3bij3 – Developing a framework for researching recommender systems and their effects

Loecherbach, F.; Trilling, D.

DOI
10.5117/CCR2020.1.003.LOEC

Publication date
2020

Document Version
Final published version

Published in
Computational Communication Research

3bij3 – Developing a framework for researching recommender systems and their effects

Felicia Loecherbach & Damian Trilling

CCR 2 (1): 53–79
DOI: 10.5117/CCR2020.1.003.LOEC

Abstract

Today's online news environment is increasingly characterized by personalized news selections, relying on algorithmic solutions for extracting relevant articles and composing an individual's news diet. Yet, the impact of such recommendation algorithms on how we consume and perceive news is still understudied. We therefore developed one of the first software solutions to conduct studies on effects of news recommender systems in a realistic setting. The web app of our framework (called 3bij3) displays real-time news articles selected by different mechanisms. 3bij3 can be used to conduct large-scale field experiments, in which participants' use of the site can be tracked over extended periods of time. Compared to previous work, 3bij3 gives researchers control over the recommendation system under study and creates a realistic environment for the participants. It integrates web scraping, different methods to compare and classify news articles, different recommender systems, a web interface for participants, gamification elements, and a user survey to enrich the behavioural measures obtained.

Keywords: news, recommender systems, computational social science, web application

News usage online has undergone considerable changes: Increasingly, the selection and presentation of news gets adapted to each user individually (Thurman & Schifferes, 2012) using *recommender systems*, algorithms that decide which articles are displayed to whom based on criteria such as past behaviour and/or ratings of similar users (Ricci, Rokach, & Shapira, 2011).

While these systems already form an integral part of news sites and social network sites, their impact on how we consume and perceive news is still understudied. Better understanding recommender systems is imperative for practitioners and academia: Media need insights into how editorial decisions can be combined with systems accommodating their audiences' wishes, while maintaining vital functions of journalism for democracy (Bhaskar, 2016; Schlesinger & Doyle, 2015). Communication researchers need a better understanding of how recommender systems affect selective exposure, political attitudes, and knowledge.

So far, the effect of algorithm-based selection on the diversity of news diets has mostly been discussed negatively, assuming that such systems limit the breadth of viewpoints and topics. However, recent studies challenge this conception by showing that, especially compared to other selection processes (e.g., by human editors), algorithms might not lead to more narrowed media diets after all (Möller, Trilling, Helberger, & van Es, 2018; Nguyen, Hui, Harper, Terveen, & Konstan, 2014). To provide researchers with a tool to contribute to this debate, this article sets out to develop one of the first research designs to tackle the issue of studying recommender systems in the context of news and political communication. We present a framework called *3bij3*. 3bij3 means three by three in Dutch, and signifies the most prominent feature of a news application developed for this purpose: It displays a 3 × 3 grid of the nine most relevant news articles, and allows to investigate different news recommenders and their impact on news usage and selection. The design derives from the necessity to use techniques from computational sciences to inform research of communicative phenomena, merging methodological innovation with theoretical approaches of political communication. It allows for implementing different selection mechanisms of news from various sources on-the-fly while tracking user behaviour. The resulting digital traces are enriched with information about the user through ratings and surveys. By that, the main contribution of this article is to offer a solution to particular challenges related to the study of recommender systems and their impact, answering the question:

How can news recommender systems and their effects be adequately researched?

Theoretical Background and Related Research

The consumption of news increasingly takes place online (Nic, Fletcher, Kalogeropoulos, Levy, & Nielsen, 2018). Instead of reading the printed

newspaper, consumers use a variety of sources to get information online, including search engines, social network sites and news websites. While incidental exposure via search engines plays an important role in news consumption, habitual usage including direct visits of news websites continues to be an important element of encountering information online (Möller, van de Velde, Merten, & Puschmann, 2019). On those news websites, different forms of personalization become increasingly prevalent. During the last decade, major news outlets significantly increased the usage of recommender systems on their platforms, tailoring the selection of articles to each individual user (Kunert & Thurman, 2019). All in all, the news consumption of today is increasingly driven by recommender systems that select and filter the information available.

While this personalization can have many beneficial effects for the user such as reduction of information overload, a commonly voiced criticism is that recommender systems also reduce the diversity of information that is encountered or consumed.

The Role of Diversity in News Recommender Systems

Diversity and Democracy

Before turning to the role of diversity specifically in news recommender systems, we briefly discuss its important role in democratic societies. In particular, a substantive body of literature has discussed the role of *homogeneity* versus *heterogeneity* in democratic settings. For instance, the homogeneity of strong ties regarding political viewpoints in someone's network is related to political participation (Lee, Kwak, & Campbell, 2015), and many have described how central the question whether media content is homogeneous or not is to communication research (see, e.g., Stroud, 2011, p. 172–173). In general, it is assumed that heterogeneous content is good for democratic discourse, and even though some have argued that exposure to opposing views can have diametrical effects, others have not been able to confirm this (Guess & Coppock, 2018). Consequently, diversity has been called 'a central value in public communication' (McQuail, 2007, p. 41).

While diversity can be examined at very different levels, including but not limited to the background of a sender, content characteristics such as genres, language, topics, viewpoints, and more (Helberger, Karppinen, & D'Acunto, 2018; McQuail, 2007), there are two aspects that are particularly prominent in discussions around the diversity of media content in democratic societies: viewpoint diversity and topic diversity. Exposure to a too narrow set of topics or viewpoints is depicted as dangerous for democracy

(Sunstein, 2009). While 'hearing the other side' as a central feature of democratic discourse implies a need for viewpoint diversity, the need of topic diversity is stressed as well by many. In particular, it is argued that democracy needs citizens that have at least basic knowledge about a broader range of topics, rather than just very specialized interests in only one or two pet topics (e.g., Geiß, Magin, Stark, & Jürgens, 2018; Moeller, Trilling, Helberger, Irion, & De Vreese, 2016). Given that viewpoint diversity is inherently tied to a specific topic, we will for now focus on the more basic topic diversity in this article: we consider a set of articles as more diverse when multiple topics occur (for a similar argument, see Haim, Graefe, & Brosius, 2018).

Diversity and Recommender Systems

In the context of recommender systems, the question of how diverse personalized content should be has been an often discussed and contested subject. On the one hand, it is a widespread fear that a reduction of diversity via personalization reduces the quality of democracies (see, e.g., the literature review by Zuiderveen Borgesius et al., 2016). Debates about 'filter bubbles' (Pariser, 2011) and 'echo chambers' (Sunstein, 2009), showing only content that reflects a users' past interests or that of their close friends, have sparked academic as well as public discussion. A vicious circle of personalization paired with selective exposure is expected to lead ultimately to fragmented societies, polarization and the spread of false information. In an environment of highly personalized media diets, it is argued that '[t]he overlap of issues, evidence, and arguments between citizens decreases' (Donsbach, 2014, p. 664).

