An Investigation of Biases in Web Search Engine Query Suggestions

Malte Bonart

TH Köln - University of Applied Sciences, Cologne, Germany

Anastasiia Samokhina Recogizer Group GmbH, Bonn, Germany

Gernot Heisenberg

TH Köln - University of Applied Sciences, Cologne, Germany

Philipp Schaer

TH Köln - University of Applied Sciences, Cologne, Germany

Purpose - Survey-based studies suggest that search engines are trusted more than social media or even traditional news, although cases of false information or defamation are known. In this study, we analyze query suggestion features of three search engines to see if these features introduce some bias into the query and search process that might compromise this trust. We test our approach on person-related search suggestions by querying the names of politicians from the German Bundestag before the German federal election of 2017.

Design/methodology/approach - This study introduces a framework to systematically examine and automatically analyze the varieties in different query suggestions for person names offered by major search engines. To test our framework, we collected data from the Google, Bing, and DuckDuckGo query suggestion APIs over a period of four months for 629 different names of German politicians. The suggestions were clustered and statistically analyzed with regards to different biases, like gender, party, or age and with regards to the stability of the suggestions over time.

Findings - By using our framework, we located three semantic clusters within our data set: Suggestions related to (1) politics and economics, (2) location information, and (3) personal and other miscellaneous topics. Among other effects, the results of the analysis show a small bias in the form that male politicians receive slightly fewer suggestions on "personal and misc" topics. The stability analysis of the suggested terms over time shows that some suggestions are prevalent most of the time, while other suggestions fluctuate more often.

Originality/value - This study proposes a novel framework to automatically identify biases in web search engine query suggestions for person-related searches. Applying this framework on a set of person-related query suggestions shows first insights into the influence search engines can have on the query process of users that seek out information on politicians.

Keywords - Google, Bing, DuckDuckGo, Query Suggestions, Bias, Politics, Text Mining, Clustering, Regression, Rank Stability Analysis

Paper type - Research Paper^a

^aPreprint, accepted for publication, 27-Sep-2019, *Online Information Review*, forthcoming, *https://doi.org/10.1108/OIR-11-2018-0341*. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Any reuse is allowed in accordance with the terms outlined by the licence. For commercial purposes, permission should be sought by contacting permissions@emeraldinsight.com.

1. Introduction

A recent survey by the PEW Research Center (Matsa and Lu, 2016) showed that today most American adults still get their news from TV (57%), but that online sources are a strong second with 38% followed by radio (27%) and print news (20%). In younger cohorts online sources exceed TV news consumption and we can see a clear trend for a rising importance of online news sources. Politics are no exception to this change in information behavior. As paraphrased by Schiffmann: "were it not for the internet, Barack Obama would not be president" (Schiffmann, 2008). In light of the ongoing discussion on "fake news" and the impact of Donald Trump's online presidential election campaign (Parkinson, 2016) the influence of the systems underlying online news and information sources should be discussed.

Information found online, might influence real-world events such as the results of elections, especially if search engines manipulate the way information is queried. Past research studies have shown that biased search rankings can shift the voting preferences of undecided voters by up to 20% and more, and such rankings can be masked in such a way that people show no awareness of the manipulation (Epstein and Robertson, 2015). While the presentation of the effect size is criticized for being exaggerated and falsely calculated (Zweig, 2017), the study nevertheless shows the importance of looking for biases in online systems concerning their real-world impact. In pre-election phases shifting the voting preferences can have a particular influence on the future political scene of an entire country (Larcinese and Miner, 2017). The influence that search engines might have on events shows the importance of studying online search behavior.

Focusing on the influence of search engines is of high importance if we look at the level of trust people show in search engines. The 2017 Global Edelman Trust Barometer (Harary et al., 2017) asked for the trust in different sources of news and information. It ranked search engines as the most trusted information sources (64%) followed among others by traditional media (57%), and social media (41%).

Specifically for person-related searches, we ask if this trust is compromised by biases inherent to the underlying retrieval, ranking, and suggestion algorithms. In the following paper, we present a general query suggestion data acquisition, processing, and analysis framework to automatically identify biases in an exploratory manner. An example of different query suggestions is shown in Figure 1.

We concentrate on the following research questions:

- **RQ1** How can we automatically identify clusters and patterns in web query suggestions for person-related searches?
- **RQ2** To what extent can metadata on the persons searched (e.g., gender, age, party membership) be used to explain possible biases?
- **RQ3** How can we measure and analyze the stability and persistence of possible biases in the suggestions?

angela merkel $ ho$	
Angela Merkel Chancellor of Germany	
angela merkel	angela merkel × Q
angela merkel news	encele medici
angela merkel biography	angela merkel biography
angela merkel jewish	angela merkel young
angela merkel age	angela merkel husband
angela merkel trump	angela merkel trump
angela merkel birthdav	angela merkel eye roll
angela merkel migration	angela merkel twitter angela merkel net worth

Figure 1. Two different query suggestions for the query on Angela Merkel (Bing on the left, DuckDuckGo on the right).

We approached these research questions by focusing on the tasks of identifying any systematic patterns in the query suggestions (RQ1) and by concluding how these suggestions are generated (RQ2). As the crawled data sets typically contain query suggestions from a longer period, we were also interested in the long-term fluctuation of the suggestion terms (RQ3).

We intend to contribute to the topic of information biases in search engines by applying a text mining approach. We focus on the statistical identification and quantification of systematic patterns in the data. Whether possible biases do influence the process of political opinion formation of individuals, is not part of this work.

In an experimental study, we applied our framework to the case of online searches for German politicians: We gathered and analyzed query suggestion data for the names of politicians from three popular web search engines during four months before the German federal election 2017. We used the names of the politicians of the German Bundestag (the German federal parliament) and enriched the crawled data with socio-demographic information, such as age, gender, home state, and party membership.

