Measuring personalization in environmental search

Alma Nilsson alma.nilsson@ru.nl
Diego Bartoli Geijo diego.bartoligeijo@ru.nl
Marta Españó López marta.espanolopez@ru.nl
Group 15 - Project 6

1 INTRODUCTION

Search engines try to optimize the ranking of query search results for users through profile-specific personalization, in other words, through basing part of the ranking on previous search sessions. This results in queries posed by different profiles getting different matches, which gives rise to a range of potential problems. The so-called "Filter bubble" problem refers to profiles not being able to access the same information as others due to information being "filtered" by previous searches. Similarly, personalization could lead to information "Rabbit holes" where a profile only finds information that complies with the viewpoint of the user, raising concern about the consumption of biased information and lack of factuality.

Politics is a field where one can expect to find distinct differences in viewpoints, therefore this project will look at the effect of political views on search results regarding climate change.[1] In particular, we will try to answer the following question: does personalization performed by a search engine on user profiles, with different political views, affect the results of queries related to environmental issues?

2 RELATED WORK

Personalization in search has been studied in the literature "Examining Personalization in Academic Web Search" [1] and "Detecting and Visualizing Filter Bubbles in Google and Bing" [2] which provided an understanding of the different techniques for personalization, as well as the various effects search personalization may have on the user's search results, e.g. "filter bubble". Most of the experimental design choices have been based on the study: "Measuring Personalization of Web Search by Anikó Hannaák" [3], which helped to establish the fundamental proceedings related to user profile creation, and to identify the different sources of unwanted noise, as well as the methods and considerations in order to avoid it, such as using VPN, or dividing the conducted search sessions in different times of the day. Also, the statistical tests used in order to quantify the personalization, namely the Jaccard similarity index, and Kendall's Tau rank correlation coefficient.

3 METHOD

Users were divided in 2 groups based on political orientation. The users profiles were trained with different queries meant to represent their political orientation and then differences in the results for the same environmental queries for both groups were measured.
The search engine *Bing* was chosen for this project due to it allowing the implementation of automated logged browsing. *Google Chrome*, which was the original choice of search engine for this project, recognizes automatic browsing and only allows non-logged browsing and thus could not be used.
For each political orientation, left-wing and right-wing, 10 user accounts were manually created. The user was arbitrarily chosen to be a woman by the name of Maria Gómez from Spain born on the 1st of September in 1992. Furthermore 2 sets of queries meant to represent political views of the far left and far right, henceforth referred to as the training set, were written. These were split into 4 groups of 15 queries each, namely *News*, *Lifestyle*, *Humanities* and *Other*. These queries were largely based on the judgment of the group and may not actually be representative of the queries made by right-wing or left-wing people. Additionally, 5 biased and 5 general environmental queries were written for evaluation of the personalization. Biased queries were rhetorical questions, not objective ones. These will be referred to as the testing set.

Custom python scripts [2] were written to run browsing sessions where a pair of users per iteration, one of each political orientation, would make 2 random queries per category. These were chosen with replacement from the training set. Furthermore 1 URL, web address, per query was randomly chosen from the top 50 % of the ranked results. The chosen URL was visited for between 5 and 10 seconds. The queries were sent on the Bing news page, not on the main one since the results of this one were considered to represent better the ones the users would be interested in. For each browsing session, a *.json* file containing information about the user, its queries, and the associated URL was created. The structure of the file is reported in appendix A.

To evaluate the personalization of the users one pre-training, one mid-training, and a post-training environmental browsing session was run on the testing set. This choice was also due to some technical problems explained in section 4. All 10 of the queries in the testing set were made by each user and the top 10 ranked results for each query were saved. Furthermore, for the final environmental session, screenshots of the top-ranked results were saved. For each environmental session, a *.json* file containing information about the user and its queries was created. The structure of the file is reported in appendix B.

In order to measure the personalization acquired from both profiles after the training, two different techniques have been used. The Jaccard similarity index, a statistic used for identifying the similarity and diversity of two sets, it measures the degree to which the different sets have elements in common, relative to the total number of elements in both sets, the formula computed on set A and set B is:

[1] We don't want to foment prejudices in any way. However, we think there are some differences due to political orientation that could be reflected in our Google searches. We are going to try to represent these differences in the most impartial way.

[2] Please find all the files used for the implementation in the GitHub repository https://github.com/Ambivalentalfa/IR_personalization_in_search

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Although this metric provides us with a measure of how similar are the results, it does not take into consideration the ranking of the URLs, so, to account for the rank, Kendall's (Tau) rank correlation coefficient has been used, and it is computed in the following way:

$$\tau = \frac{C - D}{n(n-1)/2}$$

where n is the total number of samples, C is the number of concordant pairs, and D is the number of discordant pairs among the two different sets.