On the other hand, the individual often has a need to reduce the complexity of all the information available to not overwhelm the individual (Haim et al., 2018). In fact, reducing this information overload or 'choice overload' (Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012) is seen as one of the most important features of recommender systems (Bozdag, 2013; Konstan & Riedl, 2012). Here, diversity of information also plays a crucial role due to the increase in volume and (possibly challenging) perspectives (Bawden & Robinson, 2009). News overload can lead to news fatigue and ultimately news avoidance – with one of the coping strategies being news curators recommending what to read to filter the information available (Song, Jung, & Kim, 2017). Simply providing the highest choice diversity (all news) might thus not enhance the exposure diversity (what is chosen). This does however not imply that diversity should be reduced to a minimum. It can be argued that 'people might be interested in things that they did not know they were interested in' (Bozdag, 2013, p. 217). Therefore,

having an algorithm present items from different topics could in some cases provide the user with more opportunities to explore their own interests than having them select the same content over and over again. While the question of how to develop a recommendation algorithm that can be beneficial for democracy is beyond the scope of our paper, Helberger (2019) has offered a conceptual framework to do exactly that – which then could, in fact, be empirically verified with the method we present in this paper.

Especially during the early years of news recommender development, the main focus was put on measures of precision and accuracy (see for example Bomhardt and Gaul (2005), or Bogers and Van den Bosch (2007)) when trying to extract items that are as close as possible to the user's profile. However, it has been shown that this measure is not enough to judge the quality of a recommender accurately as the danger of presenting the user with *too similar* items is inherent to this approach. The purpose of recommender systems lies not only in retrieving the best matching results but also showing a variety, including serendipitous items that surprise the users or allow them to be exposed to something unexpected (Kotkov, Wang, & Veijalainen, 2016). Therefore, it has become common to also include an element of diversity in recommending algorithms – while still providing accuracy and relevant items to the user (Bozdag, 2013; Bridge & Kelly, 2006).

The issue with studying recommender systems and crucial topics such as the role of diversity in algorithmic personalization is that it requires specific methodological efforts to research it. Questions about the extent to which users actually accept different recommenders and are affected by them cannot only be seen as mathematical problem to solve. In the following, different approaches to studying recommender systems are presented and their usability for studying questions of importance for communication science is discussed.

Past research on news recommendation

What types of recommender systems can be studied?

In general, recommenders can be divided into three types: content-based (also called 'semantic filtering', Möller et al., 2018), collaborative, and hybrid (Bridge & Kelly, 2006; Knijnenburg et al., 2012). Most recommendation systems rely on building a user profile based on explicit (e.g., ratings) or implicit (e.g., clicking behaviour) feedback – except for those that only offer recommendations on an item-to-item basis without taking the user's history into account (Möller et al., 2018).

Content-based systems 'recommend an item to a user based upon a description of the item and a profile of the user's interests' (Pazzani & Billsus, 2007, p. 339). They rely on identifying attributes of an element to judge how well it fits a user's profile. The items are annotated with specific features such as the topic and author of the article, or more formal categories such as length, making a comparison between item and user profile possible.

In contrast, collaborative recommenders use the profiles of similar users to infer which items fit a particular user. They 'automate the process of 'word-of-mouth' recommendations: items are recommended to a user based upon values assigned by other people with similar taste' (Bozdag, 2013). Identifying other users with similar taste is usually done by comparing users' reading histories and matching profiles based on behavioural patterns.

Both of the above approaches have shown to be influenced by data sparsity and the cold start problem of having no ratings or selections available at the beginning (Ricci et al., 2011). Some strategies can be used to overcome these problems: having the respondent choose from a set of topics and surveying specific attributes or using demographic similarity (age, gender, socioeconomic status etc.) as proximate measure for user similarity (Gupta & Gadge, 2015). Still, a large sample of participants is needed to make the implementation of collaborative recommenders feasible (Paliouras, Mouzakidis, Moustakas, & Skourlas, 2008) since it has to be ensured that close user profiles can be located. Otherwise 'a collaborative filtering algorithm experiences a lot of difficulties when trying to identify good neighbours in the system' (Victor, De Cock, & Cornelis, 2011, p. 656).

The most common form of recommenders today are hybrid recommenders, which combine features from both content-based and collaborative systems or other elements such as demographics, communities or editorial selections – for example to solve new-item problems in collaborative recommenders by integrating a content-based element (Ricci et al., 2011).

How have recommender systems been studied?

The evaluation of recommender systems can be done offline or online – by simulating user behaviour and interaction with the system or by conducting (experimental) user studies (Shani & Gunawardana, 2011). Offline approaches simulating user behaviour are focused on evaluating the functions and outcomes of algorithms (Bountouridis et al., 2019; Karakaya & Aytekin, 2017; Möller et al., 2018; Karimi, Jannach & Jugovac, 2018) by using measures such as accuracy, diversity, or novelty. While being a good approach for judging the performance of algorithms regarding certain

predefined measures and improving performance, such studies remain in an 'evaluation setting where recommendation approaches are compared without user interaction' (Ribeiro et al., 2014, p. 9). By doing that, they do not allow to research the questions more relevant for communication science: In how far is people's news selection and their perception of messages affected? How do algorithmic recommenders change our news consumption? Investigating and answering those questions does not necessarily call for designing the most effective and innovative algorithms – in fact it might even be especially informative to confront participants with 'failed' algorithms that lock them in filter bubbles or ignore their wishes – but to see how giving various degrees of information or user agency affects the news selection process.

For those questions, online evaluation is needed. In the computational sciences, user studies about recommendation algorithms usually have the aim to improve the algorithms' performance and find out which specific one users prefer (e.g., Garcin et al., 2014; Jonnalagedda, Gauch, Labille, & Alfarhood, 2016). However, they rarely incorporate more substantive questions related to selection or message effects or the interaction of the user with the system. Experimental studies in communication science that pay attention to those questions so far largely failed to incorporate realistic recommenders that go beyond displaying ideology-congruent articles based on an initial questionnaire (see also Beam, 2014; Dylko et al., 2017) or did not use recommenders but instead manipulated recommendation features (such as 'most-viewed' tags) (e.g., Messing & Westwood, 2014; Yang, 2016). Additionally, the so-called recommendations are usually only presented once or twice to the user. This limits the studies' actual value for researching the impact of recommendation algorithms as 'performance of most recommender systems evolves over time' (Ricci et al., 2011, p. 343), and an increasing familiarity of the users with the system also changes how they interact with it (Ricci et al., 2011). Lastly, the article content is mostly related to fictive scenarios and events, not taking into account that news consumption is inherently linked to getting up-to-date, relevant news (Garcin & Faltings, 2013; Karimi, Jannach & Jugovac, 2018). In other domains where recommendation algorithms are often used and researched, such as entertainment goods (books, movies), the aspect of timeliness is less of importance – while an old book might still be interesting, the news of yesterday is not (Karimi, Jannach & Jugovac, 2018).

Thus, incorporating (remotely) realistic algorithms with real-world data in an experimental setting that allows for testing media effects over a longer period of time is needed and currently lacking.

Towards studying recommender systems from a computational communication science viewpoint

The different approaches outlined above show varying strengths and limitations, each offering important insights. We propose to combine them, merging approaches from the domains of computational and social sciences to address existing limitations.