We contribute to the research by presenting a novel framework to automatically identify biases in web search query suggestions for person-related searches. We do not study query suggestions for individual cases but systematically for many search terms and over a long period. The presented methods do not need access to the search engine's query log or its underlying algorithms. Therefore, our approach contributes to the research on algorithm audits for search engines (Sandvig et al., 2014). In contrast to other bias quantification approaches, our methodology is datadriven and does not require a normative reference scheme against which a bias is measured (e.g., by defining concepts of sexism, political leaning or offensive speech).

The paper is structured as follows: In Section 2, we give a literature overview on query suggestions, the topics of biases in search engines and the impact on the search and decision processes of users. In Section 3, we describe the general framework for the identification of biases. It consists of data acquisition, preprocessing, and

analysis modules. In the following Section 4, we apply the framework on searches for German politicians before the federal election in 2017. In Section 5, we discuss the results of the bias and stability analysis and conclude in Section 6.

2. Literature Study

Search engines support their users by suggesting possible completions of their partially typed queries. Typing longer queries with more keystrokes increases the possibilities for inaccurate query formulations or typos. Therefore, the goal of the suggestions is first, to speed up users' searches by reducing the number of typing interactions and second, to avoid the wording problem. Usually, these suggestions are shown as drop-down menus or are inline suggestions next to the top results of the search engine result page (SERP). Different types of suggestions are possible, ranging from a single word or *term suggestion* to full *query suggestions* or *query auto completions* (Chen, 2011, Cai and de Rijke, 2016).

In today's web-based systems, algorithms generate suggestions by analyzing the query log files to find the most popular queries. Although query suggestions are available on all major platforms, details on how search engines implement the suggestions are not available. Google describes their query "prediction" method as being based on the frequency with which other users have searched for the suggested term, the location and language of the user and the popularity of recent trending searches (Sullivan, 2018, Mysen and Schwartz, 2011).

Different studies based on the AOL logs showed the positive effects of query suggestions on the search process like fewer empty result sets, fewer query drifts due to a more precise query wording of using the terms of the corpus, and a generally quicker search process (Pass et al., 2006). Kato et al. (2013) showed that users rely on query suggestions "(1) when the original query is rare, (2) when the original query is a single-term query, (3) when query suggestions are unambiguous, (4) when query suggestions are generalizations or error corrections of the original query, and (5) after the user has clicked on several URLs in the first search result page".

Query suggestions are specially designed to support users that are uncertain in their information need, and that find themselves in a so-called Anomalous State of Knowledge (ASK) (Belkin et al., 1982). In this state, they are uncertain of their own genuine information need or the right way to formulate and verbalize it. In retrieval settings where users are less familiar with the given search topic, they tend to consult external knowledge representation systems (Hienert et al., 2011). In addition to this, Kelly et al. (2009) points out that query suggestions seem particularly important in cases where users are searching for topics on which they have little knowledge or familiarity. As query suggestions are meant to support the uncertain user, the question arises if a biased suggestion mechanism can spoil this uncertainty.

Friedman and Nissenbaum (1996) were among the first who studied the phenomenon of biases in computer systems in general. They defined bias as systematic and unfair discrimination against (groups of) individuals. The process is unfair if those individuals receive a negative and inappropriate outcome. It is systematic if it does not occur randomly. For the measurement of biases in online retrieval systems, bias is defined as a systematic, inclusion, exclusion, or prominence in the selection or the content of the retrieved documents (Introna and Nissenbaum, 2000, Mowshowitz and Kawaguchi, 2002). In a narrower statistical sense, errors in the estimation or sampling process cause a systematic deviation from the true unknown distribution (Baeza-Yates, 2018). It is important to note that biases can only be relatively measured as a systematic deviation from a reference that is a fair norm or ideal distribution.

The many forms of biases that exist in the literature are due to the different underlying references utilized. Mowshowitz and Kawaguchi (2005) approximates the true distribution by combining the output of several comparable retrieval systems (e.g., a set of other web search engines). *Coverage-bias* measures the share of web pages that are indexed by a particular search engine compared to an independent web crawl (Vaughan and Thelwall, 2004). Fortunato et al. (2006) study the *popularity-bias*, the assumption that search engines amplify the popularity of all-ready well-known pages. Here, a website's traffic is compared to a theoretically deduced traffic that would occur without the impact of search engines. *Topical bias* is concerned with the actual content of retrieved web documents (Pitoura et al., 2018). Kulshrestha et al. (2018) analyzes the political leaning of a ranked retrieval set against an independently crawled set that represents the ground truth. *Political bias* of search engines was also measured by linking the URLs from a ranked list of documents to social media utterances of Twitter users whose past voting decision was known (Robertson et al., 2018).

These studies focus on biases in search engine result pages, and less research exists on slanted query suggestions. Noble (2018) collects and reports many cases of sexism and racism in search suggestions. Yenala et al. (2017) uses supervised machine learning methods to detect inappropriate suggested queries. The success of the method relies on the quality and quantity of manually labeled training examples. As Hiemstra (2017) points out, there are adverse effects that complement the general positive image of query suggestions.

Our framework, presented in the next section, detects systematic topical biases in query suggestions related to a set of search terms with common attributes. The topics are derived from the data itself in an unsupervised manner. Systematic biases are then measured as deviations from the distribution of topics. On an abstract level, this approach is comparable to the bias studies in Mowshowitz and Kawaguchi (2005). If groups of search terms with similar characteristics deviate from the overall topical distribution, this group is reported as having biased search suggestions. As we do not impose any prior concept of fairness or ideal distribution, our methods should only be applied in an exploratory manner. It is the researcher's responsibility to interpret the topics derived from the data and the reported biases afterward.



