Brainstorm - the personalized search engine for scholarly results

Abiola Oyebanjo
abiola.oyebanjo@fu-berlin.de

March 30, 2021

Abstract

Optimizing search results, though an important component of the internet and the web world, is relatively less discussed in research and practice. Few scholars have identified that search can be time and money consuming - and this can be appalling if users are unsuccessful in obtaining desired results in the process. Research on personalized search aggregation have also been used mainly for marketing and business purpose. The common method in the field is ranking users preference and matching output that serves recommendation systems. However, single user-independent ranking model are often insufficient to satisfy different users' result preferences. I propose a transfer learning, the project creates word embeddings for the universe of search queries and a model that assign labels to new ones. This labels will be synchronized with personalized and optimized search histories to give better results for academic recommendations than standard search engines. The search query data with COVID-19 intent will be used for this analysis.

1 Proposed Method

The method is to evaluate the performance of matching results from search queries with journals and social media APIs, with the results from trained models that use personalized results, and labels from word embedding. Using transfter learning, I will first pre-train an auxiliary dataset on the text classification domain to build the model. Ther training objective is to maximize the log likelihood of the sentences.t

2 Experiments

Data: The dataset will be curated from the Bing search logs (desktop users only) over a 13-month period of Jan 1st, 2020 till January 2021. Only searches that were issued many times by multiple users were included. The dataset includes queries from all over the world that had an intent related to the Coronavirus or COVID-19. Data includes (1) Date; (2) Query which is the actual search query issued by user(s); (3) Query Implicitness which is a Boolean and true if query did not mention COVID or coronavirus or sarsncov2 and false if otherwise. It also include the (4) State from where the query was issued, (5) the country and the (6) PopularityScore which is a value between 1 and 100 inclusive (1 indicates least popular query on the day/State/Country with Coronavirus intent, and 100 indicates the most popular query for the same geography on the same day

Evaluation method: I will use Google's BERT to access the library for the learning of word embeddings and classification. These words are pre-trained on a large corpus and can be plugged in a variety of downstream task models to automatically improve their performance. Features that will be optimized in the classification from each search result considering the Bing data are popularity feature of each query and query content. This will also allow me to identity search themes and top words.

Furthermore, I will use labels and scores from the models trained to establishing the similarities between the top classification associated with each search results and contents in academic journals. I will also convert the core features of interest that is scrapped from each academic journal into a vector. Content in academic journal, or social media will be filter to contain features such as recent data, most mentioned, and most cited. Using Scikit-Learn in python that provides the function, the Cosine similarity between these two documents will be calculated. Data from journals and social media will be sourced by their respective APIs.

Results: From descriptive statistics, the total search queries from all countries are 6,283,896 observations with 197,472 unique queries. The are 47,992 unique queries with more than 3 string length; and there are 3,928 unique queries with string length more than 6. As at this report, no performance have not been generated.

3 Future work

The current idea for creating unique ID per query is to randomize. Since there are 47,992 unique queries, I will create 47,922 unique ID and the randomly assign queries to unique ID across the entire data, grouped by user's location. I will also ensure that there are variations across users with some users assigned more queries than the other. It is not entirely clear if this method will bias the result. If this problem becomes inextricable, I will use the popularity score for the most popular research in the past week per location, as a precondition for optimizing each search results. For example, if a user searches ”When is Corona likely to end” in location A, and the most ranked search in the user's location is ”Who started Corona” in location A, both queries will be merged as list and checked for semantic similarities - and the final results will be used to index the user's query. In general, I can adopt both methods and compare their predictive powers.

1

References

[1] Teevan, Jaime, Daniel J. Liebling, and Gayathri Ravichandran Geetha, “Understanding and predicting personal navigation.,” In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 85-94, 01 & 2001.

[2] Zhou, Yun, and W. Bruce Croft. ”Query performance prediction in web search environments., “In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 543-550” 01 & 2007.

[3] Teevan, Jaime, Susan T. Dumais, and Daniel J. Liebling. ”To personalize or not to personalize, “modeling queries with variation in user intent.” In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 163-170” 01 & 2008.