Personalization Finder: A Search Interface for Identifying and Self-controlling Web Search Personalization

Yusuke Yamamoto
yusuke_yamamoto@acm.org
Shizuoka University
Hamamatsu, Japan

Takehiro Yamamoto
t.yamamoto@sis.u-hyogo.ac.jp
University of Hyogo
Kobe, Japan

ABSTRACT

We propose the Personalization Finder, a search interface that enables users to be aware of and control personalization of a web search. The proposed interface is intended to improve behavioral data privacy and promote critical information seeking. A preliminary user survey indicates that many web users worry that search engines personalize their search results for political topics, which can raise concerns about opinion polarization. However, the survey also indicates that few users believe that search engines personalize search results for such topics. Based on the results of an online user study, we confirmed the following: (1) Our prototype interface resulted in users spending more time looking at the web search results at deeper ranking positions than conventional web search interfaces when querying political topics, and (2) on average, users thought our prototype was useful for objective collection of information.

CCS CONCEPTS

• **Information systems** → *Search interfaces*; *Personalization*; • **Human-centered computing** → *User studies.*

KEYWORDS

Personalization; filter bubble; privacy; search user interaction

ACM Reference Format:
Yusuke Yamamoto and Takehiro Yamamoto. 2020. Personalization Finder: A Search Interface for Identifying and Self-controlling Web Search Personalization. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3383583.3398519

1 INTRODUCTION

Advances in personalization technology have enabled preference-based web searches that support efficient information seeking. However, concerns about data privacy and filter bubbles are increasing.

The European Union's recent General Data Protection Regulation[1] indicates increasing societal concern about data privacy. An Accenture Consumer Pulse Research survey[2] reported that 66% of respondents wanted web service vendors to reveal how they exploit behavioral data, although they generally agreed that personalization technologies are useful. The survey also revealed that 40% of respondents feel weird when web services accurately predict their preference and recommend favorable items. These findings suggest that web search engines should be transparent about how their search results are personalized and that the techniques employed should be controllable by users.

Personalization technology also leads to filter bubbles, i.e., a state of information isolation that personalization algorithms can cause by only providing information that reinforces existing opinions or beliefs. Web search engines often modify search results to reflect user preferences and interests, which can cause filter bubbles. Hannak et al. found that, on average, 11.7% of Google Search results varied between users because of a personalization algorithm [8], and Le et al. observed significant personalization of Google News search results for political topics [13]. Personalized web searches often provide users with information that reinforces existing beliefs [21].

Search result diversification [1], which attempts to provide broader-based and serendipitous information, can be useful to mitigate filter bubbles [15]. However, diversification approaches cannot fully solve the filter bubble problem because of selective exposure, individual's tendency to favor information which reinforces their prior belief.As White [21] reported, even if a range of opinions are provided, people often look for information that supports their beliefs in web searches. Therefore, to mitigate filter bubbles effectively, information access systems that consider selective exposure responses are required.

Here, we investigate a web search interface that enables identification and control of personalization. We consider ways to make users more aware of personalization and to promote critical information seeking. Here, we define *critical information seeking* as activities to examine information critically and proactively collect information useful for decision-making using information access systems, by reference to [24]. In addition, we recognize that personalization has advantages and disadvantages for the user. From the perspective of intellectual isolation and data privacy, personalization is negative. However, for some topics where individual preferences differ significantly, e.g., entertainment, personalization can contribute to an effective search.

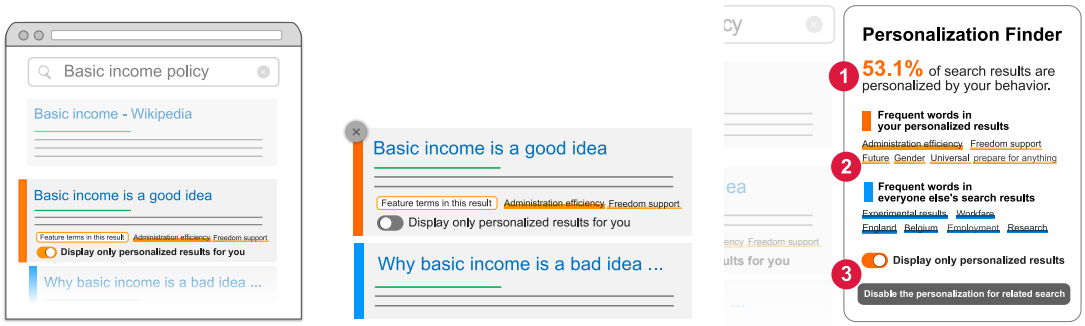
We propose the Personalization Finder, a web search interface that deals with the abovementioned issues. We conducted a user survey about web search personalization and then designed functions to identify and control the personalization of web search. As illustrated in Fig. 1, our prototype Personalization Finder provides the following functions for conventional web search engines:

- highlighting of personalized search results (Fig. 1-(a)),
- visualizing the extent of personalization (Fig. 1-(c1)),

[1]GDPR: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679
[2]https://www.accenture.com/us-en/insight-hyper-relevance-gcpr

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06…$15.00
https://doi.org/10.1145/3383583.3398519



(a) Highlighting of personalized search results on SERPs.

(b) Disclosure of Web search results hidden by the personalization.

(c) Providing various functions to grasp the personalization effect on a side bar.

Figure 1: Personalization finder features.

- exposure of search results removed from the search results because of personalization (Fig. 1-(b)), and
- the ability to disable web search personalization depending on the search results and queries (Figs. 1-(b) and (c3)).

The primary contributions of this study are as follows.

- We administered a user survey and found that people often believe that personalization is not extensively applied to web search results for topics related to politics and economics. In addition, the survey results indicate that people do not want personalization to be applied to such topics.
- Based on the survey results, we propose a web search interface that enables identification and user-control of web search personalization.
- Compared to conventional web search interfaces, our prototype interface results in users spending more time looking at the listed search results and clicking on results at deeper positions in the list.