The usage of computational methods combined with communication science theories offers considerable advantages for addressing challenges in researching recommender systems. Computational methods facilitate the collection, processing, and enrichment of large-scale content data as well as behavioural data, which in combination can be used for insightful analyses. However, at the same time methodological challenges arise: Ethical questions of data collection, validity and reliability of data, and representativeness of findings are some of the most pressing issues (boyd & Crawford, 2012; van Atteveldt & Peng, 2018) – calling for bringing these methods to use in clearly as research identified settings (i.e., with explicit consent of users). Furthermore, acknowledging the gap between behavioural data and user experience (see Knijnenburg et al., 2012) makes it important to combine different data sources (including surveys and experiments) with behavioural data to get more precise insights into social phenomena (Shah, Cappella, & Neuman, 2015). Another crucial aspect is that analyses should be embedded in context and applied with profound theoretical knowledge to gather new insights and make substantial use of any kind of data (boyd & Crawford, 2012; Kitchin, 2014).

To evaluate the user's experience with and usage of news personalization – capturing the user's perspective – implicit as well as explicit feedback should be used. Information from the clickstream (i.e. which stories where selected) gives behavioural indicators of which articles the user selects and wants to read (Ricci et al., 2011). However, such passive measures should be enriched by explicit feedback through article ratings as 'implicit feedback is more difficult to interpret and potentially noisy' (Joachims et al., 2007) and behavioural measures are not always the best indicator of user interests (Ekstrand & Willemsen, 2016).

In this paper, we present a tool for studying recommender systems in the context of news consumption. The main features of the proposed application, derived from the shortcomings identified in past research, include: (1) The creation of an environment for the controlled testing of different types of algorithmic news recommenders, allowing for performance testing as well as user interaction; (2) Using real-time, actual news to test the

recommenders, enhancing ecological validity and modelling user experience while still in an experimental setting; (3) Inclusion of different elements of personalization – recommendation as well as customization – to enable testing their impact on user experience and diversity measures and; (4) Enrichment of behavioural data with information about the user and their feedback to more closely capture the user's perspective.

Developing the app

We built a news web application with different selection mechanisms. In the past, various attempts have been made at building websites for testing the impact of different forms of personalization. Frameworks such as PNS (Personalized News Service, Paliouras et al. (2008)), PEN (Personalized News, Garcin and Faltings (2013)) or the NZZ Companion (Leuener, 2017) have been developed to test real-time recommendation systems. While PNS was one of the first academic applications of using RSS feeds for gathering content, the last two systems were implemented in cooperation with specific websites, aiming at improving their traffic and revenue. Important insights can be gained from such frameworks (and many more not mentioned here). However, the main purpose of this paper is not to necessarily improve the recommenders and the application to draw as much traffic as possible but rather to research their effects on the user and their selections and take into account important content-wise dimensions (such as diversity).

The following sections are aimed at explaining the structure of the web application and its different parts in detail. It was developed within the Flask microframework (Grinberg, 2014), which allows for building the whole application in the Python programming language. It is divided into three parts: *Content retrieval, processing, and enrichment*, mostly taking place outside of the actual application, *Recommendation and customization*, describing the different mechanisms via which articles are selected, and *Flow of user interaction*, elaborating on the intended usage and functions of the application and the final questionnaire. The overall structure with all relevant elements is depicted in Figure 1. The code of the application can be found publicly accessible on GitHub[i]. It is deployed on a remote Linux server, hosting the application and the models necessary for text enrichment and comparisons as well as an ElasticSearch and MySQL database.

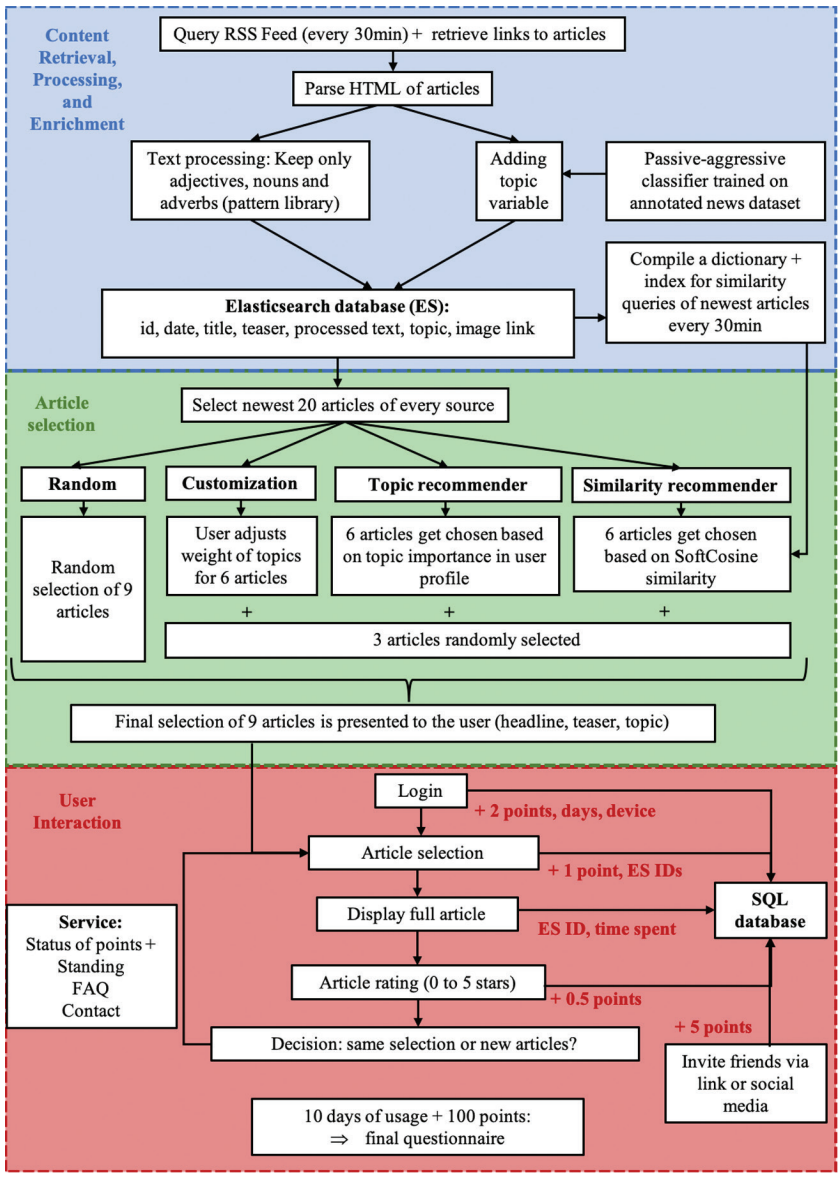

Figure 1. Overview of the framework.

Content Retrieval, Processing, and Enrichment

One of the most important issues to tackle when researching news personalization is the timely component of the content. This can be achieved by querying the RSS feeds of news sites (Trilling, 2014). We query the feeds of

pre-defined sources every thirty minutes, keeping the selection up-to-date, and used the scrapers written for a larger project (Trilling et al., 2018) to extract the whole content of the articles, including title, teaser, text, and pictures, by parsing the HTML content that we retrieved by following the links provided by the RSS feeds. The resulting data is saved in an ElasticSearch database. After this, several processing steps (such as using the pattern library (De Smedt & Daelemans, 2012) for only keeping adverbs, adjectives, and nouns) take place to structure the raw text, preparing it for subsequent algorithms. The processed text is saved as a separate key in the ElasticSearch database. Furthermore, each article gets enriched by assigning a topic to it, indicating its news section (e.g., sports, entertainment).

