Similarity of Search Results in the Datenspende BTW17 Dataset - CSS Data Challenge 2018

Johannes Nakayama *RWTH Aachen University*
Nils Plettenberg *RWTH Aachen University*
Laura Burbach *RWTH Aachen University*
André Calero Valdez *RWTH Aachen University*

The filter bubble theory, popularized by Eli Pariser in 2011 is subject of academic discussion. With political polarization seemingly increasing, this issue increases in relevance. As recommender algorithms in news-related data bases or search engines such as Google are often intransparent, the project Datenspende: BTW17 aimed at providing data from the users point of view. Google search results for the major German political parties and their leaders of over 1000 participants were collected in order to make sophisticated analyses possible. Based on the data provided, we conducted exploratory tests to investigate how search results of different users differed. As we deemed the order of the search results relevant as well, we compared how the first three, six and nine results were different respectively. We found that differences were the strongest for the first three results, with large differences between the search terms as well, indicating that for some queries filter bubbles could exist.

Keywords: Datenspende BTW17, Algorithmwatch, Search Bias

Introduction

In 2011, Eli Pariser popularized the idea that there might be a "filter bubble", i.e., that algorithms determine which kinds of contents we are shown on the web whenever we are using systems that use recommender algorithms (Pariser, 2011). He suspected that people were rather going to be shown contents that confirm their views, thus shielding them from challenging opinions. As cognitive biases in humans are reasonably well documented, this issue grows in relevance. People tend to believe information more readily if it supports their current world view (Hart et al., 2009). This effect even increases when the information is received sequentially (Jonas et al., 2001). Under the assumption that Google users usually follow several links of their search results, highly personalized Google results would therefore probably contribute to political polarization. The "Datenspende: BTW17" adresses this issue by providing data on Google search results.

The data source

The project "Datenspende: BTW17" was a large scale data mining initiative with the purpose of collecting google search results throughout the month of the German federal election of 2017. Over 1000 people participated by installing a browser plugin, which was specifically developed for the project, on their computers. In the participants' browsers, it conducted a regular google search and a google news search for a set of predefined terms every day throughout september 2017, the month of the German federal election. The terms in question were the names of the major political parties in Germany and their leaders:

Method

As a first approach, we only examined the searches, that were conducted with the participant logged into Google and where the language the search was conducted in was German. We also excluded the google news queries and focussed on the regular searches. To determine how similar the searches for a specific term were, we developed a function, that compares the first n results of a search result list with the first n results of every other search result list for a given *term*. We then computed the similarity by measuring set overlap (Singhal et al., 2001) for the first three, six, and nine results respectively to see if there were significant differences in the order the search items were presented in. The similarity scores are given as an average percentage of how many items every search result list for a given term shared with every other search result list.

We only looked into the data set for september 30th 2017, but the method applies for the other data sets as well and we intend to look into them as we proceed.

Results

We first look at the data set in general and try to provide a high level understanding of the quality of the data. We then look at the similarities of the data-set from a top-3, top-6, and top-9 perspective.

Descriptive Data Analysis

Our dataset contained $n = 28704$ search results. For each google search result list, a hash key was provided. Two results that were exactly the same produced the same hash key. Some hash keys even appeared close to or over a hundred times. Approximately half of the searches (14,640 out of 28,074) were conducted with the participant logged into Google. The search type ("search" = usual google search, "news" = google news search) was quite evenly distributed as well ("search": 14,080, "news": 14,624). In the vast majority of cases, the language of the search was conducted in was German (see Fig. 1).

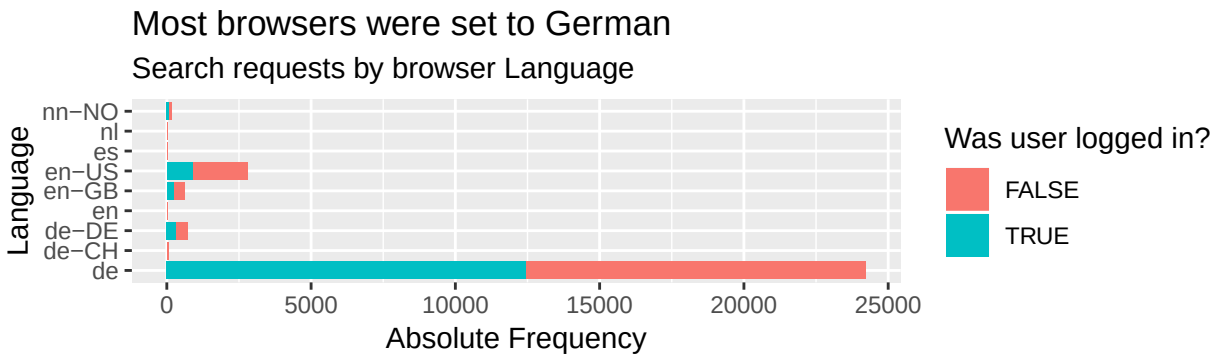

Figure 1: Comparison of browser languages

Results from Similarity Analysis

We next look at the Similarity of the results lists looking at the top three, six, and nine results first. Since users often only look at top results, we found these to be of primary interest. The results are shown in the following figure