2 RELATED WORK

Several studies have reported that people often accept web information without consideration of credibility. For example, Morris et al. claimed that many people trust information from social network services more than search engine results [16]. Even if people pay attention to suspicious information, they often misjudge its credibility because of cognitive biases. Ieong et al. [9] revealed the existence of domain biases whereby users believe that specific domains provide more relevant web pages. White et al. [21] suggested that if a user has a strong opinion about a specific search topic, he/she is unlikely to change his/her opinion based on search results. Y. Yamamoto et al. have proposed the *web access literacy scale* to assess user competency for critical information seeking on the web, including tolerance for cognitive bias [24]. T. Yamamoto et al. have revealed that people with strong attitude toward critical information seeking are more likely to use search terms such as "truth" or "evidence" in their queries than those without such attitude [22].

Some researchers in the fields of human–computer interaction and information retrieval have developed methods of promoting critical information seeking. Munson et al. [17] proposed a web browser extension to indicate whether users' browsing histories are politically balanced. Hamborg et al. [7] developed NewsBird, a news aggregation system that presents international news from various perspectives. Yamamoto et al. [23] proposed Query Priming to activate critical information seeking in web searches. Liao et al. [14] revealed that indication of the opinion stance and expertise of the information sender can mitigate the echo chamber effect.

Many researchers have warned of the risk of filter bubbles on social media. Garimella et al. [5] analyzed large-scale Twitter data and found that Twitter users were exposed to political opinions consistent with their own opinions. Nagulendra et al. [18] designed an interactive system through which social networking users became aware of personalization mechanisms. However, Nguyen et al. [19] insisted that collaborative filtering-based personalization can reduce the risk of filter bubbles by recommending varying content.

3 USER SURVEY

We conducted an online survey to understand what people think about personalization in web searches and to explore design concepts for our prototype system. The survey was conducted in Japanese between August 16 and 18, 2019. We recruited 470 participants using Lancers.jp, a Japanese crowdsourcing service[3] (male = 55.1%, female = 43.8%, NA = 0.1%). Most participants were 30 to 50 years of age (20s = 14.9%, 30s = 31.7%, 40s = 32.3%, 50s = 16.2%, others = 4.9%). Each participant was paid 50 Japanese yen (approximately $0.50). On average, the survey took 6.6 min to complete.

3.1 Procedure

Prior to answering any questions, the participants were informed of the following in writing: (1) that web search engines collect behavioral data during web searches and (2) how web search personalization works and changes search results. The information also explained that web search results returned by conventional search engines could differ among users. After reading the description, each participant answered demographic questions and six questions about personalization in web searches.

3.2 Questionnaire Items

This survey attempted to understand the following issues: (1) perceptions of web search personalization, (2) willingness to use a

[3]Lancers.jp: https://www.lancers.jp/

user-controlled personalization function, (3) advantages and disadvantages of personalization, and (4) how to improve personalization.

To assess participants' perception of web search personalization, we posed the following question (Q1): "*What percentage of web search results do you think conventional web search engines personalize for a specific topic?*" For this question, the participants selected an answer from six categories (e.g., *less than 10%*). In addition, participants answered this question for the following topics: politics, economics, entertainment, sports, science/technology, shopping, health, nature, and culture[4].

To assess participants' willingness to use web search personalization, we asked Q2: "*How willing are you to use the personalization function when you use search engines such as Google?*" For this question, the participants used a six-point Likert scale. The participants also answered this question for the same topics as those given in question Q1.

To assess the advantages and disadvantages of personalization, we asked Q3(4): "*What do you think is the major advantage (disadvantage) of the personalization of web searches?*" In addition, we asked the following question (Q5) to investigate the trade-off between personalization and behavioral data provision: "*To use the personalization function, you need to provide your search/browsing behavior logs to web search engine vendors. What do you think about providing your behavioral data for personalization?*"

To investigate potential improvements in personalization, we asked Q6, "*If you could improve web search personalization, what functions would you hope for?*" Note that for questions Q3–Q6, each participant selected an item from a provided list (Table 1).

3.3 Results

3.3.1 Perception of web search personalization. As illustrated in Fig. 2, the participants thought that the extent of web search personalization varied according to the search topic. Furthermore, participants thought that search results for *political* and *economic* topics were personalized much less intensively than those for *shopping* and *entertainment*. According to Hannak et al. [8], politics is one of the most personalized topics. However, 34.9% and 29.1% of the participants thought that conventional web search engines personalize less than 10% of the items in a search result list for *political* and *economic* topics, respectively. Conversely, only 3.8% and 4.3% of the participants thought that less than 10% of search results for *shopping* and *entertainment* topics, respectively, were personalized.

3.3.2 Willingness to use the personalization. As illustrated in Fig. 3, 62.3% and 50.4% of the participants were unwilling or very unwilling to use the personalization function when searching for *political* and *economic* topics, respectively. Conversely, 56.4% and 55.7% of the participants were willing or very willing to use the personalization function for *shopping* and *entertainment* topics, respectively.

3.3.3 Advantages and disadvantages of the personalization. As presented in Table 1, most participants thought that the best attribute of the personalization was to promote efficient information seeking. Among the participants, 39.8% thought that personalization was

[4]These topics were selected from the top categories of Yahoo Japan News.

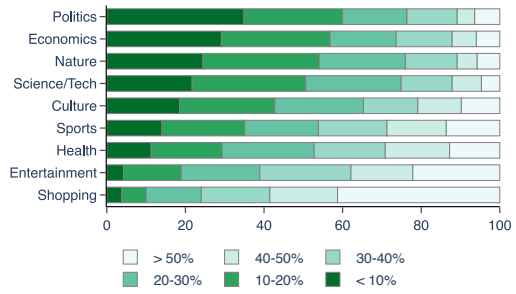


Figure 2: Perception of web search personalization.

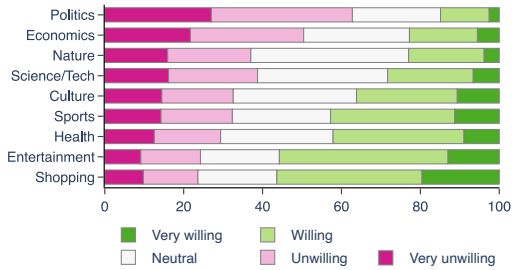


Figure 3: Willingness to use web search personalization.

useful for reducing search costs. In addition, 35.5% thought that personalization enabled them to find information that was favorable but difficult to search for.

Conversely, the survey revealed that many participants disliked the following three aspects of personalization. First, 43.2% worried that, if they used the personalization, they would lose opportunities to encounter diverse content (obtaining a narrower view, 22.8%, and losing opportunities to see unexpected information, 20.4%). Second, 29.8% did not want to provide their search/browsing behavior logs to search engine vendors. Third, 18.5% disliked personalization because they felt that the search engines were controlling their access to information and their subsequent behavior.