To create the topic tag, we chose to use supervised machine learning by applying a passive-aggressive classifier trained on data collected in another project, following the exact steps outlined in Burscher, Vliegenthart, and De Vreese (2015) to train the classifier. While the original dataset had 31 different issue categories, these were grouped into nine overarching categories for the application, derived from typical newspaper categories (domestic news, foreign news, economy, entertainment, crime, science, environment, immigration, sports). The trained classifier showed acceptable recall and accuracy (average F1 score[ii] of .68 with none of the categories below .55, indicating sufficient performance for automated coding of article topics. The trained topic classifier is used to assign a topic key to each retrieved article in the database, after processing the text as described in Burscher et al. (2015) and transforming it with a tf·idf vectorizer.

Recommendation and Customization

Currently, the application includes four different groups of selection mechanisms for articles: one random group, one group with customization and two recommendation algorithms. The recommenders used in this specific test instance of the framework are limited to content-based recommenders – but could be extended to other kinds of algorithms with collaborative elements when given a larger sample size and longer testing period. The different groups allow for comparing different types of personalization. They are all used to select nine stories out of the most recent articles in the ElasticSearch database (the last 20 published by each news site). The amount of nine articles reduces the choice for the user to a manageable amount and allows for presenting all options on one screen – an important factor to consider due to selection and positioning effects (Teppan & Zanker, 2015).

For the first group, a random sample is retrieved, making this a baseline or control group. In the other conditions, three articles are retrieved randomly (updated every selection round) while the other six are selected based on specific rules. The ratio of six to three was chosen to (1) have a clear majority of the articles selected by the recommendation algorithm or customization to see its effects but also to (2) give enough other options to make it possible for respondents to choose non-recommended content. This procedure ensures that there always is an element of serendipity and randomness in the selection, avoiding trapping the user in an impermeable bubble that only shows content and topics that are very similar to the past selections. However, adjusting those ratios when further testing the system and interpreting the results should be taken into account.

Customization

The customization condition by default also randomly displays all articles – until the user actively intervenes. This can be done by selecting between one and three topical categories in the side-menu that subsequently appear more often. The user can choose between nine different generic news topics (such as domestic politics, foreign politics, economy, sports, entertainment). After the choice, the displayed articles change as follows: (1) Three random articles are displayed, and (2) the remaining six articles are drawn from the selected topic categories. When the person selected one topic, all six user-selected articles are from this topic. Choosing two topics leads to three articles each from those categories, etc.

Which specific articles from the topics are displayed is chosen randomly. These topic settings remain in place until the user changes them again.

From a methodological point of view, one could object that this violates the principle of randomization in social-scientific experiments. After all, a participant's selection directly influences the treatment they will receive. Contrarily, more and more studies on selective exposure advocate such designs (e.g., Arceneaux & Johnson, 2013; Gaines & Kuklinski, 2011; Trilling, Van Klingeren, & Tsfati, 2017). They argue that other designs, in which participants are forced to make a selection they would never have to make in real life, make it hard to draw valid conclusions. This seems even more true in the context of recommender systems.

The customization condition mimics what is quite common in current news aggregation services and news websites where users can select or follow topics – but it does so on the most basic and most obvious level. This makes it easy to understand for the user, giving immediate visible feedback for every change and action of controlling the system. This allows to study

how variables such as user agency and the feeling of being in control affect satisfaction with the system and interaction with it. Because it may be the case that the very nature of 'being in control' increases user satisfaction, simple between-group comparisons are not sufficient. It is necessary to systematically investigate whether – when the same actual articles are displayed – satisfaction differs, depending on whether the selection was based on user input or, for instance, on random selection. Another option would be to compare two groups in which the customization interface is available, but where only in one group the article selection is actually based on the choices the users made. More generally, next to straightforward questions asking about user satisfaction, more subtle and sophisticated ways of tapping into user experiences need to be explored.

Topic-based recommender

The first recommender condition is based on a similar principle, however without explicit user action but rather using implicit measures (i.e. past selections) to infer preferences for topics. In order to build an initial (very small) user profile that gives an indication of interests, the first three times the respondent is shown the news website only a random selection is displayed and the choices made are recorded. After this, those initial choices are used to recommend other items to the user. Firstly, the topics of all articles the user has selected in the past are retrieved. Secondly, three topics are randomly selected from this list, giving frequently appearing topics a higher chance of getting into the final selection. For each of the three topics, two articles are shown. Thus, in case a user only selected articles from one topic in the past they get shown six articles of that topic, with more diverse past preferences the selection also gets more diverse. Due to the random element included it is however not possible to get completely 'locked' for the user – there are always stories that were not selected based on past behaviour and those are updated with every iteration. Thus, even if the initial selections did not include articles of interest for the user, a third of articles presented still remain randomly selected, offering ample opportunities to explore other topics. Apart from specifically selecting certain topics to 'trick' the algorithm in case its workings are figured out by the user, no active control over the selection of stories is given. This behaviour could indicate a distrust in algorithmic personalization or a dissatisfaction with the lack of user agency given in the study – and thus be a very interesting avenue to explore when studying how users interact with phenomena such as 'filter bubbles'.

Similarity-based recommender

The third group deviates from the usage of the topic variable in determining the most relevant articles for the user. Instead, it uses word vectors for determining similarity between documents. The general procedure has been applied to recommender systems in the form of pairwise cosine similarity in various studies. For this, each document is represented in terms of the Vector Space Model (VSM) as a vector of term weights, and the similarity between two documents is estimated by taking the cosine of the angle between the vectors.

However, this approach suffers from limitations with regard to similarity detection between documents: The VSM features are considered to be independent – thus two words are seen as entirely different, no matter what they are. This is a problem when we want to infer what an article is about. As Sidorov, Gelbukh, Gómez-Adorno, and Pinto (2014, p. 492) explain: 'For example, words "play" and "game" are of course different words and thus should be mapped to different dimensions in SVM [*sic*]; yet it is obvious that they are related'. Thus, Sidorov et al. (2014) introduced a new measure, termed 'soft cosine measure' (p. 491) which can be used to calculate the soft similarity between documents. It makes use of word vectors that have to be derived from a larger training corpus into the original cosine similarity formula. By this, the equivalency of words can more accurately be detected since it tries to take the context words appear in into account.[iii]

The Python library *gensim* (Řehůřek & Sojka, 2010) offers an implementation of the soft cosine measure. To use it, word vector embeddings are needed, produced by a word2vec model, a technique developed by Mikolov, Corrado, Chen, and Dean (2013).

We used word embeddings that were trained on a dataset of all the print issues from several Dutch newspapers (De Telegraaf, NRC Handelsblad, De Volkskrant, Algemeen Dagblad) between 2000 and 2015, thus overall representing the Dutch newspaper market[iv]; however, one can easily use different ones, for instance in another language. The gensim implementation allows for creating a sparse similarity matrix which serves as input for making an index for similarity queries (i.e. retrieving the most similar articles to the articles a user has read before). A Python script was used to perform all the steps necessary for this every 30 minutes on the newest content available in the database – outside of the application since it otherwise would severely impact performance and user interaction with the system.