4 EXPERIMENTAL SETUP

All the sessions were run on the same machine, a Dell PC with Windows 10. The ensure Madrid location *Freedome VPN* [3] was used. To interact with Bing *selenium* library from *python* and the chrome driver for windows were used, since Bing was launched on Chrome browser.

We had some trouble identifying which was the perfect structure for our training sessions. The most important problem was that the sessions took a lot of time, from 3 to 5 hours each, and we didn't have a machine that we could use just for that every day. If we try to use the machine for other tasks, while doing the sessions, this often caused connection problems and aborts of the sessions. It took us 3 days of tries to find a viable experimental setup. In the end we set the values indicated in the section 3, and the python *multithread* library was used in order to run two user sessions at the same time. The session duration was reduced to approximately 1 hour, allowing us to complete 2 every day without working problems. The sessions were held one in the morning and one in the afternoon at variable times. We had 3 days of unstructured training and 3 days, for a total of six sessions, of structured training.

Due to these initial technical problems, the first part of the training is not structured and well documented, since most of the sessions didn't come to a positive end. The mid-training environmental browsing session, indicated in the section 3, was carried out right before starting the structured training part so we can compare the differences in the results with no training, with unstructured training, and with structured training. Some exploratory analysis measures on the data collected during the structured training sessions are reported in appendix C.

5 RESULTS AND CONCLUSIONS

To assess how similar are the two differently trained profiles, the Jaccard index and Kendall's coefficient are applied to the resulting files containing the outcome URLs from the environmental queries, the output from the computed statistics can be appreciated in the following table 1. From it, we can observe how after each session, both the Jaccard index and Kendall's coefficient for all the sets increase. For both statistics, the closer the value is to one, the fewer differences have been found between the sets, meaning that over time the search results from the user accounts have become more similar. Although it is not the expected outcome, this demonstrates that the profiles got trained, but not in the manner we expected, as each time, for a given query the users got more similar results instead of more distinct ones.
This behavior of the user's account personalization was also appreciated when the Jaccard statistic was computed with the comparison of the top ten most common URLs from each of the profile types, for each of the three sessions in which the environmental queries were launched. The Jaccard indexes obtained were 0.667, 0.818, and 1.0 for each session respectively. From this, again, it can be seen how the results become more similar until the end when the top ten most common links in both profile types become identical.

	Jaccard	**Kendall's**
Session 1 env RW-LF 1	0.7378	0.3507
Session 2 env RW-LF 2	0.8467	0.5083
Session 3 env RW-LF 3	0.9709	0.8049
Session 1 RW	0.7711	0.3689
Session 2 RW	0.8955	0.6673
Session 3 RW	0.9570	0.7591
Session 1 LW	0.7214	0.3384
Session 2 LW	0.8084	0.4739
Session 3 LW	0.9814	0.82024

Table 1: The first three rows contain the computed statistics for the comparison of the environmental sessions, "Session 1" being the one before training the profiles, "Session 2" the middle one, and "Session 3" the final one, after training the profiles. The following six rows contain an internal comparison among the URLs from the same type of profile but different sessions. RW being right-wing and LW left-wing profile.

The similarity between the initial environmental run and environmental run 2 is larger than the similarity between the initial environmental run and environmental run 3, as can be seen in Figure 1, indicating that there is some manner of personalization going on. This personalization, however, does not seem to be based on the queries. Contrary to the hypothesis made by this project they are changing to become more alike, as can be seen in Figure 2. This could be due to factors that the two groups share being more influential in the personalization of the rankings than that of the queries, at least when measuring personalization within the time span of this project. The users all shared the same name, age, and location. The latter being factors that are generally accepted to be major contributors to personalization. These factors, however, were constant throughout the period of the experiment, and thus other factors such as the IP address, and the time and structure of the browsing sessions might have been the cause of the unified personalization seen in Figures 1 and 2. It is possible that personalization based on queries will take effect after a period of time that is longer than the duration of our experiment. A further possibly influential factor for the personalization might have been that all users were interested in the same main areas, as specified in the *Method* section. So the detail of the query may not play such an important role when personalizing, or not as much as the broader

[3] https://www.f-secure.com/en/home/products/freedome

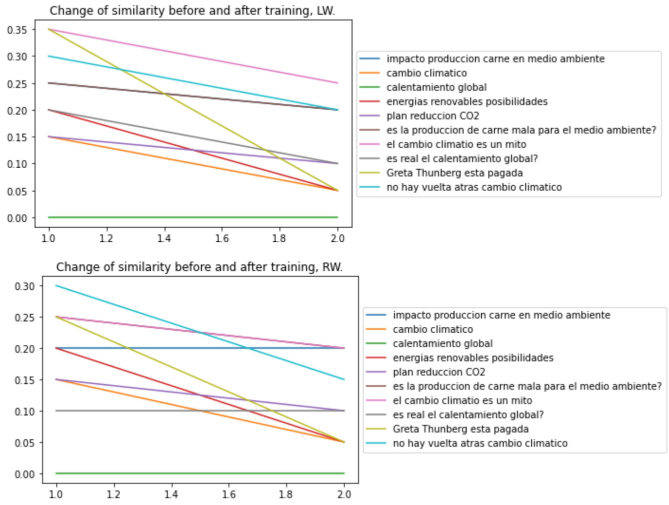

Figure 1: The similarity of environmental run 2 and 3 to the initial environmental run for all queries, for left-wing and right-wing users respectively.

field to which the query belongs would. Lastly, there is the possibility that the results simply mean the hypothesis was wrong. The fact that personalization in general has been previously shown to occur, does not mean that this specific type of personalization exists.