3.3.4 Trade-off between the personalization and data provision. 61.1% of participants were willing to provide a portion of their search/browsing behavior logs to search engine vendors to support personalization. However, 31.1% did not want to provide any behavioral data.

3.3.5 Improvements in the personalization. 22.3% of participants wanted a function to disable personalization, whereas 17.2% wanted to control the personalization metrics. 16.6% wanted search engines to display various content to broaden their viewpoint, and 13.6% wanted a function to check non-personalized web search results as well as personalized results, and 21.3% wanted a function to delete their behavior logs.

4 SYSTEM DESIGN

Our goal was to design a system that enables web users to be aware of the personalization of web searches, to reflect on their preferences and biases, and to manipulate personalized web search results to enable critical information seeking. In this section, we present

Table 1: Opinions concerning web search personalization.

Question	Percentage
Q3. Advantage of web search personalization	
I can decrease the cost to search for information.	39.8%
I can find information that is favorable but difficult to search for.	35.5%
I can find only the information that I want to read.	10.0%
There are no advantages.	8.1%
I can obtain insights into my preferences.	6.2%
Others.	0.4%
Q4. Disadvantage of web search personalization	
I need to provide behavioral data to search engine vendors.	29.8%
I obtain a narrower view of things.	22.8%
I lose opportunities to see unexpected information.	20.4%
I feel as if search engines control my access to information and my behavior.	18.5%
My ability to search for information gets weaker.	5.3%
There are no disadvantages.	2.8%
Others.	0.4%
Q5. Collection of behavioral data for personalization	
I don't mind providing a portion of my search/browsing history.	61.1%
I don't want to provide all of my search/browsing history.	31.1%
I don't mind providing all of my search/browsing history.	7.9%
Q6. Desired function improvement to personalization	
Disabling web search personalization.	22.3%
Deletion of behavioral data used for personalization.	21.3%
Self-control of personalization metrics.	17.2%
Presenting diverse information so as not to have a narrow point of view.	16.6%
Presenting non-personalized search results normally hidden to users.	13.6%
Presenting reasons for the personalization.	8.7%
Others.	0.2%

system requirements based on the preliminary survey results. Then, we describe our approach to system implementation.

4.1 Requirements

The first requirement is to expose the personalization effect on a list of web search results. The survey results revealed that quite a few people do not know that personalization algorithms are frequently applied to search results concerning politics and economics. The results also indicated that most participants do not want personalization applied to those topics. Conventional web search results do not indicate whether a personalization algorithm has been applied. Therefore, functions to inform users that results have been personalized and the degree to which they are personalized are required.

The second requirement is to expose web search results that are hidden because of personalization. The survey results suggest that people often worry that personalization algorithms prevent them from encountering diverse content. To enable critical information seeking, it is crucial for users to check both favorable and unfavorable results. Therefore, we need to consider how a system can draw users' attention to hidden search results because some users are not willing to investigate search results inconsistent with their beliefs [21].

The third requirement is a user to be able to disable web search personalization. The survey results also indicated that some people want to completely disable personalization. However, the survey results also suggest that people are not always opposed to personalization and may want to use it for specific topics. Therefore, an appropriate system needs to let users flexibly control web search personalization.

Finally, it is desirable to let users easily remove their behavioral data used to personalize a specific web search. The survey results revealed that many people want to hide or delete a portion of their search history because of privacy concerns.

4.2 Concept

We designed a prototype of the Personalization Finder to meet the above requirements. Our prototype is designed as a browser extension that works with conventional web search engines. Figure 1 illustrates features of our prototype. Note that we define *personalized search results* as those that do not occur in a list of search results when the personalization algorithm is not used for the same query (the opposite is defined as *non-personalized search results*). Furthermore, *hidden search results* are defined as results omitted from the presented result list because of personalization.

To enable users to see that web search personalization is occurring, the proposed prototype highlights personalized search results for a given query. As illustrated in Fig. 1-(a), personalized search results are identified by an orange bar to the left of the entry. Then, some of the terms featured in each personalized result are displayed below its snippet. Inclusion of these terms is intended for helping users to learn which factors are emphasized by the personalization. This highlighting function helps users identify personalized search results.

To summarize the personalization effect, we analyze the list of original search results and note how many of the search results are personalized as *personalization degree* in the sidebar (Fig. 1-(c1)). To make users more aware of the search results removed because of personalization (i.e., hidden search results), the proposed prototype displays one of these results below each personalized result (Fig. 1-(a)). By default, to avoid information overload, only a small portion of the hidden results is displayed with a transparent gradation. However, users may be more inclined to look at hidden search results due to psychological reactance [3]. If users click on the orange bar next to their personalized results (Fig. 1-(a)), the prototype reveals all hidden search results (Fig. 1-(b)). The prototype also provides a function to reveal all hidden results simultaneously (an orange toggle button on the sidebar illustrated in Fig. 1-(c3)). In the sidebar, the prototype displays the term sets featured in both the personalized and the hidden search results; this allows users to understand the difference between the two sets of results.

In addition, the prototype provides two functions to mitigate and disable the personalization of web searches. If users click the "delete" button on the upper-left corner of each personalized result, the result is removed from the search result list (Fig. 1-(b)). At the same time, if we can access user search histories, the prototype can delete the search histories used to personalize the removed results. This function is intended to prevent such deleted histories from affecting the personalization in subsequent web searches. The proposed prototype provides a function to perform the same operation for all personalized results for a query (gray button in Fig. 1-(c3)).

4.3 Hypotheses

In a user study, we investigated the effect of the Personalization Finder on search behavior and decision making on the web. We expect Personalization Finder to provide opportunities for users to be aware of web search personalization and its influences; thus,

users will search the web more critically. We hypothesized the following:

H1: *With Personalization Finder, users will spend more time searching the web more carefully compared to with conventional web search interfaces.*

H2: *Personalization Finder encourages users to visit more webpages to obtain more information.*

H3: *Personalization Finder will change prior beliefs more significantly through web searches than a default web search.*

We expect that Personalization Finder can mitigate personalization of web searches; however, personalization is not always a problem. The user survey revealed that people should be aware that personalization of web searches for politics can cause harmful information isolation or social polarization. Exposure to diverse political opinions is considered quite important [6]. For entertainment topics, e.g., music and movies, individual preferences differ; thus, the effectiveness of personalization is expected to be high. Therefore, we propose another hypothesis:

H4: *The effect of Personalization Finder will be greater for political search topics than entertainment topics.*

5 USER STUDY

This section describes a user study to evaluate our prototype search system for identifying and controlling web search personalization and promoting critical information seeking. The user study was conducted in Japanese (on August 21 and 31, 2019).