Figure 2. Bias identification framework with the three modules for acquiring, processing and analyzing query suggestions from different web search engines. The analytic procedures in the last module address the research questions RQ1-RQ3 as proposed in section 1.

Manipulated and biased rankings can have effects on the user. Otterbacher et al. (2018) look at gender biases in image search and how the users' perception can reinforce stereotypes. In their well-known experiments, Epstein et al. (2017) present manipulated rankings of political result pages and measure changes in the voting preferences of the users. The effect of manipulated search engines was also confirmed for medical-related searches in the context of vaccination beliefs (Allam et al., 2014). In this paper, we neglect the users' perspective and focus on the task of bias detection and measurement.

3. Automatic Identification of Biases in Web Query Suggestions

This section describes the framework to detect and analyze potential biases in query suggestions (see Figure 2). It consists of three modules that can be applied to different use cases. In this prototype, we relied on text mining libraries for German language corpora, but the different procedures can be easily modified to address other languages as well. In the following, we will describe each of the three modules.

3.1. Module 1: data acquisition

The bias analysis is based on a set of initial search terms t_i with i = 1, ..., N which share a set of P meta-attributes $\{x_{i,1}, \ldots, x_{i,P}\}$. For example, t_i can be the name of a prominent personality with some meta-attributes such as *age*, *gender* or *country* of birth. A specialized web crawler was developed to fetch the query suggestions related to the search terms over a longer period. It is capable of systematically browsing the auto-complete API of three different search engines: Google, Bing, and DuckDuckGo. This is done by sending an HTTP request with the search term to each search engine twice per day.

The HTTP response of each search engine API delivers a list of query suggestions with a different number of items typically ranging from 4 to 10. Both the original

session id	queryterm	datetime	suggestterm	position
2086154	Angela Merkel	2017-08-03 05:46:17	biography	0
2086154	Angela Merkel	2017-08-03 05:46:17	husband	1
2086154	Angela Merkel	2017-08-03 05:46:17	young	2
2086154	Angela Merkel	2017-08-03 05:46:17	trump	3
2086154	Angela Merkel	2017-08-03 05:46:17	meme	4
2086154	Angela Merkel	2017-08-03 05:46:17	eye roll	5
2086154	Angela Merkel	2017-08-03 05:46:17	twitter	6
2086154	Angela Merkel	2017-08-03 05:46:17	photo	7
2086154	Angela Merkel	2017-08-03 05:46:17	snl	8
2086154	Angela Merkel	2017-08-03 05:46:17	and donald trump	9

Table 1. Illustration of the query suggestions data set retrieved by the DuckDuckGo API.

query term and the corresponding query suggestions are stored in a database. A script parses the raw request data and extracts the suggestions for further processing. Most of the suggestions are of the form {<search term>+<suggestion>}. Here, the search term entered into the input field is similar to the prefix of the suggestion. In this case of true auto-completion, the search term was stripped off. An example of the parsed suggestions in the database for the term "Angela Merkel" and from the English DuckDuckGo API is shown in Table 1.

3.2. Module 2: data preprocessing

The second module of the framework deals with cleaning the suggestions, with the problem of mapping single query suggestion terms to a base word form and with the construction of a high dimensional word vector that represents the semantic meaning of the query term. This allows for the grouping of stand-alone words into semantic clusters.

The suggestions are cleaned from special characters, numbers, and German umlauts. In order to tackle the word form problem, we rely on lemmatization. The number of existing German implementations for lemmatization is limited, and we used the text analysis module *pattern.de* (Gesmundo and Samardžić, 2012) that contains a built-in parsing function. It annotates each word with its base form. The language model the parser is built on is described in Schneider and Volk (1998).

The process of vector transformation implies taking a corpus of text as input and building a vector space of several hundred dimensions. Each unique word in the input is assigned a corresponding vector in that space. Řehůřek and Sojka (2010) provide a Python re-implementation of the *word2vec* model (Mikolov et al., 2013) that is used in this module. We relied on the "GermanWordEmbeddings" toolkit as a pre-trained model. Its corpus included more than 600,000,000 words and was

trained with the German Wikipedia and news articles written in German (Mueller, 2015).

3.3. Module 3: data analysis

3.3.1. Cluster analysis

After transforming each query suggestion into a high dimensional numeric word vector, an unsupervised cluster analysis was applied. For words with a similar meaning, the distance between their word vectors is small. Therefore, by clustering the vectors, semantically similar words are grouped. Clustering enables the identification of topical categories in the search terms' suggestions. The k-means clustering algorithm was chosen here (Kanungo et al., 2002). The number of clusters K is determined by relying on established heuristics such as the total within-cluster variation (Pelleg and Moore, 2000), the Silhouette or the Calinski-Harabaz score (Rousseeuw, 1987).

For each initial search term t_i with n_i unique query suggestions, we can sum up the relative number of its suggestions belonging to each cluster topic. This provides a measure of how well each topic is represented in the suggestions for this search term. Hence, consider the cluster score $y_{i,k}$ with $k \in \{1, \ldots, K\}$. It is defined as $y_{i,k} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(s_{i,j} = k)$, where $s_{i,j}$ is the topical category of the j^{th} suggestion and the i^{th} search term. Since the cluster score is larger for search terms with more query suggestions, we normalized the score by the term's total number of query suggestions n_i . This results in a relative cluster score in the range $y_{i,k} \in [0, 1]$.