In the application, on every request the past articles a user read are retrieved and for each of these articles, the three most similar new articles are determined. Now having a list containing three new articles per past article,

the most frequently occurring articles are selected and subsequently presented to the user. This procedure can be seen as superior to just averaging the similarities of one article with all past articles to find the most fitting ones, as similarities could 'cancel each other out' – if a person for example has an interest in sports and politics, a new sports article could show high similarity to a sports article read in the past but low similarity to a politics article, leading to medium similarity on average. Selecting the articles that appear to be the most similar to the highest number of past articles is seen as the best solution to map best the users' past behaviour.

Flow of User Interaction

While the user is interacting with the application, their behaviour is recorded at several steps along the way. After a registration including ethical consent, the user can login on the website using a username and password. Each log in is saved to a MySQL database, including information about the device being used to access the website. This allows for selecting the appropriate design for screen size by having a responsive layout ensuring the highest flexibility for the users. After this, the user is led to the main page of the application, showing nine articles either in a 3x3 format or below each other (Figure 2).

Figure 2. User interface main page (left: desktop, right: mobile) showing the navigation menu on the left side and the selection of stories (with colored topic tag, headline, and short teaser) on the right side.

For every article, the title and a short teaser are displayed as well as a colour-marked indication of the topic category on top of the article. In a side bar, options for getting information about the project (FAQs), contacting

the researcher, checking the percentage of completion of the study, and inviting friends to use the application are given. Additionally, the customization group gets the option to select topics from nine different categories. When selecting an article, the user is taken to a detail page where title, time of publication, teaser, picture, and text are presented. Below the article, a 5-star rating system is given to collect instant feedback about the article from the user. Lastly, in case the display of the article did not work as intended (e.g. due to scraping errors), the user is given the option to report the story instead of rating it. This way, low ratings due to faulty presentation (which could lead to misinterpretation) are prevented. All the different actions are recorded in the MySQL database and, adding an element of gamification, the user receives points for all actions (as displayed in Figure 1).

Points are rewarded for logging in, selecting articles, rating articles, and inviting friends to use the application – with limitations on the number of points one can get per day for logins and reading stories to ensure that participants are not able to collect all points in one go. Only if the application is used for a certain period of time (at least on 10 different days) and enough interaction has taken place (100 points), does a link to the final questionnaire appear on the website, allowing the participant to finish the study. The main purpose of the point system here is to give the participants an indication of how far along with the study they are and to also show them their standing in relation to other users (element of competitive gamification). This general points system could also be expanded and be used for tracking other relevant factors, i.e. by giving points for time spent reading articles. Furthermore, it could also be used for giving (extra) rewards to further improve interaction.

To finish the study, participants have to fill in a questionnaire that allows us to measure, for instance, their satisfaction with the system or any outcome variable (next to their selection behaviour, which is automatically stored) the researcher is interested in.

An important aspect of the overall design is its modularity: Each step described above can be replaced by a different strategy or implementation while keeping the general structure intact. One could for example think of changing the way of retrieving and annotating articles by collaborating with media outlets and using different classifiers. In particular, this means that it is simple to add or replace language-specific elements for studies in different locales. Likewise, the proposed content retrieval method could be combined with various selection mechanisms and algorithms in the second step. Thus, the elements as presented here can be seen as an outline of the different modules of the design – open to flexible adaptation to different research questions and contexts.

Testing and Further Adaptation Hard- and Software

Requirements

Two different forms of testing were used to test the application: During a period of 8 weeks in April to June 2018 the application was first pre-tested to identify possible problems and improve the visual presentation. 19 participants used the application for two weeks, gathering an amount of 100 points to finish the study. Taking a closer look at the soft- and hardware required to implement the application for this period, the main elements to run it continuously on a remote server are described in Grinberg (2014), including the nginx, gunicorn and supervisor packages. In addition, the two databases (ElasticSearch, MySQL) and their memory and storage usage are important factors to consider.

For this particular instance of the application, the online versions of four major Dutch newspapers (Algemeen Dagblad, De Telegraaf, NRC Handelsblad, De Volkskrant) and one online-only news website (nu.nl) were chosen as sources to ensure a broad range of topics and sufficient supply of up-to-date content. Overall, between 114 and 749 stories were collected per day, summing up to around 40,000 documents after eight weeks, taking 7 GB of space on the disk. In contrast, the MySQL database took much less space (200 MB). In addition, the application itself needs around 1.6 GB (mostly due to the word2vec model and vectorizer needed for content annotation). Considering all additional files such as system files, libraries and other software, 60-70 GB of disk space would be needed to ensure that the application can run on this small scale for a year. Furthermore, 10 GB RAM and one core have been proven sufficient to handle the workload. Overall, the application ran smoothly the whole testing period, with only few reported articles by the users (mainly due to minor style issues) and no further complaints, proving that the application is a working system.

In a second step, automated testing using the web automation tool Selenium[v] was used to run several scripts with the softcosine algorithm to 'stress-test' the performance of the website and get insights into whether a long-term usage with larger samples would be possible. For this, a different configuration with 64 GB RAM and 16 cores was used to model a research setting aimed at higher sample sizes. The results are depicted in Figure 3. In total, 40 different selenium agents were run simultaneously. Given that users will not use the system at exactly the same time, this can be seen as the equivalent of having a sample of hundreds of users.

Every session, an agent selected 6 stories and every time retrieved a new set of articles, resulting in retrieving 54 stories each session. After that, they paused for 2 hours to approximately model what could be normal browsing

behaviour. This behaviour was repeated for 10 sessions to reach a selection of 60 articles per agent. The loading times (in milliseconds) for retrieving new sets of stories (the point where the different databases and calculations are heaviest) were recorded.

As can be seen, the number of milliseconds it takes to load the news stories remains constant between the different sessions, with slight increases for the last two selections within each session. Thus, for a small- to midsize usage with up to 40 users accessing the website and continuously requesting new stories no visible deceleration of the website occurs to participants. At least after selecting 60 stories – enough to successfully end the study within a period of two weeks (540 stories retrieved per agent; 21,600 stories in total), participants to not experience loading times of more than one second for the first four new retrievals of stories, going up to under 4 seconds if even more is read in one session. However, it might be necessary to use different configurations for running the application for large-size samples (>200) and over longer periods of time (>6 months), requiring more computing power and memory. Nonetheless, it has been shown that in this particular research setting the waiting times for participants in the condition that requires most computational power do not go above on average 4 seconds. Thus, the tool is also feasible for using more advanced recommendation algorithms over a longer period of time with many active participants.

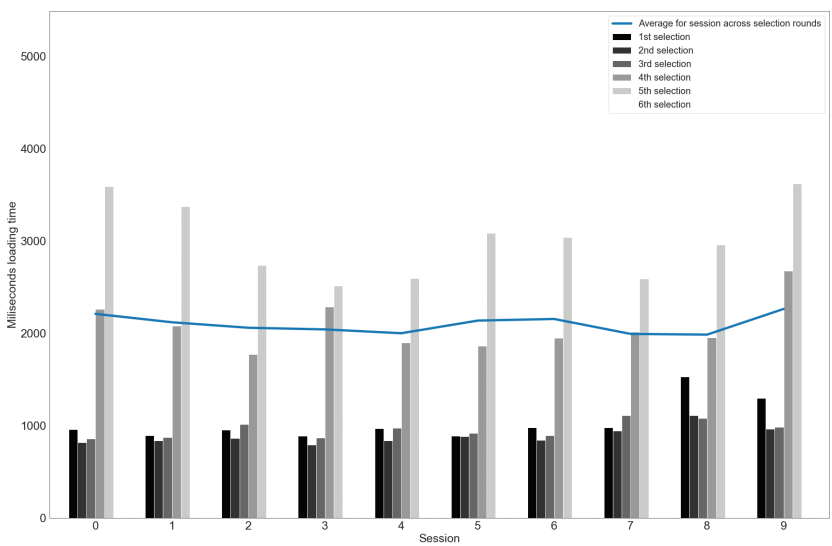

Figure 3. Automated performance testing.