5.1 Methodology

5.1.1 Participants. We recruited 300 participants using Lancers.jp, a Japanese crowdsourcing service. We excluded 80 participants from the analysis because they unintentionally used web search engines that we did not allow or because they completed the tasks without a web search. Therefore, we analyzed 220 participant responses (113: male, 100: female, and 7: NA). Most participants were in their 30s and 40s (10s = 1.8%; 20s = 22.3%; 30s = 29.5%; 40s = 31.4%; 50s = 11.8%; others = 3.2%). All participants who completed the tasks received 350 Japanese yen (approximately $3.50). On average, the participants finished all tasks within 36.6 min.

5.1.2 Search topics. We prepared four topics for the search tasks. Table 2 lists the four topics. Two of them involved political topics. The selected topics are often controversial but are essential for the future in Japan. Therefore, they require careful and objective consideration. The other two topics involved entertainment. For any question, the answer often reflects a person's preferences and sense of values; therefore, if one performs web search on each topic using conventional web search engines, the search results would be personalized.

Here, we define *web search personalization degree (P-degree)* as the ratio of personalized search results in the 100 web search results of a searcher per query. We computed P-degree by comparing each participant's search results with 100 non-personalized search results against the input query. Table 2 presents the mean P-degree for each topic. It indicates that web search result lists obtained from Google Search were intensely personalized.

5.1.3 Design and procedure. In the user study, we adopted a 2 × 2 mixed factorial design to examine the effects of two factors, i.e., search topic type and search UI condition. The search topic type factor had two levels: *politics* and *entertainment*. The search UI condition also had two levels: *control* UI, which provides a list of search results for a given query in the same manner as a conventional web search engine, and *experimental* UI, which is an extension of the control UI that provides functions to expose and control web search personalization.

After the participants agreed to a consent form on the crowdsourcing service, they were transferred to our website for the user study. Then, we randomly allocated each participant to a UI condition. Here, 111 participants used the control UI, and 109 participants used the experimental UI.

First, we asked the participants to download and install a Google Chrome extension developed for this study. The participants were also asked to use the Google Chrome browser with the extension during the study. Then, the participants read a description of the task flow and search system. We explained that our extension worked on Google Search and provided the same search results for a query as those from Google Search. Furthermore, we indicated that Google generally personalizes web search results considering user interest and search/browsing behavior. For participants in the experimental group, we described each function of the prototype using visual materials.

Next, we asked participants to perform a practice task to become familiar with the tasks and the search system. The participants were told to assess the advantages and disadvantages of the basic income policy using the allocated search system.

Then, the participants performed four search tasks for the topics in Table 2. The task order was randomized for each participant. Each search task comprised three steps. First, the participants were told to report their interest, knowledge, and prior belief for each search task topic. The participants ranked their interest and knowledge using a five-point Likert scale (1: Not at all; 5: Extremely). For their prior belief on each *political* topic, we asked if they agreed or disagreed with the topic[5] on a seven-point Likert scale (-3: Extremely disagree; 3: Extremely agree). For prior belief on each *entertainment* topic, we asked participants to imagine their opinion about the topic[6] and report their confidence in the opinion on a five-point Likert scale (1: Not at all; 5: Extremely). Table 2 indicates that the participants were not very familiar with the search topics and did not have strong prior beliefs on average.

Second, the following scenario was presented:

> *"Nuclear power phase-out policy" is one of the most controversial topics in Japan. Please imagine that you have been asked to answer if you agree or disagree with the nuclear power phrase-out policy[7]. Assume that you are now about to collect information about the nuclear power phase-out policy via a Web search to formulate your*

[5] For political topics, we asked the questions like "Do you agree or disagree with nuclear power phase-out policy?"
[6] For entertainment topics, we asked the questions like "What music uplifts you?"
[7] For entertainment topics, the early part of a scenario was like "Please image that you want to listen to uplifting music."

Table 2: Search topic (query), participant's impression of a topic, and web search personalization degree (P-degree). Interest, Knowledge, and Prior belief use a five-point scale (1: Not at all; 5: Extremely). Prior agreement uses a seven-point scale (-3: Extremely disagree; 3: Extremely agree). Numbers in the table indicate the mean and standard deviation (in parentheses).

Topic type	Topic (query)	Interest	Knowledge	Prior agreement	Prior belief	P-degree (%)
Politics	Nuclear power phase-out	2.89 (0.81)	2.09 (0.72)	0.19 (1.40)	—	49.9 (7.1)
	Constitutional Revision	2.86 (0.82)	2.09 (0.70)	0.06 (1.29)	—	51.3 (6.7)
Entertainment	Uplifting music	2.89 (1.06)	2.08 (0.82)	—	2.63 (0.95)	65.2 (8.2)
	Movies that you watch at home on a holiday	2.75 (1.11)	2.21 (0.83)	—	2.43 (0.98)	62.5 (4.5)

> *opinion. Seek a displayed list of Web search results and visit several web pages on the list. When you come to a satisfactory conclusion, stop the Web search and report your final opinion with reasons in the answer form below.*

Then, each participant clicked a link to start a web search. The browser opened a search engine results page (SERP) on the Google Search website for a fixed query (e.g., nuclear power phase-out). One hundred search results per page were displayed to each participant. Each participant browsed the list of search results and attempted to determine their answer to the task questions. For political search tasks, the participants indicated whether they agreed or disagreed with the topic question using a seven-point Likert scale (-3: Extremely disagree; 3: Extremely agree). For entertainment tasks, the participants provided free-form answers.

Third, participants were asked to report the extent to which their prior belief changed because of the search task using a five-point Likert scale (1: Not at all; 5: Extremely).

We then administered an exit questionnaire (Table 4) to obtain feedback about the search systems. Participants in the experimental group answered additional questions to examine the functions of the prototype system. All participants also reported opinions about the search system in free form. Finally, we asked demographic questions related to sex, age, and education.