3.3.2. Regression analysis

The proportional cluster scores represent the distribution of a search term's suggestions over the topical categories. Significant differences in the cluster distributions for a group of search terms can indicate possible systematic biases. For example, in the case of searches for the names of prominent personalities, the cluster analysis could reveal two clusters: A cluster with suggestions related to family members (e.g. *mother, brother, daughter, ...*) and a cluster related to social media and other online resources (e.g. *wiki, twitter, instagram, ...*). Suppose that compared to males, female personalities had a significantly higher score for the first cluster, meaning that more suggestions from the first topic show up if users search for them online. Our framework is capable of discovering these kinds of systematic biases.

To detect significant deviations in the cluster scores regarding the metaattributes of the search terms, a multiple regression analysis was performed. The cluster scores for one cluster $y_{i,k}$ over all search terms form the independent variable and the meta-attributes $x_{i,p}$ are the input to the model. Since we were only interested in detecting broad systematic patterns we relied on the standard linear regression model for the ease of interpretation (Hastie et al., 2001). Formally, for some clusters $k \in 1, \ldots, K$ the model can be represented by the regression equation (Equation 1): An Investigation of Biases in Web Search Engine Query Suggestions 9

$$y_{i,k} = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_P x_{i,P} + \epsilon_i, \tag{1}$$

where ϵ_i is the independent and identically distributed error term and $i = 1, \ldots, N$ are the observations indices. A significant positive or negative coefficient β_p then indicates a biased topical distribution for this cluster and the attribute $x_{i,p}$.

3.3.3. Stability analysis

The cluster scores in the regression analysis are calculated by considering the unique query suggestions over the entire observation period. By relying on such a crosssectional analysis, every query suggestion is assigned an equal weight. A suggestion that appears only once during the observation period is assumed to show the same importance as a suggestion which is prevalent most of the time. To see whether the time independence assumption is reasonable, it is important to check for the degree of stability of the rankings.

As a supplement to the bias identification framework, we separately analyze the long-term ranking stability of the query suggestions gathered in module 1. This is done by comparing each term's suggestion ranking at each time-point with its earliest ranking in the sample and its respective preceding ranking. Formally, rankings of query suggestion lists are incomplete, top-weighted, and indefinite. The importance of the items decreases with the depth of the list. To compare the ranking of two lists, we rely on the rank-biased overlap (*RBO*) statistics (Webber et al., 2010). It is capable of measuring the similarity of two indefinite and possibly nonconjoint rankings. The statistics can take values between 0, indicating no similarity, and 1 for identical lists. The degree of top-weightiness, hence the importance of items at the start of the list, can be controlled by the parameter p. For $p \to 0$ the first item in both lists gets assigned all the weight while for $p \to 1$ all list items become equally important.

4. The Search for Politicians – The Case of the German Bundestag in 2017

In the previous section, we described the framework for exploratory bias detection on an abstract level. Here, we apply it to the names of German politicians and their respective suggestions in web search engines. The underlying data is archived in Zenodo (Samokhina et al., 2019).

Political activity on the web is already generating massive amounts of data, such as individuals' political conversations, donations, online formats of news and politics, political blogs, online public speeches and openly available information about political activities and involved individuals. Hence, the search for politicians is a common task as citizens might inform themselves about a candidate before elections. However, as most persons with publicity, politicians do not always become



Figure 3. Venn diagram with the sizes of the sets of overlapping and unique search terms for each of the three search engines. The total number of unique search terms is 10,357.

the object of public interest only because of their primary work, but due to other aspects such as their private life, family affairs or fashion style and appearance.

4.1. Data acquisition and processing

First, the data acquisition phase was initiated by collecting the names of all politicians of the 18th German federal parliament (German Bundestag) in the constitution of late 2016. Additional socio-demographic information for each politician, such as the age, the home town, and the party affiliation, was collected too. This information corresponds to the framework's meta-attributes, for which a possible bias can be detected. The data set included 36% female politicians. The average age was 54. Most members of the parliament originated from the German federal state of North Rhine-Westphalia (22%), which is the largest populated, while the smallest amount of members of the parliament originated from the German federal state of Bremen (1%). The biggest party in the sample was the CDU ("Christian Democratic Union") with a proportion of 40%. Since there was only one member in the group "non-attached member" in the *party* category, the politician of this group was excluded from further analysis.

Searches for Bing and Google were carried out using the German language setting. English settings have been used in the DuckDuckGo search engine. The data was crawled for four months between 2017/02/13 and 2017/06/19. The data set of query suggestions included more than 10,000 unique suggestions from all three search engines. Regarding the query suggestions for the politicians, we observed some differences between the search engines (see Figure 3): First, the number of unique search terms was largest for Google, indicating a higher diversity in terms

cluster 1	cluster 2	cluster 3
717 suggestions related	1,251 suggestions re-	1,072 suggestions con-
to politics and eco-	lated to personal and	taining location infor-
nomics	emotional topics	mation (cities)
parliamentary office prosecutor minister of transport integration commissioner	pregnant toupet cartoon simple	giessen kassel radeberg rudolstadt
member of bundestag	wife	heiligenhafen
minister	brother	forchtenberg
treasurer	tie	fulda
middle east policy	airplane	moers

Table 2. Selected query suggestions from each of the three clusters identified by the k-means procedure. The terms were translated from German to English.

for this provider. Secondly, the three search engines had all small overlapping sets of suggestions with only 85 unique suggestion terms shared by all three engines. Therefore, by combining the three different search engines and their query suggestions, we could create a larger and diverse data set with more unique suggestions.

After cleaning and lemmatization, 3,824 unique single word suggestions were leftover, and the vector-transformation algorithm could identify 3,040 words. Word embedding vectors for each of the suggestions were used to perform a cluster analysis. The heuristics suggested a cluster size of K = 3. By manually evaluating the clusters, we assigned a label that best describes the topic of each cluster (see Table 2).