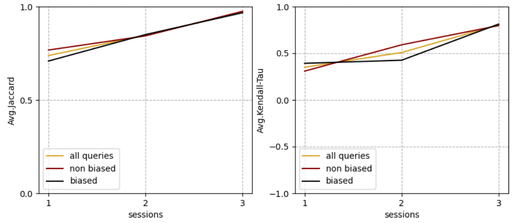

Figure 2: Similarity for left-wing and right-wing users measured by the Jaccard similarity index and the Kendall-Tau index rank correlation coefficient. The figures show the similarity between the biased and unbiased environmental queries, both individually and combined. Find internal similarities in Appendix D

This project leaves room for improvement. The user groups could have included more variation, for example, by adding both male and female users to see if the changes are different between the two sexes. Furthermore, the browsing sessions could have included a larger amount of, and more time spent at, URLs for each query. Lastly, the duration of the experiment could have been greatly expanded as well as including some variation within the time and duration of browsing sessions which most likely would result in more personalized profiles.

REFERENCES

[1] Jia Tina Du Sara Salehi and NY USA 103–111. Helen Ashman. 2015. Examining Personalization in Academic Web Search. In Proceedings of the 26th ACM Conference on Hypertext Social Media (HT '15). Association for Computing Machinery, New York.

[2] Christopher A. Brooks Tawanna R. Dillahunt, Samarth Gulati. 2015. Detecting, Visualizing Filter Bubbles in Google, and NY USA 1851–1856. Bing. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15). Association for Computing Machinery, New York.

[3] Arash Molavi Kakhki Balachander Krishnamurthy David Lazer Alan Mislove Aniko Hannak, Piotr Sapiezynski and NY USA 527–538. Christo Wilson. 2013. Measuring personalization of web search. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). Association for Computing Machinery, New York.

APPENDIX

A STRUCTURE BROWSING SESSION (TRAIN)

```
{"user id": {
  "username": "username value",
  "password": "password value",
  "session id": {
    "query": [ "link visited" ]
    ....
  }
}
 ...
}
```

B STRUCTURE BROWSING SESSION (TEST)

```
{"user id": {
  "username": "username value",
  "password": "password value",
  "session id": {
    "query": [ "links visited" ]
    ....
  }
}
 ...
}
```

C EXPLORATORY ANALYSIS TRAINING DATA

In the table below are reported the number of users with no links visited (for connection problems), the ratio of unique queries searched and the ratio of unique links visited, for each browsing session.

	Session	Wing	Missing val.	Uniq.queries	Uniq.links
1	1	left	0.0	0.9625	0.9625
2	1	right	0.0	0.9875	0.9625
3	2	left	0.2	0.8	0.8
4	2	right	0.1	0.875	0.9
5	3	left	0.0	0.975	1.0
6	3	right	0.0	1.0	1.0
7	4	left	0.0	0.9625	1.0
8	4	right	0.0	0.9875	1.0
9	5	left	0.0	0.975	1.0
10	5	right	0.0	0.9625	1.0
11	6	left	0.0	0.975	1.0
12	6	right	0.0	0.875	0.9

D STATISTICS AND MEASUREMENTS OF SIMILARITY

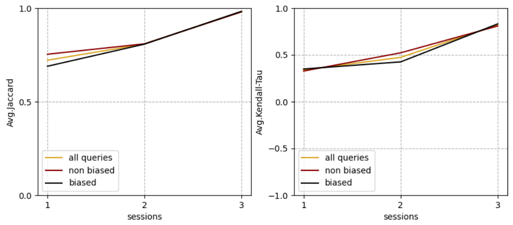

Figure 2: Internal similarities across environmental runs for left wing users

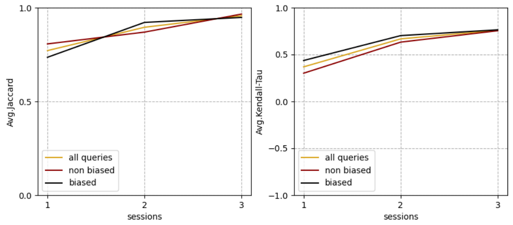

Figure 2: Internal similarities across environmental runs for right wing users