5.1.4 The search system. We developed Personalization Finder as a Google Chrome extension that worked on the Google Search website. This extension manipulated Google Search results and the SERP HTML. The extension also monitored participant behavior during the user study.

For the control and experimental groups, when the participants started Google Search from the link on our experiment website, the link fetched 100 organic search results for a specific query (*original Google results*). The original Google results were those that Google personalized and provided to participants, which varied for each participant. Then, the extension removed vertical search results (e.g., news and images) from the original results to simplify the user study analysis. The extension did not permit participants to modify their initial query because we wanted to ignore the *carry-over effect*, which the last querying affect immediate search results [8].

For the experimental group, the extension provided functions (Fig. 1) to identify and control web search personalization. To highlight personalized search results, the extension needed to obtain a list of non-personalized search results for queries not personalized by Google's personalization algorithms. Here, we used the Google Custom Search JSON API[8] to obtain approximately a list of non-personalized search results. Once each participant started a web search for a query, the extension fetched 100 non-personalized search results using the Google API behind the UI. Then, the extension highlighted each personalized search results comparing the original Google search results with the API results. Furthermore, the extension indicated the percentage of personalized search results in a list of the 100 original search results as *personalization degree* in the sidebar.

To present terms featured in each personalized search result (Fig. 1(a)), the prototype extracted three terms with the greatest term frequency-inverse document frequency (TF-IDF) weight. To present terms featured in a set of personalized/hidden search results (Fig. 1(c)), 20 terms with high TF-IDF weights that did not appear in personalized/hidden search results were selected. To allocate a hidden search result for each personalized search result (Fig. 1b)), the prototype selected a result from the Google API search results in the ranking order.

In the user study, we disabled the function to delete search histories for personalization because we did not have permission to access participant behavior logs on Google.

5.2 Statistical Analysis

We analyzed the collected search behavior logs using generalized linear mixed models (GLMM) [2], which distinguish *fixed effects* caused by experimental conditions from the *random effects* caused by variations between the random samples, e.g., participants and tasks. GLMMs are popular tools for modeling various user behaviors in information retrieval and human–computer interactions, where GLMMs replace traditional ANOVA analyses [11]. We constructed Bayesian models for the behavioral data using a GLMM. Although this approach is less familiar than the frequentist approach, we used it because it handles data uncertainty better, yielding probability distributions of the target parameters.

Here, we assumed that the session and dwell times on the SERPs followed Weibull distributions.For countable data, such as the number of click-throughs, we adopted a GLMM with a Poisson distribution. For the probability of clicking Web search results, we adopted a GLMM with a beta distribution, and we used a linear mixed model with Gaussian distribution for belief change.

We trained the GLMMs with random effects for each response variable. The fixed effects were *UI condition* and *topic type*, and the random effects were *participant* and *task*. Following [2], we modeled the following maximal mixed model for each response

[8]Google Custom Search JSON API: https://developers.google.com/custom-search/v1/overview

variable using the R package `brms`:

$$\begin{aligned} Y &\sim \text{UI} + \text{Topic} + \text{UI} : \text{Topic} + (1|\text{Task}) \\ &+ (1 + \text{UI} + \text{Topic} + \text{UI} : \text{Topic}|\text{Participant}), \end{aligned}$$

where `Y` is the observed response, `UI` is a binary indicator of whether a participant used our prototype for the tasks, and `Topic` is a binary indicator of whether a task was a political topic. Here, `(x|y)` means that `y` is a random effect of `x`.

We adopted two approaches to examine the effects of UI condition and topic type. First, we checked the *high density interval (HDI)*, which summarizes the posterior distribution of a parameter such that each point inside the interval has a probability density greater than that of any point outside the interval [12]. When zero lies outside the 95% HDI of the coefficients of the target variable, Bayesian statistics interpret this as the variable with a significant effect on the outcome. Conversely, frequentists conduct a significance test of the null hypothesis at a specified significance level of $\alpha = 0.05$. Second, we calculated the Bayes factors (ratio of the likelihood of two competing hypotheses) to compute the result of a Bayesian hypothesis test [10]. The Bayes factor of hypothesis H_a versus H_b (BF_{H_a/H_b}) represents how strongly the data support H_a over H_b. According to [10], if $BF_{H_a/H_b} > 10$, we can interpret H1 as being more strongly supported by the data than H2.

To analyze the exit questionnaire, we used the Wilcoxon signed-rank test. Significant effects are reported at a significance level of $\alpha = 0.05$.

5.3 Results

5.3.1 Search duration. To test **H1**, we first analyzed how long participants spent on search tasks (i.e., session time). As indicated in Table 3, we found that the 95% HDI of the interaction coefficient between UI condition and task topic excluded zero (equivalent to $p < 0.05$ in the frequentist approach). Figure 4(a) illustrates the mean and 95% credible interval of the session time on the posterior distribution. As can be seen, the GLMM model predicted that there was a very little difference between the simple effect of the control group session time and that of the experimental group for the entertainment topics (mean delta = 3.0s). For political topics, the mean of the simple effect of the control group was 52.0s longer than that of the experimental group. In the analysis of Bayes factors for the fixed-effect parameters, the Bayes factor (BF) for the interaction coefficient $\beta_{u\times t} < 0$ versus $\beta_{u\times t} > 0$ was BF = 42.76.

To investigate how much time participants spent examining web search result lists, we analyzed dwell time on SERPs. Table 3 indicates that the 95% HDI of all fixed-effect coefficients contained zero; however, in the analysis of the Bayes factors for fixed-effect parameters, we confirmed that the Bayes factor for the UI condition coefficient $\beta_u > 0$ versus $\beta_u < 0$ was BF = 36.56. On the posterior distribution, the mean dwell times on SERPs of the experimental group were longer than those of the control group for both entertainment and political topics (entertainment: 40.1s versus 32.2s; politics: 47.7s versus 37.0s).

These results suggest that even though the prototype did not increase session time (including web page browsing), it caused users to spend more time browsing search results or reading information on the SERPs for both political and entertainment topics.