The cluster analysis provided useful insights into the topical distribution of the search suggestions. The first cluster contained suggestions referring to political and economic topics. These suggestions were related to the formal position and to trending political news events during that time. The second cluster included many suggestions related to personal and emotional topics, including references to tabloid journalism. The final cluster solely consisted of names of cities and regions. We assume, they possibly refer to the politicians' electoral districts and the appearances in election campaigns.

4.2. Bias identification

The cluster assignments were used to calculate the topical distribution for each politician. We then performed a regression model on the proportion of unique suggestions belonging to cluster 1 and cluster 2. Cluster 3 containing German geo-

graphical locations did not promise any insights, as users searching for geographical locations in combination with the names of politicians could have been querying for the place of birth, the electoral district or other topics that are out of reach without the specific context. Therefore, we focussed further investigations on clusters 1 and 2 only. For groups with similar socio-demographic characteristics, we tested whether the group status had a significant positive or negative effect on the respective topical distribution. In this case, our framework reports a bias for this group (compare to section 3.3.2).

Table 3 shows the regression results for cluster 2, which consists of query suggestions related to personal and emotional topics, and cluster 1, which comprises suggestions related to politics and economics. We found that the model for cluster 1 was not statistically significant. Since its p-value was quite large, the F-test could not reject the joint null hypothesis that all coefficients are zero.

Hence, we focused on the significant second model, which predicts the proportion of query suggestions belonging to the cluster with terms related to personal and emotional topics. The base category for the attribute *constituency* is "Baden-Württemberg". The base category for the attribute *party* is "CDU". As an illustration, consider a 40-year-old female politician from "Baden-Württemberg" belonging to the party "CDU": This model predicts an average cluster value of $0.353 + 40 \times (-0.002) = 0.273$. Hence, approximately 27% of the unique query suggestions for this hypothetical politician belong to the cluster associated with personal and emotional terms.

We found that both the gender and age, are negatively associated with the amount of personal and emotional query suggestions. Compared to an average female politician of the Bundestag, a male politician had on average 2.4 percentage points fewer suggestions in cluster 2. Neglecting the fact that some suggestions were more visible than others and assuming that the performance of the word-embedding method works equally well for all types of words, this indicates that users of search engines were slightly more exposed to topics related to personal and emotional terms in combination with a female politician. Nevertheless, the effect is rather small.

Roughly, the same negative effect can be seen for the age of a politician: Keeping the gender or other socio-demographic variables constant, on average, the number of queries belonging to cluster 2 decreases by 2 percentage points by steps of 10 years. The most substantial effect was found regarding the constituency of the politician: A politician from the federal state of Bremen has 11 percentage points fewer suggestions in cluster 2 compared to the base category. However, as this state is the smallest in Germany, only 6 politicians were included in our sample. Finally, we also found a positive effect for the parties "Die Grünen" (The Green Party) and "CSU" (The Christian Social Union in Bavaria). Compared to a politician from the "CDU", on average, a politician from "Die Grünen" had roughly 4 percentage points more suggestions in cluster 2.

Table 3. Linear Regression model for fitting the proportion of the politicians' query suggestions belonging to cluster 1 or cluster 2. Shown are the coefficients along with the significance value of the test for non-zero coefficients, the F-test for overall significance and the adjusted R^2 measure. Statistically significant values (p < 0.05) are marked bold.

	Dependent variable				
	Cluster 1: Politics and Economics		Cluster 2: Personal and Emotional		
Independent variable	Estimate	P-value	Estimate	P-value	
β_0	0.242	0.000	0.353	0.000	
Gender Male	-0.009	0.371	-0.024	0.025	
Age	0.000	0.396	-0.002	0.000	
State Baden-Württemberg	reference category				
State Bayern	0.008	0.720	-0.033	0.198	
State Berlin	0.001	0.968	0.032	0.244	
State Brandenburg	-0.025	0.385	-0.060	0.064	
State Bremen	-0.033	0.483	-0.111	0.035	
State Hamburg	-0.004	0.904	0.011	0.767	
State Hessen	-0.022	0.303	-0.034	0.145	
State Mecklenburg-Vorpommern	0.079	0.023	-0.071	0.065	
State Niedersachsen	-0.001	0.968	-0.002	0.914	
State Nordrhein-Westfalen	0.000	0.997	0.005	0.770	
State Rheinland-Pfalz	-0.004	0.874	-0.001	0.967	
State Saarland	-0.047	0.226	-0.060	0.169	
State Sachsen	-0.024	0.297	-0.016	0.526	
State Sachsen-Anhalt	-0.002	0.946	-0.038	0.231	
State Schleswig-Holstein	0.028	0.281	-0.026	0.373	
State Thüringen	-0.008	0.792	-0.019	0.556	
Party CDU		reference category			
Party CSU	-0.019	0.463	0.057	0.048	
Party DIE LINKE	0.010	0.531	-0.009	0.635	
Party GRÜNE	0.015	0.350	0.039	0.033	
Party SPD	-0.006	0.598	-0.004	0.734	
F-statistic	0.865	0.638	2.944	0.000	
Adjusted R^2	0.000		0.0	0.061	
Degrees of freedom	605				

While our proposed framework showed a possible way of identifying biases, we could only find weak effects in this specific use case. The overall level of variance was significant: The R-squared statistics showed that only around 6% of the variation



Figure 4. The colored lines represent the three day average for the politician's stability statistics (RBO value). The dashed lines indicate the crawling period for the suggestions used in the section on bias identification. The shaded areas are a standard deviation margin around the mean.

in the dependent variable was explicable by the independent model variables for the second cluster. While there were some interesting significant effects, e.g., for the gender of the politician, the effect size was rather small. The model for the first cluster showed no overall significance at all.