Adapting 3bij3 for different studies

3bij3 can be easily adapted to be used in different contexts. In particular, we made sure that each part of 3bij3 can be easily modified by editing configuration files, or by replacing whole components. In this section, we discuss the most important modification possibilities to give other researchers a guideline how to use the tool in future research.

Adapting the news sources

3bij3 retrieves its articles from an ElasticSearch database. We used the tool Inca (Trilling et al., 2018) to continuously populate this database with articles retrieved via a combination of RSS-feeds and web scraping, and Inca makes it easy to add any custom scraper. But also, any other scraper, for any other news source, could be used, as long as the scraped data are then put into an ElasticSearch database. In essence, as long as title, text and potentially an image link for all articles that 3bij3 needs to access are available, storing them in ElasticSearch is sufficient to make 3bij3 use them.

This also implies that it does not matter in which language the articles are written. The only language-specific resource is a file with pre-trained word embeddings in word2vec format. We used Dutch-language embeddings provided by Kroon et al. (2019), but pre-trained embeddings in all major languages are readily available, and some are even included in libraries such as gensim (Řehůřek & Sojka, 2010) or spacy (Honnibal & Montani, 2017). For language-specific settings, such as stopword removal or stemming, the language can be provided in 3bij3's settings.

Adapting the visual appearance

In order to change the language of the user interface, or to customize the layout, the user simply can change the HTML and CSS templates used by 3bij3, which should be straightforward to do. If researchers want to present additional cues (source cues, popularity cues, . . .), they need to adapt the HTML code and make sure that the corresponding key is present in the ElasticSearch database. This only requires minimal adjustments to 3bij3's code.

Adapting the recommendation algorithms

3bij3 imports its underlying recommendation algorithms from one Python file that provides implementations for the recommender systems described in this article. If researchers wish to test the effects of a different recommendation algorithm, they can modify this file accordingly. Admittedly, this asks more knowledge than the modifications described in the previous

paragraphs, but still should be no insurmountable hurdle for computational social scientists with moderate levels of programming knowledge. This means that 3bij3 frees the researcher from the burden of constructing a user interface, a data handling backend, and so forth – all of this is provided. Instead, researchers can focus on implementing the algorithm to be studied.

Conclusion: A Working System With a Research Agenda

This paper set out to propose an overarching framework for studying different forms of personalization and especially recommender systems online in an ecologically and externally valid way. It addressed several challenges that became apparent from past research, especially in the context of news and political communication: So far, the communication science perspective remained rather limited with regard to including actual, realistic recommendation algorithms in experimental settings, as for example the studies by Beam (2014) and Dylko et al. (2017) showed. This issue was addressed by building a web application in which the articles shown to the user are presented in a realistic setting with the possibility to implement various types of recommendation algorithms for testing. This makes the framework and application very flexible and usable for various questions concerning (news) recommender systems and their effects. Furthermore, the amount of content (and different cues) shown to the user can be efficiently controlled for and varied if necessary, preserving the experimental character of this research.

Going one step further, the actual latest articles from different newspapers are retrieved, processed, and enriched in an automated way to be employed in the application. By that, participants are no longer presented with mock stories far away from their normal news consumption, but with up-to-date content that actually is of interest to them. This can be seen as a crucial factor for effectively studying news recommendations (Garcin & Faltings, 2013).

The approach we proposed helps studying this impact from different theoretical perspectives, and can contribute to several streams of communication science. In the gatekeeping literature, the notion of 'algorithmic gatekeeping' (Napoli, 2015) has emerged, sparking both empirical studies and even calls for a 'normative evaluation of machines as news gatekeepers' (Nechushtai & Lewis, 2019, p. 303). Thorson and Wells (2016) suggest to speak of 'curated flows', in which not only journalists, but also citizens, different stakeholders, and algorithms perform gatekeeping functions (see

also Bruns, 2018). With our suggested method, we hope to have offered an approach to scholars interested in algorithmic gatekeeping that allows them to study its usage and effects in a more controlled environment than when studying existing black-box algorithms like Google News and similar, without compromising the realistic setting that cannot be achieved in traditional lab experiments with simulated stimuli. The same holds true for scholars interested in the interplay between algorithms and selective exposure (see Möller et al., 2018). But also framing scholars can benefit from our approach: Peperkamp and Berendt (2018) highlight that news recommender systems can influence the propagation of different frames. To the best of our knowledge, there are no empirical studies yet that test framing effects in algorithmically personalized news media environments. Lastly, the proposed tool can be used to advance studying the phenomenon of algorithms reducing the diversity of the news diet: While simulation frameworks have investigated the capacity of algorithms to reduce diversity (Bountouridis et al., 2019) and Bodó, Helberger, Eskens, and Möller (2019) note that users (at least in the Netherlands) are overall not diversity-averse in nature, they cannot give insights into the actual behaviour of users when being exposed to algorithmically selected news. Testing user behaviour when being confronted with a 'filter bubble' (i.e. a diversity-reducing algorithm) can be seen as one of the practical applications of the proposed tool.

The short evaluation above showed that the aims of the overall framework could be implemented with the first prototype of the application. The general system is working and the basic elements form a solid structure to build on (usable on a broad variety of devices, browsers, and operating systems) and can in the future be applied to large-scale field experiments over extended periods of time. However, to further develop and use it for substantial analyses, several measures have to be taken, apart from more practical advices as given above: Firstly, a development of more advanced recommendation algorithms modelling specific aspects such as diversity need to be implemented – a step requiring collaboration with other fields such as computational sciences (articles discussing the implementation of diversity-sensitive recommenders are for example Karakaya and Aytekin (2017), Bridge and Kelly (2006)) as well as building on a larger sample of respondents. This would allow for collaborative and hybrid recommenders and incorporating user feedback to better model the algorithms actually being used by news companies to research their effects.

Secondly, when using the application on a larger scale, collaborations with news media (as was done by Garcin and Faltings (2013) or Leuener (2017)) could be a good way forward – or, to prevent narrowing the content

to one specific outlet, redirecting the participants to the actual source websites (such as for example Paliouras et al. (2008) did). By that, the news usage experience would also become more realistic while solving issues with scraping and parsing content from websites with ever-changing layouts – however, with the downside of partially losing control over the experimental setting, given that participants leave the website and are exposed to other links or stories that differ between the various outlets.