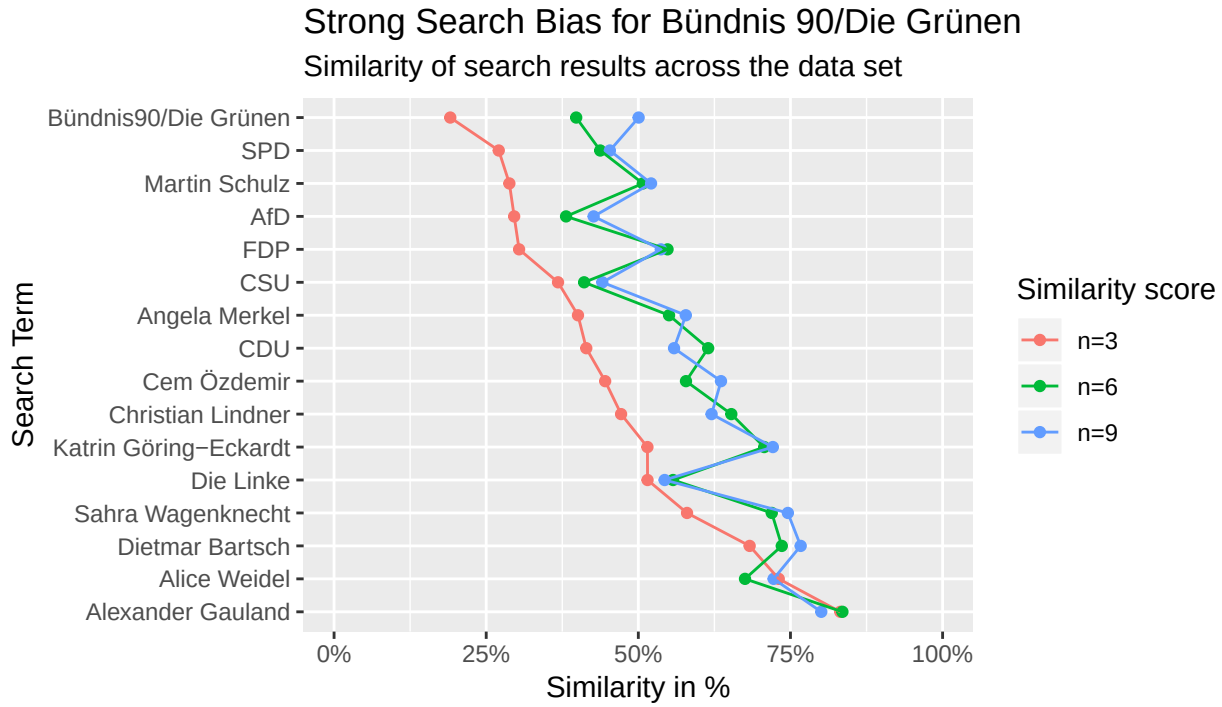

Figure 2: Similarity of all results for given search keyword, sorted by n=3

As can be seen (see Figure 2), the similarities for the first three results differ greatly among terms (range of 64.1 %) with "Bündnis90/Die Grünen" differing the most (19.1 % overlap on average) and "Alexander Gauland" differing the least (83.2 % overlap on average). The range for the first six results is 45.4 %, with "AfD" differing the most (38.1 % overlap on average) and "Alexander Gauland" differing the least (83.5 % overlap on average). Lastly, for the first nine results the range is 37.4 %, with "AfD" differing the most (42.7 % overlap on average) and "Alexander Gauland" differing the least (80.1 % overlap on average).

Discussion

In this short article, we looked at the Datenspende BTW17 Dataset very selectively and tried to estimate, whether or not a "filter bubble" is present depending on the resulting search term. For this purpose we measured the average set overlap of all results from september 31st 2017 for all individual search terms. We found relatively strong differences in the resulting overlaps. We found very high overlaps "Alexander Gauland" and "Alice Weidel" and low overlaps for "Bündnis 90/Die Grünen" and "SPD".

These results indicate that not all search results are equal with respect to a possible filter bubble. However, these results must be interpreted carefully. Our method does not allow to identify subgroups in overlaps. These subgroups would be individual filter bubbles. Our method only shows an overall tendency for results diversification in the the top-3, top-6, or top-9 results. In these results, strong filter bubble effects are present especially for "Bündnis 90/Die Grüne".

Limitations

So far we did not look into the actual results, but only looked at the returned hash-keys. We further ignored language settings for the similarities and looked only at German results. Results that are *only* similar in nature are seen as different through this method. As future research we will look at the actual linked documents and the full data set.

References

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555.

Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology*, 80(4):557.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.