4.3. Stability of rankings over time

In the last analysis step, we tested the long-term stability of the query suggestions in Google for the German politicians over time as mentioned and introduced in section 3.3.3. For each measurement date, we calculated the rank-biased overlap similarity of each ranking with its first and its previous ranking in the sample. The former gives insights into the overall stability, whereas the latter approximates a rate of change of the suggestions. The median of the suggestions for the lists was 9 with a maximum of 10 list entries. Therefore, for the calculation of the RBO score, we chose the weighting parameter to be p = 0.90. Doing so puts most of the weight (83%) to the first nine list items and results in an expected evaluation depth of 10 items as shown by Webber et al. (2010).

Figure 4 shows the aggregated three-day average over the rankings of all politicians. The shaded area indicates the variability of the data in the form of a standard deviation margin around the mean. The overall similarity decreased over time. Nevertheless, after one year, the average similarity roughly converged to $\overline{RBO} = 0.45$. In early 2018, each ranking compared to the ranking at the beginning still had around 45% of their suggestions in common. As the figure shows, the rankings seemed to change faster at the beginning of the observation period. After July 2017, the RBO decreased only slowly from around 0.5 to 0.4. This finding indicates that some query suggestions were much more prevalent, while other suggestions fluctuated and exchanged more often. The successive rate of change was small, having a stability value of $\overline{RBO} = 0.98$ on average. This rate stayed relatively constant, despite some negative shocks at several time points. Possibly, these temporal shocks occurred due to political events or changes in the Google query suggestion algorithms.

5. Discussion

This work is the first that systematically investigated biases in web query suggestions for person-related searches. We proposed an analytic framework to identify possible biases in a data-driven and exploratory way and applied it to the case of web searches for German politicians. The modularity of the presented framework simplifies possible extensions and improvements of specific procedures. Based on the research questions formulated in the introduction, in this section, we will discuss the outcomes and limitations of our work and suggest starting points for further research.

5.1. Topical clustering of query suggestions

RQ 1: How can we automatically identify clusters and patterns in web query suggestions for person-related searches?

The query suggestions are grouped into topical clusters by first transforming single words into a high dimensional word vector that represents the semantic meaning of the word. Second, a clustering algorithm classifies geometrically close vectors, and therefore semantically similar words together. Future research can work on the improvement of the utilized word embedding methodology. This includes the implementation and evaluation of other embedding models such as *fastText*, which can also process unseen and misspelled words not included in the training data (Bojanowski et al., 2016). In the current framework, only single word suggestions are clustered. By averaging or summing up the corresponding word vectors for longer query suggestions, it is possible to derive a vector representation for complete queries (Zamani and Croft, 2016).

In the experiment on searches for the names of German politicians, the cluster analysis identified three distinct groups of semantically similar query suggestions. The three clusters were related to formal topics from politics and economics to more personal and emotional issues, and the names of cities and regions in Germany. Thus, regarding the first research question, for this case, our methodology has been proven successful. Further experiments are needed to evaluate the general validity

and applicability of the clustering methodology. Promising use cases might be the search for prominent personalities, well-known business persons, and politicians participating in other national elections. For the clustering itself, we determined the number of topics by relying on several data-driven statistical heuristics. This might not always be the ideal approach if the use case or specific research questions can provide a more reasonable number of topics.

5.2. Identification of biases

RQ 2: To what extent can metadata on the persons searched (e.g., gender, age, party membership) be used to explain possible biases?

As argued in Section 2, a bias measures a systematic deviation from an ideal or fair distribution. In the case of retrieval systems, this can be the deviation from a somehow theoretically derived fair ranking, or from an ideal retrieval set that, for example, does not contain cases of inappropriate language, hate speech or racism. The challenge here is to define and find this ideal norm.

In our framework, we study systematic topical biases in query suggestions for groups of search terms that share similar characteristics. For instance, this can be the gender in searches for persons' names. The topics are derived from the query suggestions by using the exploratory cluster analysis. Each suggestion is labeled with a topic, and for each search term, the relative number of unique suggestions in each topical category is summed up. For specific groups of search terms, biases are then measured as significant deviations from the overall distribution of topics. Thus, the outcome of the framework is a set of statistically significant estimators that indicate a systematic deviation for particular groups.

Whether these deviations are unfair in a theoretically justified way cannot be answered by this analysis and is subject to the researcher's interpretation. However, the proposed procedures can be applied to various search terms and are not restricted to specific use cases. They support the researcher in spotting broad biases hidden in a large amount of data.

In our experiment, the regression analysis found some weak biases for female politicians as they receive a slightly higher proportion of query suggestion terms related to personal and emotional topics (cluster 2). Among others, another biasing factor that appeared in the data was the age. The older the politicians are, the fewer personal query suggestion terms appeared. Due to the main focus on the German Bundestag, these results are limited as we used 629 different names of politicians to form our query suggestions data set. Considering the overall weak explanatory power of our regression results, the second research question $RQ \ 2$ remains an open issue. Future research should try out more use cases and test other methods of statistical inference. For example, beta regression models are designed for the modeling of proportions and frequencies and might perform better in this case (Cribari-Neto and Zeileis, 2010).

5.3. Persistence of biases

RQ 3: How can we measure and analyze the stability and persistence of possible biases in the suggestions?

The bias analysis relies on a cross-section of unique search suggestions. Hence, any time dependencies are neglected here. Especially for prominent search terms, the generation of query suggestions is highly dynamic. As the retrieved lists of suggestions might change fast, future work should incorporate these dynamic effects into the analysis. For example, for each suggestion, the visibility in the ranking measured in days can be calculated. These statistics can be used as a weight when calculating the relative cluster scores such that prevalent items achieve a more significant influence.