Thirdly, when using our application, researchers need to think about legal implications, in particular copyright restrictions. We are not in the position to give legal advice at this place, but would like to provide some suggestions. Most importantly, as Van Atteveldt, Strycharz, Trilling, and Welbers (2019) point out, in many jurisdictions, research exceptions may be applicable. Next to that, in many jurisdictions, snippets (like those provided by search engines or previews when posting on social media) are not subject to copyright laws. One could therefore think of just providing these snippets, and then redirect users to the original outlets. While this restricts the functions 3bij3 can offer (such as direct user feedback on an article), it still may be enough for some use cases. Also, it can be worth exploring a collaboration with one or more news providers. Given that many publishing houses are working on the development of news recommender systems as well, this can be a win-win situation.

Lastly, a resulting improved application would have to be used by participants over a longer period of time (several months or more) to actually capture changes in aspects such as content diversity which are assumed to take longer to come into effect (Möller et al., 2018) and meaningfully evaluate recommendation algorithms. That said, the proposed framework offers a great flexibility for testing different recommenders and their effects in a realistic setting, giving the opportunity to further explore the effects of tailored news environments on the news diet and perception of individuals.

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperation. We would like to thank Rens Vliegenthart who gave us access to the annotated dataset from the NWO-VENI project 'The contingency of media's impact on national parliaments: a comparative study' (https://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/26/5226.htm), which allowed us to re-implement the classifier described in Burscher et al. (2015).

Notes

i https://github.com/FeLoe/3bij3
ii The F1 score is defined as the harmonic mean of recall and precision.
iii The usage of word embeddings has been proposed as specifically useful for news recommendation systems due to its potential to capture implicit semantics and word context (Peng, Liu, & Lin, 2016). The usage of neural networks for recommender systems is becoming increasingly popular (Zhang, Yao, Sun, & Tay, 2019), making this recommender an example for a more realistic recommendation algorithms.
iv The word embeddings were provided by Anne Kroon (also described in Kroon et al. (2019)).
v https://selenium-python.readthedocs.io

References

Arceneaux, K., & Johnson, M. (2013). *Changing minds or changing channels? Partisan news in an age of choice*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226047447.001.0001

Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, *35*(2), 180–191. doi: 10.1177/0165551508095781

Beam, M.A. (2014). Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, *41*(8), 1019–1041. doi: 10.1177/0093650213497979

Bhaskar, M. (2016). In the age of the algorithm, the human gatekeeper is back. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/sep/30/ age-of-algorithm-human-gatekeeper

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital Journalism, 7*(2), 206–229. https://doi.org/10.1080/21670811.2018.1521292

Bogers, T., & Van den Bosch, A. (2007). Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems* (pp. 141–144). Minneapolis, MA: ACM Press. doi: 10.1145/1297231.1297256

Bomhardt, C., & Gaul, W. (2005). NewsRec, a personal recommendation system for news websites. In C. Weihs & W. Gaul (Eds.), *Classification—the ubiquitous challenge* (pp. 394–401). Berlin: Springer. doi: 10.1007/3-540-28084-7_45

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., & Hauff, C. (2019). SIREN: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 150–159). Atlanta, GA: ACM Press. doi: 10.1145/3287560.3287583

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. doi: 10.1080/1369118X.2012.678878

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. doi: 10.1007/s10676-013-9321-6

Bridge, D., & Kelly, J.P. (2006). Ways of computing diverse collaborative recommendations. In *Lecture Notes in Computer Science* (pp. 41–50). Berlin: Springer. doi: 10.1007/11768012_6

Bruns, A. (2018). *Gatewatching and news curation: Journalism, social media, and the public sphere*. New York: Peter Lang. doi: 10.3726/b13293

Burscher, B., Vliegenthart, R., & De Vreese, C.H. (2015). Using supervised machine learning to code policy issues. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 122–131. doi: 10.1177/0002716215569441

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(6), 2063–2067.

Donsbach, W. (2014). Journalism as the new knowledge profession and consequences for journalism education. *Journalism*, *15*(6), 661–677. doi: 10.1177/1464884913491347

Dylko, I., Dolgov, I., Hoffman, W., Eckhart, N., Molina, M., & Aaziz, O. (2017). The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure. *Computers in Human Behavior*, *73*, 181–190. doi: 10.1016/j.chb.2017.03.031

Ekstrand, M.D., & Willemsen, M.C. (2016). Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 221–224). Boston, MA: ACM Press. doi: 10.1145/2959100.2959179

Gaines, B.J., & Kuklinski, J.H. (2011). Experimental estimation of heterogeneous treatment effects related to self-selection. *American Journal of Political Science*, *55*(3), 724–736. doi: 10.1111/j.1540-5907.2011.00518.x

Garcin, F., & Faltings, B. (2013). PEN recsys. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge* (*NRS*) (pp. 3–9). New York, NY: ACM Press. doi: 10.1145/2516641.2516642

Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., & Huber, A. (2014). Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems* (pp. 169–176). Foster City, CA: ACM Press. doi: 10.1145/2645710.2645745

Geiß, S., Magin, M., Stark, B., & Jürgens, P. (2018). "Common Meeting Ground" in Gefahr? Selektionslogiken politischer Informationsquellen und ihr Einfluss auf die Fragmentierung individueller Themenhorizonte. *Medien & Kommunikationswissenschaft*, *66*(4), 502–525. doi: 10.5771/1615-634X-2018-4-502

Grinberg, M. (2014). *Flask web development*. Sebastopol, CA: O'Reilly Media.

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science, online first*. doi: 10.1017/S0007123418000327

Gupta, J., & Gadge, J. (2015, Jan). Performance analysis of recommendation system based on collaborative filtering and demographics. In *International Conference on Communication, Information Computing Technology* (pp. 1–6). doi: 10.1109/ICCICT.2015.7045675

Haim, M., Graefe, A., & Brosius, H.B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism*, *6*(3), 330–343. doi: 10.1080/21670811.2017.1338145

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, online first*. doi: 10.1080/21670811.2019.1623700

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information Communication and Society*, *21*(2), 191–207. doi: 10.1080/1369118X.2016.1271900

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems* (*TOIS*), *25*(2), 1–27. doi: 10.1145/1229179.1229181

Jonnalagedda, N., Gauch, S., Labille, K., & Alfarhood, S. (2016). Incorporating popularity in a personalized news recommender system. *PeerJ Computer Science*, *2*. doi: 10.7717/peerj-cs.63

Karakaya, M.Ö., & Aytekin, T. (2017). Effective methods for increasing aggregate diversity in recommender systems. *Knowledge and Information Systems*, *56*(2), 1–18. doi: 10.1007/s10115-017-1135-0

Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems–Survey and roads ahead. *Information Processing & Management*, *54*(6), 1203–1227. doi: 10.1016/j.ipm.2018.04.008

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. doi: 10.1177/2053951714528481

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504. doi: 10.1007/s11257-011-9118-4

Konstan, J.A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1/2), 101–123. doi: 10.1007/s11257-011-9112-x

Kotkov, D., Wang, S., & Veijalainen, J. (2016). Knowledge-Based systems: A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*, 180–192. doi: 10.1016/j.knosys.2016.08.014

Kroon, A., Trilling, D., Fokkens, A., Loecherbach, F., Moeller, J., van Atteveldt, W., & van der Velden, M. (2019). Deriving semantics from dutch media corpora: The Amsterdam word embedding model. Paper presented at *Etmaal van de communicatiewetenschap*. Nijmegen, Netherlands.