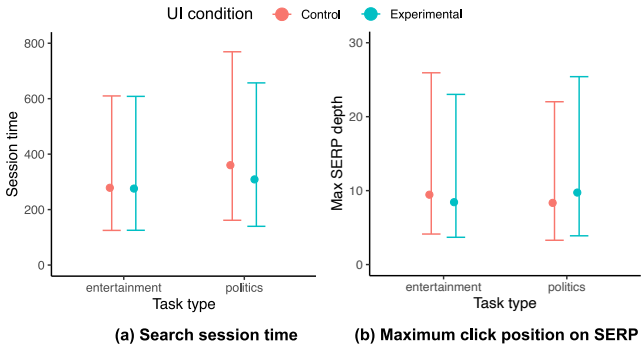


Figure 4: Mean and 95% credible intervals of (a) the session time and (b) the max SERP depth on the posterior distribution.

5.3.2 Click-through. To test hypothesis **H2**, we analyzed how many links participants clicked in their SERP list (i.e., click-through count). Table 3 indicates that the 95% HDIs of *UI condition*, *topic type*, and interaction contained zero. The results do not indicate that *UI condition* or *topic type* had significant effect on the click-through count. Furthermore, the Bayes factor analysis indicated that neither fixed effect was significant.

To investigate what the participants clicked on, in detail, we analyzed what percentage of participants' clicked search results were personalized or hidden results[9]. Here, we refer to such search results as *manipulated search results (MSRs)*. Even though the control group did not know which search results were personalized, we could check whether the clicked web search results were personalized. We then investigated how disclosure of personalized/hidden search results attracted participant attention and affected clicking behavior by analyzing the click-through ratio of MSRs.

The GLMM analysis revealed that 95% HDI of the UI condition coefficient exceeded zero ($\beta_u = 0.33$) (Table 3), and the estimated mean of MSR click probability of the experimental group was greater than that of the control group for the entertainment and political topics (entertainment: 18.5% versus 16.2%; politics: 16.7% versus 14.6%). In the analysis of Bayes factors for fixed-effect parameters, we confirmed that the Bayes factor for the UI condition coefficient $\beta_u > 0$ versus $\beta_u < 0$ was 261.3. These results indicate that the participants were more likely to click on MSRs than control participants when the prototype highlighted and revealed MSRs.

According to [22], users with a stronger attitude about critical information seeking are more likely to click lower-ranked search results. Therefore, we analyzed *max SERP depth*, the maximum rank position of clicked URLs in a search result list per query. We found that 95% HDI of the interaction coefficient between UI condition and task topic excluded zero. We confirmed that the Bayes factor for interaction coefficient $\beta_{u\times t} > 0$ versus $\beta_{u\times t} < 0$ was over BF = 40.9. From the analysis of the two-way interaction effect, we found that the UI condition effect varied according to the topic type (Fig. 4(b)). For entertainment topics, the expected mean max. SERP depth of the experimental group was 1.01 less than that of the control

[9]Note that in this study, participants could click on hidden search results only when they used our prototype functions to disclose them in Fig.1-(b) and (c).

Table 3: Statistics on search behavior and belief change for each UI condition and task type. The entertainment and politics present show the mean (with SD) of each response. The estimated coefficient column presents the mean and 95% HDI of each coefficient via the Bayesian GLMMs. Bold numbers indicate that the 95% HDI of a coefficient of a parameter did not contain zero. Numbers with an asterisk indicate that the Bayes factor was over 10.

	Entertainment		Politics		Estimated coefficient		
Response	*Control*	*Experimental*	*Control*	*Experimental*	*UI*	*Task type*	*Interaction*
Task completion time (s)	345.1 (254.8)	319.9 (227.1)	477.9 (441.2)	382.6 (289.1)	-0.01 [-0.19, 0.16]	0.26 [-0.75, 1.40]	**-0.14*** [-0.29, -0.01]
Dwell time on the SERP (s)	41.4 (39.8)	54.2 (102.9)	48.5 (53.7)	61.3 (78.9)	**0.22*** [-5.4e^{-5}, 0.45]	0.14 [-1.43, 1.73]	0.03 [-0.19, 0.25]
Click-through count	3.55 (2.69)	3.42 (2.33)	3.19 (1.96)	3.53 (2.58)	-0.02 [-0.20, 0.14]	-0.11 [-1.58, 1.38]	0.11 [-0.05, 0.27]
Click probability for MSRs	0.16 (0.25)	0.26 (0.32)	0.07 (0.20)	0.17 (0.29)	**0.33*** [0.09, 0.56]	-0.23 [-1.50, 0.83]	-0.02 [-0.44, 0.38]
Max SERP depth	15.0 (18.7)	13.0 (16.3)	12.7 (14.5)	15.2 (16.7)	-0.11 [-0.38, 0.14]	-0.12 [-1.50, 1.17]	**0.27*** [1.2e^{-3}, 0.53]
Belief change	2.67 (1.19)	2.92 (1.15)	2.24 (0.90)	2.36 (0.92)	**0.25*** [0.01, 0.48]	-0.45 [-2.85, 1.52]	-0.13 [-0.39, 0.15]

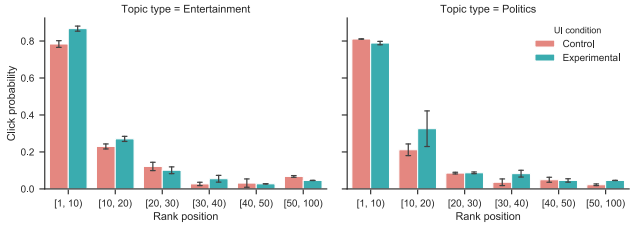


Figure 5: Click probability of search results by rank position. [x, y) indicates the rank of a search result whose rank ranges from x to y.

group (8.53 versus 9.54). This indicates that the prototype caused participants to look at search results with higher rank positions. For political topics, the expected mean of the experimental group was 1.37 greater than that of the control group (9.83 versus 8.44).

We observed a similar tendency for click probability by rank position. According to Fig. 5, when searching for political topics, more participants in the experimental group clicked on search results with lower positions, especially in the 10–19th positions, than those in the control group. Conversely, for entertainment topics, participants in the experimental group were more likely to click on the top 10 results than participants in the control group. These results indicate that if participants searched for political topics with the prototype, they were likely to seek SERPs at lower positions.