Considering these dynamics, an important question is, whether biases in the ranking remain stable over time. In the experiment, the aggregated measures of stability showed that the rate of change in the rankings stayed relatively constant. After one year, comparing each ranking to its counterpart from the earliest time point, the rankings still showed an average similarity of roughly 45%. There seems to be a general pattern in the data: While some trending suggestions only appeared for a short period, other suggestions were present for most of the time. It seems that these trending suggestions are often related to news events about the person searched. Studying which kind of news trigger enough searches such that the suggestion algorithms depict them might help in understanding how biases emerge online. The terms' search volume might be an essential factor in explaining the variety of long-term persistence observed. For example, when looking at the data from the experiment, it seems that well-known politicians such as Angela Merkel have a larger number of unique suggestions and more variations over time compared to other, more unknown politicians.

6. Conclusion

The discussion about biases is not purely academic but has got a real impact on searchers. By measuring search engine bias with retrieval measures, it was shown that fairer and less biased search systems tend to perform better (Wilkie and Az-zopardi, 2017). The design of current search engines requires users to be aware of possible biases. They must submit specific queries and thoroughly assess the result pages. As long as search engines are commercial black-box systems, this situation is unlikely to change. Throughout this article, we emphasized that research, focusing on algorithmic biases in retrieval systems, should consider the case of query suggestions. They already steer and influence the decision-making process during the early phase of query formulation. Therefore, we support the demand for a free web index (Lewandowski, 2015) and wish for services and tools that simplify the study of query suggestion mechanisms. However, as a free web index is not in sight, monitoring and auditing of web search engines and their query suggestions are required to identify potential biases.

18 REFERENCES

We see our work as a first step into a systematic, algorithmic audit of query suggestions in web search engines. In a case study, we studied biases in query suggestions when searching for the names of politicians. We presented a framework, which is capable of automatically collecting and analyzing query suggestions for large sets of search terms over a long period. The suggestions are grouped into topical clusters, and possible biases can be identified. Nevertheless, it is the researchers' responsibility to interpret the topics derived from the data and the reported biases afterward.

References

- Allam, A., Schulz, P. J. and Nakamoto, K. (2014), 'The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output', *Journal of Medical Internet Research* 16(4:e100).
- Baeza-Yates, R. (2018), 'Bias on the web', Communications of the ACM **61**(6), 54–61.
- Belkin, N., Oddy, R. and Brooks, H. (1982), 'ASK FOR INFORMATION RE-TRIEVAL: PART I. BACKGROUND AND THEORY', Journal of Documentation 38(2), 61–71.
- Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2016), 'Enriching word vectors with subword information', arXiv preprint arXiv:1607.04606.
- Cai, F. and de Rijke, M. (2016), 'A Survey of Query Auto Completion in Information Retrieval', Foundations and Trends® in Information Retrieval 10(4), 273–363.
- Chen, L. (2011), 'Term suggestion with similarity measure based on semantic analysis techniques in query logs', *Online Information Review* **35**(1), 9–33.
- Cribari-Neto, F. and Zeileis, A. (2010), 'Beta regression in r', Journal of Statistical Software 34(2).
- Epstein, R. and Robertson, R. E. (2015), 'The search engine manipulation effect (seme) and its possible impact on the outcomes of elections', *Proceedings of the National Academy of Sciences* 112(33), E4512–E4521.
- Epstein, R., Robertson, R. E., Lazer, D. and Wilson, C. (2017), 'Suppressing the Search Engine Manipulation Effect (SEME)', Proceedings of the ACM on Human-Computer Interaction 1(CSCW), 1–22.
- Fortunato, S., Flammini, A., Menczer, F. and Vespignani, A. (2006), 'Topical interests and the mitigation of search engine bias', *Proceedings of the National* Academy of Sciences of the United States of America 103(34), 12684–12689.
- Friedman, B. and Nissenbaum, H. (1996), 'Bias in Computer Systems', ACM Trans. Inf. Syst. 14(3), 330–347.
- Gesmundo, A. and Samardžić, T. (2012), Lemmatisation as a tagging task, *in* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2', Association for Computational Linguistics, pp. 368–372.

Harary, A., Bersoff, D. M., Adkins, S., Bruening, J., Lvovich, S., Bonifaz, G. and Heume, K. (2017), 2017 Edelman TRUST BAROMETER - Global Results, Business, Edelman Intelligence. (accessed 2019-09-24).

URL: https://www.slideshare.net/EdelmanInsights/2017-edelman-trustbarometer-global-results-71035413

- Hastie, T., Tibshirani, R. and Friedman, J. H. (2001), The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations, Springer series in statistics, Springer, New York.
- Hiemstra, D. (2017), 'Query autocompletions considered harmful'. (accessed 2019-09-24).

URL: http://searsia.org/blog/2017-02-09-autocomplete/

- Hienert, D., Schaer, P., Schaible, J. and Mayr, P. (2011), A novel combined term suggestion service for domain-specific digital libraries, in S. Gradmann, F. Borri, C. Meghini and H. Schuldt, eds, 'Research and Advanced Technology for Digital Libraries International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011. Proceedings', Vol. 6966 of Lecture Notes in Computer Science, Springer, pp. 192–203.
- Introna, L. D. and Nissenbaum, H. (2000), 'Shaping the web: Why the politics of search engines matters', *The Information Society* 16(3), 169–185.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y., Member, S. and Member, S. (2002), 'An efficient k-means clustering algorithm: Analysis and implementation', *IEEE Transactions on Pattern Analysis* and Machine Intelligence 24(7), 881–892.
- Kato, M. P., Sakai, T. and Tanaka, K. (2013), 'When do people use query suggestion? A query suggestion log analysis', *Information Retrieval* 16(6), 725–746.
- Kelly, D., Gyllstrom, K. and Bailey, E. W. (2009), 'A comparison of query and term suggestion features for interactive searching', *Proceedings of the 32nd in*ternational ACM SIGIR conference on Research and development in information retrieval pp. 371–378.
- Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P. and Karahalios, K. (2018), 'Search bias quantification: investigating political bias in social media and web search', *Information Retrieval Journal* pp. 188 – 227.
- Larcinese, V. and Miner, L. (2017), The Political Impact of the Internet on US Presidential Elections, STICERD - Economic Organisation and Public Policy Discussion Papers Series 63, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE. (accessed 2019-09-24).

