CS47300 Fall 2020 Assignment 5

Due 14 November 2020 11:59pmEST

Updated 12 November 2020 11:25amEST (total retrieved in 5.1)

1 Data Privacy

1. How does personalization in search results impacts data privacy? If a search engine that tries to personalize your search results based on a profile you have provided wants to release search log data (queries and results) to the public, would user ID anonymization be enough? Explain why or why not. Hints: AOL Query Log Debacle.

2. What if users do not trust search engines and do not want to know/store more about users. However, you might already know enough in this class regarding how to form a user profile using context, search history, and query intent, improving search performance. Thus, there is a trade-off between protecting the personal information of users and giving better search results. Can we get the best of both? We want better search results without giving too much user information to search engines (not trusting search engines). Propose an idea of how this can be achieved? You need to describe the approach and how the expected trade-off can be achieved. Please be concise; half a page should be adequate. This is an active research area; you might find this SIGIR 2016 paper gives you some ideas.

2 Filter Bubbles

TikTok is a very popular platform where users can create, share and watch short videos. User's "information needs" are met through recommendations, based on profiles and collaborative filtering rather than explicit queries.

1. Have you used TikTok? How do TikTok recommend videos for you? How do you feel these are different from other recommender systems (for example, Amazon)?

2. Do you think TikToc is likely to produce filter bubbles? Give some reasons, and describe possible advantages and disadvantages of this.

3. How might you improve the recommender system to avoid generating filter bubbles?

If you are not familiar with TikTok, you may describe and answer the above questions another system you are familiar with that is based on recommendations and/or profile-based filtering rather than ad-hoc queries.

3 Bias

Give an example of how an ad-hoc retrieval system might show bias in answering a query. Address:

1. What group is the system biased towards/against?

2. What sort of harms might arise?

3. How might this have happened?

4. How you might modify an information retrieval system to address this problem, and how might that affect scores for how "good" the ad-hoc retrieval system is. (E.g, precision/recall/rank-based metrics.)

4 Fake News Detection

Analyze the following statements, in terms of their likelihood of being fake news.

1. "Neural networks are a bubble, the next big learning algorithm will make them irrelevant, as happens with all things."
2. "Neural networks are a bubble, the next big learning algorithm will make them irrelevant. "

   You should discuss:

   - How would you go about determining the legitimacy of such a statement?
   - Would you expect a consensus on the statement?
   - What are some difficulties about identifying the truth of such a statement?
3. For fake news, do you think false positives (wrongly identifying an article as fake) or false negatives (failing to identify a fake article) to be a worse problem. Briefly discuss why.
4. How does this relate to using recall and precision as measures of the ability to detect fake news?

5 Federated Search

You are constructing a federated search based on the idea of constructing a sample dataset, and using query results on the sample to determine both which sources to run a full query on, and how to merge the results.

Part 1: Sample Dataset Construction You are given 100 sample documents from Server A (which has a total of 100,000 documents), 50 from Server B (which has a total of 1000), and 50 from Server C (which has a total of 10,000).

Server D tells you how many total documents contain a term. You issue a query for "Information", and retrieve the top 30 documents but are told there are 500 containing the word. You then issue a query for "Virus" and retrieve the top 30 documents, you are told 50 match. There is one document that appears in the top 30 results for both queries.

Estimate the total number of documents in Server D.

You then do the same two queries for Server E. Server E gives you 50 documents containing "Information" and 50 containing "Virus", there are 10 documents that appear in both top 50 lists. Estimate the total number of documents in Server E.

Part 2: Source selection You run a query "Indiana Covid-19 statistics". the sample and get 1 matching document from Server B and 3 matching documents from Server C. The ranking is $D_{Cs1}, D_{Bs1}, D_{Cs2}, D_{Cs3}$.

What servers would you recommend querying, and what query would you use, if your goal is to get reasonable results for "Indiana Covid-19 statistics" without using too much bandwidth / computing resources.

Part 3: Results merging Suppose you issue the full query Indiana Covid-19 statistics to only Servers B and C. You get 10 documents from Server B, with the matching sample document D_{Bs1} 3rd on the list. From server C, you get 10 documents, D_{Cs1} is third on the list and D_{Cs2} is 5th on the list.

Come up with a means of merging the documents. You don't have scores, you only have rankings. Give your combined ranked list of 20 documents, and show how you determined the proper ordering. Note that this will require you to make some assumptions and estimates that combine different approaches we discussed in class.