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, 1–22. doi: 10.1080/17512786.2019.1567271

Lee, H., Kwak, N., & Campbell, S.W. (2015). Hearing the other side revisited: The joint workings of cross-cutting discussion and strong tie homogeneity in facilitating deliberative and participatory democracy. *Communication Research*, *42*(4), 569–596. doi: 10.1177/0093650213483824

Leuener, R. (2017). NZZ Companion: How we successfully developed a personalised news application. *Medium*. Retrieved from https://medium.com/@rouven.leuener/nzz-companion-how-we-successfully-developed-a-personalised-news-app-d3c382767025

McQuail, D. (2007). Revisiting diversity as a media policy goal. In W. A. Meier & J. Trappel (Eds.), *Power, performance and politics: Media policy in Europe* (pp. 41–57). Baden-Baden: Nomos. doi: 10.5771/9783845202938-41

Messing, S., & Westwood, S.J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, *41*(8), 1042–1063. doi: 10.1177/0093650212466406

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations* (*ICLR*), 1–12. doi: 10.1162/153244303322533223

Moeller, J., Trilling, D., Helberger, N., Irion, K., & De Vreese, C. (2016). Shrinking core? Exploring the differential agenda setting power of traditional and personalized news media. *Info*, *18*(6). doi: 10.1108/info-05-2016-0020

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication and Society*, 1–19. doi: 10.1080/1369118X.2018.1444076

Möller, J., van de Velde, R.N., Merten, L., & Puschmann, C. (2019). Explaining online news engagement based on browsing behaviour: Creatures of habit? *Social Science Computer Review*, 1–17.doi: 10.1177/0894439319828012

Napoli, P.M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy*, *39*(9), 751–760. doi: 10.1016/j.telpol.2014.12.003

Nechushtai, E., & Lewis, S.C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298–307. doi: 10.1016/j.chb.2018.07.043

Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 677–686). doi: 10.1145/2566486.2568012

Nic, N., Fletcher, R., Kalogeropoulos, A., Levy, D.A., & Nielsen, R. K. (2018). *Reuters institute digital news report 2018*. Reuters Institute for the Study of Journalism. Retrieved from http://www.digitalnewsreport.org/survey/2018

Paliouras, G., Mouzakidis, A., Moustakas, V., & Skourlas, C. (2008). PNS: A personalized news aggregator on the web. *Studies in Computational Intelligence*, *104*, 175–197. doi: 10.1007/978-3-540-77471-6_10

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin. doi: 10.3139/9783446431164

Pazzani, M.J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Berlin: Springer. doi: 10.1007/978-3-540-72079-9_10

Peng, H., Liu, J., & Lin, C.Y. (2016). News citation recommendation with implicit and explicit semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 388–398). doi: 10.18653/v1/P16-1037

Peperkamp, J., & Berendt, B. (2018). Diversity checker. In *Adjunct publication of the 26th Conference on User Modelling, Adaptation and Personalization* (*UMAP*) (pp. 35–41). New York, NY: ACM Press. doi: 10.1145/3213586.3226208

Ribeiro, M.T., Ziviani, N., Moura, E.S.D., Hata, I., Lacerda, A., & Veloso, A. (2014). Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology*, *5*(4), 1–20. doi: 10.1145/2629350

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1–35). Boston, MA: Springer US. doi: 10.1007/978-0-387-85820-3_1

Schlesinger, P., & Doyle, G. (2015). From organizational crisis to multi-platform salvation? Creative destruction and the recomposition of news media. *Journalism: Theory, Practice & Criticism*, *16*(3), 305–323. doi: 10.1177/1464884914530223

Shah, D.V., Cappella, J.N., & Neuman, W.R. (2015). Big Data, digital media, and Computational Social Science. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi: 10.1177/0002716215572084

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257–297). Boston, MA: Springer US. doi: 10.1007/978-0-387-85820-3_8

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computacion y Sistemas*, *18*(3), 491–504. doi: 10.13053/CyS-18-3-2043

Song, H., Jung, J., & Kim, Y. (2017). Perceived news overload and its cognitive and attitudinal consequences for news usage in South Korea. *Journalism & Mass Communication Quarterly*, *94*(4), 1172–1190. doi: 10.1177/1077699016679975

Stroud, N.J. (2011). *Niche news: The politics of news choice*. New York, NY: Oxford University Press.

Sunstein, C.R. (2009). *Republic. com 2.0*. Princeton, NJ: Princeton University Press.

Teppan, E.C., & Zanker, M. (2015). Decision biases in recommender systems. *Journal of Internet Commerce*, *14*(2), 255–275. doi: 10.1080/15332861.2015.1018703

Thorson, K., & Wells, C. (2016). Curated flows: A framework for mapping media exposure in the digital age. *Communication Theory*, *26*(3), 309–328. doi: 10.1111/comt.12087

Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites: Lessons from a longitudinal study. *Journalism Studies*, *13*(5-6), 775–790. doi: 10.1080/1461670X.2012.664341

Trilling, D. (2014). Weg vom manuellen Speichern: RSS-Feeds in der automatisierten Datenerhebung bei Onlinemedien. In K. Sommer, M. Wettstein, W. Wirth, & J. Matthes (Eds.), *Automatisierung in der Inhaltsanalyse* (pp. 73–89). Köln: Herbert von Halem.

Trilling, D., Van Klingeren, M., & Tsfati, Y. (2017). Selective exposure, political polarization, and possible mediators: Evidence from the Netherlands. *International Journal of Public Opinion Research*, *29*(2), 189–213. doi: 10.1093/ijpor/edw003

Trilling, D., van de Velde, B., Kroon, A.C., Loecherbach, F., Araujo, T., Strycharz, J., Jonkman, J. (2018). INCA: Infrastructure for content analysis. In *14th International Conference on e-Science* (*IEEE*) (pp. 329–330). Amsterdam. doi: 10.1109/eScience.2018.00078

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Toward open computational communication science: A practical roadmap for reusable data and code. *International Journal of Communication*, *13*(1), 3935–3954.

Van Atteveldt, W., & Peng, T.Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 1–12. doi: 10.1080/19312458.2018.1458084

Victor, P., De Cock, M., & Cornelis, C. (2011). Trust and recommendations. In *Recommender systems handbook* (pp. 645–675). Boston, MA: Springer. doi: 10.1007/978-0-387-85820-3_20

Yang, J. A. (2016). Effects of popularity-based news recommendations (most-viewed) on users exposure to online news. *Media Psychology*, *19*(2), 243–271. doi: 10.1080/15213269.2015.1006333

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys* (*CSUR*), *52*(1), 5. doi: 10.1145/3285029

Zuiderveen Borgesius, F.J., Trilling, D., Möller, J., Bodó, B., de Vreese, C.H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1), 1–16. doi: 10.14763/2016.1.401

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

About the authors

Felicia Loecherbach works at the Department of Communication Science, Vrije Universiteit Amsterdam.
Correspondence address: Vrije Universiteit Amsterdam, Faculty of Social Sciences, Department of Communication Sciences, De Boelelaan 1105, 1081 HV Amsterdam (f.loecherbach@vu.nl)

Damian Trilling works at the Amsterdam School of Communication Research, University of Amsterdam.
Correspondence address: Postbus 15791, 1001 NG Amsterdam (d.c.trilling@uva.nl)