URL: https://ideas.repec.org/p/cep/stieop/63.html

- Lewandowski, D. (2015), 'Living in a world of biased search engines', Online Information Review 39(3).
- Matsa, K. E. and Lu, K. (2016), 10 facts about the changing digital news landscape, Technical report, Pew Research Center. (accessed 2019-09-24).
 - URL: http://www.pewresearch.org/fact-tank/2016/09/14/facts-about-the-

20 REFERENCES

changing-digital-news-landscape/

- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), 'Efficient estimation of word representations in vector space', CoRR abs/1301.3781.
- Mowshowitz, A. and Kawaguchi, A. (2002), 'Assessing bias in search engines', Information Processing & Management 38(1), 141–156.
- Mowshowitz, A. and Kawaguchi, A. (2005), 'Measuring search engine bias', Information Processing & Management 41(5), 1193–1205.

Mueller, A. (2015), 'GermanWordEmbeddings'. (accessed 2019-09-24). URL: http://devmount.github.io/GermanWordEmbeddings/

Mysen, C. C. and Schwartz, S. E. (2011), 'Dynamic query suggestion'. Patent number US8027990B1. (accessed 2019-09-24).

URL: https://patents.google.com/patent/US8027990B1/en

- Noble, S. U. (2018), Algorithms of Oppression: How Search Engines Reinforce Racism, 1 edition edn, NYU Press, New York.
- Otterbacher, J., Checco, A., Demartini, G. and Clough, P. (2018), Investigating User Perception of Gender Bias in Image Search: The Role of Sexism, *in* 'The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18', ACM Press, Ann Arbor, MI, USA, pp. 933–936.
- Parkinson, H. J. (2016), 'Click and elect: how fake news helped Donald Trump win a real election | Hannah Jane Parkinson', The Guardian. (accessed 2019-09-24).
 URL: https://www.theguardian.com/commentisfree/2016/nov/14/fake-newsdonald-trump-election-alt-right-social-media-tech-companies
- Pass, G., Chowdhury, A. and Torgeson, C. (2006), A Picture of Search, in 'Proceedings of the 1st International Conference on Scalable Information Systems', InfoScale '06, ACM, New York, NY, USA.
- Pelleg, D. and Moore, A. (2000), X-means: Extending k-means with efficient estimation of the number of clusters, *in* 'In Proceedings of the 17th International Conf. on Machine Learning', Morgan Kaufmann, pp. 727–734.
- Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S. and Weikum, G. (2018), 'On Measuring Bias in Online Information', ACM SIGMOD Record 46(4), 16–21.
- Řehůřek, R. and Sojka, P. (2010), Software Framework for Topic Modelling with Large Corpora, *in* 'Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks', ELRA, Valletta, Malta, pp. 45–50.
- Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D. and Wilson, C. (2018), 'Auditing Partian Audience Bias within Google Search', *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW), 1–22.
- Rousseeuw, P. (1987), 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', J. Comput. Appl. Math. 20(1), 53–65.
- Samokhina, A., Bonart, M., Schaer, P. and Heisenberg, G. (2019), 'Query autocompletions for German politicians of the 18th Bundestag'. (accessed 2019-09-24).

URL: *https://doi.org/10.5281/zenodo.3462046*

- Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. (2014), 'Auditing algorithms: Research methods for detecting discrimination on internet platforms', *Data and discrimination: converting critical concerns into productive inquiry* 22.
- Schiffmann, B. (2008), 'The Reason for the Obama Victory: It's the Internet, Stupid', *WIRED*. (accessed 2019-09-24).

URL: https://www.wired.com/2008/11/the-obama-victo/

- Schneider, G. and Volk, M. (1998), Adding manual constraints and lexical lookup to a brill-tagger for german, *in* 'Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation, Saarbrücken'.
- Sullivan, D. (2018), 'How google autocomplete works in search'. (accessed 2019-03-13).

URL:

https://blog.google/products/search/how-google-autocomplete-works-search/

- Vaughan, L. and Thelwall, M. (2004), 'Search engine coverage bias: evidence and possible causes', *Information Processing & Management* 40(4), 693–707.
- Webber, W., Moffat, A. and Zobel, J. (2010), 'A similarity measure for indefinite rankings', ACM Transactions on Information Systems 28(4), 1–38.
- Wilkie, C. and Azzopardi, L. (2017), Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable?, in 'Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17', ACM Press, Singapore, Singapore, pp. 2375–2378.
- Yenala, H., Chinnakotla, M. and Goyal, J. (2017), Convolutional Bi-directional LSTM for Detecting Inappropriate Query Suggestions in Web Search, *in J. Kim*, K. Shim, L. Cao, J.-G. Lee, X. Lin and Y.-S. Moon, eds, 'Advances in Knowledge Discovery and Data Mining', Lecture Notes in Computer Science, Springer International Publishing, pp. 3–16.
- Zamani, H. and Croft, W. B. (2016), Estimating Embedding Vectors for Queries, in 'Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval', ICTIR '16, ACM, New York, NY, USA, pp. 123–132.
- Zweig, K. (2017), 'Watching the watchers: Epstein and Robertson's Search Engine Manipulation Effect'. (accessed 2019-09-24).

URL: https://algorithmwatch.org/en/watching-the-watchers-epstein-and-robertsons-search-engine-manipulation-effect/