5.3.3 Belief change. To test **H3**, we analyzed the *subjective belief change scores* reported by participants after each search task. Using GLMM analysis, we found that the 95% HDI of the UI condition coefficient exceeded zero ($\beta_u = 0.25$ in Table 3). We confirmed that the Bayes factor for the UI condition coefficient $\beta_u > 0$ versus $\beta_u < 0$ was BF = 52.0. In a simple effect analysis, the estimated mean of the belief change score of the experimental group was greater than that of the control group for both topic types (entertainment: 2.91 versus 2.67; politics: 2.34 versus 2.22). These results indicate that the prototype is more likely to induce greater belief changes through web searches than the baseline interface. However, the effect was less than expected.

5.3.4 Exit questionnaire. Table 4 summarizes the exit questionnaire results. The results for questions Q1, Q2, and Q3 indicate that most participants had positive impressions of the prototype relative to identifying personalization (a score of 3 indicates "moderately"). The following free participant comments reflect this result: P1: *"I sometimes feel that the personalization limits my choices and controls my behavior. Your prototype is useful to understand the dark side of search engines."* P2: *"Highlighting the personalized search results was quite useful. The color highlighting enabled me to easily judge which Web search results were only (not) presented to me."* P3: *"I appreciated the function to expose personalized search results and to reveal hidden results. This function was more useful to check other people's opinions and biases in my own opinion rather than to identify my favored results."*

As described in the previous section, few participants used the function to remove PSR from the SERP (i.e., de-personalization function). However, most participants reported that if the function could delete related search behavioral data from web search engines, then they would be eager to use such a function (Q4's score: 3.52).

Compared to the baseline UI, the participants thought that the prototype was more useful for objective collection of information (Q5's score: 3.28 versus 2.95; $p < 0.001$). The following comment supports this conclusion: P4: *"I knew that Google personalized search results. However, I was surprised that my results were personalized more intensely than I expected. ... I was happy to check both the personalized results and the hidden ones using your prototype."* However, the participants thought there was no difference between the prototype and baseline interface for objective decision making (Q6: $p = 0.098$). For usability, we found no difference between the prototype and baseline interface (Q7: $p = 0.236$). Q8 revealed that the participants were more eager to use the prototype system than the baseline method (Q7: $p < 0.05$); this indicates that the participants preferred the prototype.

Several participants indicated complaints or suggestions for system improvements, e.g., *"I couldn't understand what the "related words" displayed on the sidebar were and why they appeared there."* (P2)

6 DISCUSSION

Based on the survey results, we designed a search interface to identify and mitigate web search personalization. The prototype attempted to make users aware that web search results and their preferences are often biased. Furthermore, the prototype allowed users to control web search results for critical information seeking. The user study results demonstrated that the prototype affected participants as follows: (1) Users spent more time examining SERPs during search tasks; (2) the prototype attracted users to PSR and promoted clicking these results; and (3) participants using the prototype examined the search result list to higher (lower) rankings when researching political topics (lower (higher) rankings for entertainment topics). These findings indicate that **H1** was supported by the data.

The GLMM analysis revealed that the prototype did not affect the number of visited web pages. Prior to the user study, we expected that individuals using the prototype would visit more web pages to check information revealed to everyone else; however, the analysis indicated that the prototype did not encourage users to click on more links on SERPs. Therefore, **H2** was not supported.

The analysis revealed that the prototype induced belief change relative to search topics; however, the effect was smaller than expected. In particular, for political search tasks, even when using the prototype, the users did not change their prior stance to an opposite stance. According to the exit questionnaire, the participants felt that the prototype was more useful for collecting information more objectively than the conventional search interface (Table 4). However, they also reported that there was a little difference between the prototype and the baseline for making objective decisions concerning the tasks (Table 4). Thus, we conclude that **H3** was not strongly supported. We conclude that **H4** was partially supported because the prototype had a completely different influence on only the maximum depth of clicked search results.

We make the following interpretations via the user study: (1) Participants using the prototype viewed information on SERPs for a long time and with care. Especially for political topics, they widely viewed the list of web search results on SERPs. (2) Even though participants using the prototype attempted to objectively seek a list of search results, they did not visit many web pages and were willing to visit search results highlighted as personalized by the prototype. (3) Consequently, their prior beliefs did not change significantly.

As White [21] reported, if users have strong belief about a search topic, they are unlikely to revise their belief because of web search results. Liao et al. [14] revealed that if users found that belief-inconsistent information was created by people with high expertise, then they could willingly overcome their tendency to click belief-consistent information. Our user study revealed that the prototype made users aware of web search personalization and to examine search result lists longer. Following Liao's study, one possible improvement to the prototype would be to provide supplementary information on non-personalized and hidden search results such that users would be willing to objectively check unfavorable information.

Usability improvements are another remaining issue. The prototype provided an interactive function to disclose hidden search results; however, only 13% of participants used this function.Other interactive functions were used less frequently than the function that discloses hidden results. We suspect that some participants ignored or forgot these functions. As Spink [20] described, few users use interactive feedback functions in information retrieval (only 4% of users used the feedback function of information retrieval systems in Spink's study). However, as participant P3's comment suggested, some participants believed that it was useful to disclose search results hidden by the personalization algorithm. Furthermore, the exit questionnaire suggests that people would be more willing to use the prototype than the conventional web search interface, although the prototype involves some overhead to manage personalized results. Therefore, we conclude that the functions in the prototype were not useless, but we must improve their usability.

One limitation of our study concerns personalization detection accuracy. To determine which search results were personalized by Google, the prototype analyzed the difference between the user search result list and Google Search API result list. However, Google Search API results are slightly older than original Google Search results; therefore, the prototype may have incorrectly judged non-PSR as personalized results because non-personalized results were new and did not exist in the API result list. The objective of this study was to investigate how the prototype affected search behaviors under the assumption that the prototype could detect PSR. Therefore, we consider that personalization detection accuracy was not a significant factor. However, accuracy of personalization detection should be considered for the practical application of the prototype in future.

Another limitation involves explaining the reason for web search personalization. Some users may want to know why and how search results are personalized. The prototype analyzed the difference between personalized and hidden results and displayed terms featured in personalized results as the personalization factor. We admit that this approach is quite straightforward. In fact, as reflected by P2's comment, some participants wondered what the terms the prototype displayed as personalization factors meant. As Dodge et al. [4] reported, to judge the quality of black-box machine learning, including web search personalization, provision and explanations of overviews of the trained model and scrutiny of individual cases is important. It is difficult to explain the computed personalization model to end-users in a natural language; therefore, if we can access user search histories, then we can demonstrate how the search result list changes if searchers remove search histories associated with PSR.

