EECS 549
October 11, 2023
Haley Johnson

Final Project Proposal

Introduction

For nearly two decades, Google was so dominant in information retrieval that the company's name was synonymous with search itself. Critics, however, have noted that advertisements, personalization (Sullivan, 2022) and AI generated content (Langley, 2023) have degraded the quality of Google search. Activist and writer Cory Doctorow argues that internal pressure to continually expand and monetize their search business has "weakened Google's own ranking system to the point where a link's presence among the top results for a Google search is correlated with scamminess, not relevance" (2023). These problems have pushed some users to turn to social media platforms, such as Reddit and Tik Tok, for their information retrieval needs (Sullivan, 2022; Lorenz, 2023). Social media platforms are not optimized for general purpose search and often lack the sophisticated controls that search giants have — if consumer dissatisfaction was solely about Google, then it would be logical for consumers to turn to alternative search engines like Big or DuckDuckGo. This shift suggests that it isn't just Google — traditional search as a whole is failing to meet user's information needs.

Elan Ullendorff, a writer and adjunct professor at the University of Pennsylvania, argues that consumers' increased dissatisfaction with Google and general purpose search will lead to humans dominating the next era of search (Lorenz, 2023). "Human curation is always going to be the best way to get recommendations, we all want to connect and express and listen to other people. Unless we're looking for super cold hard facts, the way people say things is as much a part of communication as the things people are saying," he argues (ibid.). Social media search, in the eyes of some users, offers information that is more authentic (ibid.) and trustworthy (Sullivan, 2022). Dissatisfaction with traditional search, then, may be understood as a failure to capture these important aspects of relevance

While consumer's dissatisfaction with Google and increasing use of social media search has been widely reported, much of this evidence is anecdotal. The effectiveness of social media search still needs to be rigorously evaluated. In other words, this project seeks to investigate if consumers really prefer social media, or if they just dislike Google. This project compares the outputs of a custom social media search engine against Google for a list of predefined queries. directly address a query, provide novel information, or appear trustworthy or authentic. Comparing social media search to traditional search will provide insight into how search giants are failing to meet consumers demands and shed light on how other dimensions of relevance — namely authenticity and trustworthiness — are driving changing preferences.

Task Definition

Despite growing dissatisfaction with Google, reporting suggests that most users do not treat social media platforms as a general purpose search engine (Sullivan, 2022; Lorenz, 2023). Instead, they use social search for a particular kind of query: open ended questions where the user would like to see multiple human answers. This project will create a search engine indexed on question and answer data from Reddit. The question and answer format naturally orients this information retrieval project towards subjective, exploratory queries where responses from real humans are most relevant — the exact kind of queries where Google search commonly fails. Consequently, this project is not a general purpose search engine. Rather, it is focused on a particular query format and a particular kind of information needed.

In summary, this system is explicitly designed to help users find answers to queries where:

1. Real human judgements are important to the user's information need
2. There are a range of relevant answers
3. Users want to see multiple answers

Queries may include:

- What is a show you watch again and again?
- Americans: if you could move anywhere in the country, where would you go?
- What is something that is common knowledge in your field but generally misunderstood by the public?

Queries that are less suitable for this system include:

- What is the capital of Japan?
- How many members of congress are there?
- Driving distance to Chicago

Documents in this collection are **a reddit post and all of the comments in the dataset associated with it.** Note that some posts do not have any comments.

Documents in this collection include:

- The post titled "What is a simple explanation of the Monty Hall paradox?" (post ID 'jiflka') and all associated comments
    - Sample comment: "Basically the starting position has 1/3 but after the first round each door has a half percent. The door you chose before had 1/3 and the other door has ½. Hope it makes sense l there is a lot more to it actually but that is very resumed"
- The post titled "If life becomes a RPG videogame right now. What class would you be based on your decisions, experiences, job or career?" (post ID 'k090lv') and all associated comments
- The post titled "Does Joe have 253 electoral votes or 264 electoral votes ? I thought he just needed 6 to win but now I read it's 253 ?" (post ID "joxnbe") and all associated comments

Data

This project uses a dataset of posts and comments from r/askreddit. r/askreddit was created in 2008 and has over 43 million subscribers, as of October 8, 2023. Posts follow a question and answer format and cover a range of topics including product recommendations ("what life changing thing can you buy for less than $100"), advice ("what is the best advice you've ever received? The advice that has impacted your life the most") and hypothetical scenarios ("if you could telepathically say something that all 7.8 Billion people on earth could hear at once what would it be?").

This project will use a dataset of 631,000 posts and their respective comments from r/askreddit. The dataset was created in October 2021 by SocialGrep, a social media analytics company. The dataset contains unique 10,000,000 comments. The median number of comments per post is 4 and the mean is 15.3. In other words, the dataset has a strong left skew, where some posts have very high numbers of comments and others have much fewer. I have done some light preprocessing to combine posts with comments, but otherwise the data is in a tidy format and ready to use

Relevance scores will be generated by hand. For a predefined list of 300 evaluation queries, I will identify the 10 most relevant documents in my colllection. This will be the ground truth that I'll use when evaluating the information retrieval system. To compare the performance of social search against traditional search, I will use the Google Search API. Relevance scores for the results of the Google Search API will also be calculated manually using the same method.

