contact and receive feedback directly from local businesses and if they'd be willing to directly input their inventories into the tool*

Weeks 13 & 14: Finalizing the search engine and presentation

Assignment 5: Finalize the search engine tool and prepare a presentation on the process and results of the independent study