Another limitation is related to user study conditions. We recruited Japanese participants using a crowdsourcing service. Cultural differences may exist in how our prototype works. In addition, laboratory and online studies should be conducted for more effective analysis.

Finally, our prototype has a limitation relative to handling search behavioral data in search engines. We proposed a function that removes behavioral data and PSR. Even though this function was not

Table 4: Mean rating scores (with SDs) of the exit questionnaire. A five-point Likert scale was used for each question (1: Not at all; 5: Extremely). Significant differences from the baseline are shown in bold.

Question	Control	Experimental
Q1: Was the provided information useful to determine how many search results were personalized?	—	3.05 (0.79)
Q2: Was the provided information useful to determine which search results were personalized?	—	3.10 (0.77)
Q3: Was the provided information useful to determine what search results were hidden from you?	—	3.11 (0.75)
Q4: If the de-personalization function could delete related search behavioral data, would it be useful?	—	3.52 (0.89)
Q5: Was the search system useful for the objective collection of Web information?	2.95 (0.70)	**3.28*** (0.88)
Q6: Was the search system useful to make an objective decision with respect to the tasks?	2.82 (0.75)	2.98 (0.79)
Q7: Was the search system easy to use?	3.15 (0.89)	3.26 (0.96)
Q8: Are you eager to use the search system in the future?	2.81 (0.90)	**3.09*** (0.90)

actually implemented, the exit questionnaire revealed that several participants desired such functionality (score = 3.52 in Table 4). As participant P1 noted, some people worry about search engine vendors collecting and using behavioral data. Thus, if possible, it is desirable for users to be able to easily remove such data.

7 CONCLUSIONS

This study proposes Personalization Finder, which enables user awareness of web search personalization and management of personalized/hidden search results to allow critical information seeking. Analysis of our user study confirmed that our prototype made users to spend more time examining search result lists to a deeper ranking than conventional web search interfaces when searching for political topics. In addition, we found that on average, users thought the prototype was useful for collecting web information.

The prototype promoted critical information seeking at the behavioral level. However, the prototype did not affect user belief change. Therefore, further study is required to determine how we can support users to mitigate biases even if prior beliefs are strong. Even though some issues remain to be solved, our findings contribute to mitigation of filter bubbles and concerns about data privacy in web searches.

ACKNOWLEDGMENTS

The work was supported in part by the Grants-in-Aid for Scientific Research (16H01756, 18KT0097, 18H03244, 18H03243, 18H03494) from the MEXT of Japan.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying Search Results. In *Proc. of the 2nd ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, 5–14.

[2] D. Barr, R. Levy, C. Scheepers, and H. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 3 (2013), 255–278.

[3] S. Brehm and J. Brehm. 1981. *Psychological reactance: A theory of freedom and control*. Academic Press.

[4] J. Dodge, Q. Liao, Y. Zhang, Rachel K. Bellamy, and C. Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proc. of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, 275–285.

[5] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proc. of the 2018 World Wide Web Conference (WWW '18)*. 913–922.

[6] R. Garrett and P. Resnick. 2011. Resisting political fragmentation on the Internet. *Daedalus* 140, 4 (2011), 108–120.

[7] F. Hamborg, N. Meuschke, and B. Gipp. 2017. Matrix-Based News Aggregation: Exploring Different News Perspectives. In *Proc. of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL '17)*. 69–78.

[8] A. Hannak, P. Sapiezynski, A. Molavi, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. 2013. Measuring Personalization of Web Search. In *Proc. of the 22nd International Conference on World Wide Web (WWW '13)*. 527–538.

[9] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. 2012. Domain Bias in Web Search. In *Proc. of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, 413–422.

[10] R. Kass and A. Raftery. 1995. Bayes factors. *J. Amer. Statist. Assoc.* 90, 430 (1995), 773–795.

[11] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H. Yoon. 2017. What Snippet Size is Needed in Mobile Web Search?. In *Proc. of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, 97–106.

[12] J. Kruschke. 2010. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, Inc.

[13] H. Le, R. Maragh, B. Ekdale, A. High, T. Havens, and Z. Shafiq. 2019. Measuring Political Personalization of Google News Search. In *Proc. of the 30th World Wide Web Conference (WWW '19)*. 2957–2963.

[14] Q. Liao and W. Fu. 2014. Expert Voices in Echo Chambers: Effects of Source Expertise Indicators on Exposure to Diverse Opinions. In *Proc. of the 32nd SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 2745–2754.

[15] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. 2019. The impact of result diversification on search behaviour and performance. *Information Retrieval Journal* (2019), 1–25.

[16] M. Morris, J. Teevan, and K. Panovich. 2010. What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior. In *Proc. of the the 28th ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, 1739–1748.

[17] S. Munson, S. Lee, and P. Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proc. of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM '13)*. 419–428.

[18] S. Nagulendra and J. Vassileva. 2014. Understanding and Controlling the Filter Bubble Through Interactive Visualization: A User Study. In *Proc. of the 25th ACM Conference on Hypertext and Social Media (HT '14)*. ACM, 107–115.

[19] T. Nguyen, P. Hui, F. Harper, L. Terveen, and J. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proc. of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, 677–686.

[20] A. Spink, B. Jansen, and Cenk O. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet research* 10, 4 (2000), 317–328.

[21] R. White. 2013. Beliefs and Biases in Web Search. In *Proc. of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, 3–12.

[22] T. Yamamoto, Y. Yamamoto, and S. Fujita. 2018. Exploring People's Attitudes and Behaviors Toward Careful Information Seeking in Web Search. In *Proc. of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, 963–972.

[23] Y. Yamamoto and T. Yamamoto. 2018. Query Priming for Promoting Critical Thinking in Web Search. In *Proc. of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, 12–21.

[24] Yusuke Yamamoto, Takehiro Yamamoto, Hiroaki Ohshima, and Hiroshi Kawakami. 2018. Web Access Literacy Scale to Evaluate How Critically Users Can Browse and Search for Web Information. In *Proc. of the 10th ACM Conference on Web Science* (Amsterdam, Netherlands) *(WebSci '18)*. 97–106.