The following table summarizes some of the information about the dataset shared above:

Statistic	Value
Mean number of comments per post	15.3
Median number of comments per post	4
Maximum number of comments on a post	51,673
Number of unique posts	631,063
Earliest post in dataset	Feburary 18, 2010
Latest post in dataset	November 30, 2020

Related Work

In a review of information retrieval in microblogging services, Efron (2011) identifies two key types of search on social media: "broadcasting questions to their followers in hopes that people in their social network will answer them" and "conducting searches over existing microblog data

in hope for discovering relevant information that has already been written." This project addresses the later type of search, as we are only looking at existing posts on r/askreddit and not trying to evaluate how effectively users can crowdsource answers on this subreddit. However, user's ability to solicit quality information and their ability to search for existing information are interrelated and reflect the underlying quality of crowdsourced contributions. Thankfully, r/askreddit is a well established and highly active community for question and answer style queries. Nevertheless, future work in generative search (such as ChatGPT) may consider focusing on user's ability to broadcast questions to their social network.

Previously, building a collection of question-answers pairs has been the main barrier to creating information retrieval systems optimized for question-answer formats (Ji et al., 2014). Other works have tried to circumvent this problem by generating question-answer pairs from text (Chen et al., 2011; Nouri et al., 2011). The r/askreddit dataset, however, has the advantage of being a topically diverse dataset made up entirely of real questions. We expect that results from our system will provide a more accurate sense of how Q&A-based information retrieval systems perform than work that uses augmented/computer generated data.

There is growing evidence for the efficacy of alternative search interfaces and formats. One study asked subjects to complete a list of tasks using ChatGPT or Google and found that participants perceived ChatGPT as being easier to use and providing higher quality answers (Xu et al., 2023). Interestingly, participants also reported similar levels of trust in ChatGPT and Google Search, despite ChatGPT's known issues with truthfulness and accuracy (ibid). These results suggest that user experience may be more important than the quality of search results to users. It also indicates that users may prefer the natural language format of ChatGPT over seeing a simple list of ranked search results. This project will provide further insight into alternative search formats.

Evaluation & Results

I will predefine a list of 300 queries to use to compare Google and social search. For each query, I will manually assign relevance judgements on a scale of 1-5, where 0 is not relevant and 5 is most relevant, for the first 10 results returned by each information retrieval system. If there are less than 10 results returned, then I will label all results. I'll calculate mean-average and precision and normalized cumulative discounted gain for both social search and Google for each query.

As discussed earlier, the social search system this project will create is optimized for a particular type of query. I will be taking the underlying information need behind these queries into account when calculating relevance. In particular, relevance, for the sake of the project, should encompass:

- How directly the result addresses the question
- The degree to a result explains the answer or offers additional context for the answer

    - For open-ended, ambiguous queries, having some sort of explanation can greatly improve the relevance. For example, if a question is “what is the best hike in Shenandoah National Park” the answer “Old Rag — the rock scrambles are lots of fun and unlike any other hike in the park” is much better than a simple “Old Rag”
- Novelty of the results
    - In an open-ended query having a diverse range of responses is important. A good document should include multiple points of view

In addition to evaluating my results against Google, I’ll also use two more simple baseline measures. First, I’ll compare how the social search system performs compared to randomly returning a result from the collection. Secondly, I’ll compare the social search system to BM25 with no hyperparameter tuning. This will likely perform better than just selecting a random document, but is still relatively simple.

Work Plan

I have already identified my data set, done preliminary preprocessing and gotten developer access to the Google search API.

This work plan roughly outlines what task I will work on each week.

- Week of October 15: Write code to preprocess / tokenize all documents
- Week of October 22: Write code to index documents
- Week of October 29: Write code to rank results
- Week of November 5: Create set of evaluation queries, begin labeling relevance scores for top 10 documents in collection
- Week of November 12: Continue labeling relevance scores, begin retrieving results from test queries from Google API and labeling relevance
- **Wednesday November 15th: Project Update Due**
- Week of November 19: Write code to calculate MAP and NCDG for all query sets
- Week of November 26: Begin writing report
- Week of December 3: Finish writing report, flex week to catch up on any parts of the project that have ran long
- Week of December 10: Polish report & report
- **Wednesday December 13th: Final Project Due**

Works Cited

Chen, G., E. Tosch, R. Artstein, A. Leuski, and D. R. Traum (2011, May). Evaluating conversational characters created through question generation. Twenty-Fourth International Florida Artificial Intelligence Research Society Conference

Doctorow, C. (2023, May 14). Google's AI hype circle. Medium. https://doctorow.medium.com/googles-ai-hype-circle-6158804d1299

Efron, M. (2011, June). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology, 62*(6), 998-1008. https://doi.org/10.1002/asi.21512

Ji, Z., Lu, Z., & Li, H. (2014) An Information Retrieval Approach to Short Conversation. arXiv preprint arXiv:1408.6988. Retrieved from https://arxiv.org/pdf/1408.6988.pdf

Langley, H. (2023, September 20). Google recently cut 'people' from its Search guidelines. Now, website owners say a flood of AI content is pushing them down in search results. Business Insider. https://www.businessinsider.com/google-search-helpful-content-update-results-drop-ai-generated-2023-9

Lorenz, T. (2023, July 20). Google it? People now are searching with TikTok or Reddit. The Washington Post. https://www.washingtonpost.com/technology/2023/07/20/google-search-problems-mount/

Nouri. E., Artstein, R., Leuski, A., & Traum, D. (2011, November). Augmenting Conversational Characters with Generated Question-Answer Pairs. AAAI Symposium on Question Generation

Sullivan, M. (2022, February 17). Is Reddit a better search engine than Google? Fast Company. https://www.fastcompany.com/90722739/is-reddit-a-better-search-engine-than-google

Xu, R., Feng. Y., & Chen, H. (2023, July). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience. arXiv preprint arXiv:2307.01135v1. Retrieved from https://arxiv.org/abs/2307.01135