Experimental measures of news personalization in Google News*

Vittoria Cozza[1,2], Van Tien Hoang[3],
Marinella Petrocchi[1], and Angelo Spognardi[1,4]

[1] IIT-CNR, Pisa, Italy
[2] DEI, Polytechnic University of Bari, Italy
[3] IMT School for Advanced Studies, Lucca, Italy
[4] DTU Compute, Lingby, Denmark
{v.cozza,m.petrocchi,a.spognardi}@iit.cnr.it,
vantien.hoang@imtlucca.it

Abstract. Search engines and social media keep trace of profile- and behavioral-based distinct signals of their users, to provide them personalized and recommended content. Here, we focus on the level of web search personalization, to estimate the risk of trapping the user into so called Filter Bubbles. Our experimentation has been carried out on news, specifically investigating the Google News platform. Our results are in line with existing literature and call for further analyses on which kind of users are the target of specific recommendations by Google.

Keywords: Filter bubbles; web search results; news publishers

1 Introduction

Search engines and social media provide Internet users the opportunity to discuss, get informed, express themselves and interact for a myriads of goals, such as planning events and engaging in commercial transactions. The management of relational networks, e.g., for researches, data gathering and sharing, raises significant questions about the quality of the information retrieved. In his popular work on Filter Bubbles [13], Pariser was one among the first ones to theorize the phenomenon according to which users are unknowingly trapped in "protective" bubbles, created by search engines and social platforms to automatically filter contents. As an example, the author reports how some posts gradually disappeared from his Facebook news feed, probably driven by his historical navigation activities once logged into the popular social platform.

Remarkably, while the most active users of social media act as gatekeepers, the new guardians of information, by personalizing its spread, the same media and search engines track the user navigation and filter the search results. In such a way, gatekeepers and social media become "dangerous intermediaries" [12], with the natural potential consequence of narrowing the world view. Confined

* Partly funded by the Registro.it project *MIB* (My Information Bubble).

in comfortable micro-arenas, the potential risk is loosing the communicative potential of the web, in which information management is performed in a bottom-up fashion [8]. The practice of filtering re-ordering is testified by the service providers themselves. As an example, the Google patent on personalization of web search [10] states the existence of mechanisms linking the re-order of search results to the preferences in the user profiles. In last recent years, Academia spent a significant effort in measuring the level of such personalization, giving raise to seminal work like the one in [6] on Google, which showed that criteria mostly influencing personalization of search results are geo-localization and login into the platform.

Personalization affects also online advertising. Indeed, recently, the traditional advertising approach has moved towards a targeted one: the ad is shown only to online users with a specific profile - location, gender, age, e-shopping history are among the monitored aspects. Although personalised ads have the significant advantage to guide the customer mostly towards products she likes, concerns were born since the ads system could 1) hide to the user other potential interesting products [14]; and 2) expose user private information [1].

In this paper, we focus on news aggregators. The target of our analysis is Google News and the goal is to measure the level of personalization of results returned to different kinds of users when they search over a news dataset. Google News offers a panoramic view on several articles on different subjects, redirecting the users on publishers' accounts, when they can select a news to read. The proposed articles range over a variety of topics like business, technology, entertainment, sports and many others. Moreover, the platform offers the capability to personalize the kind of news shown both to logged and not logged users, based, e.g., on the frequency of news sources, and on specific topics.

We consider two aspects of news personalization, defined by the related literature as *expected* and *unexpected passive personalization* [3]. Expected (resp., unexpected) personalization is an explicit (resp., not claimed) personalization, described (resp., undocumented) by Google in reports and other documentations. The term *passive* means that such personalization has not been directly configured by the user through the appropriate functionalities offered by Google News. To measure personalization, we compare the results provided by Google News to logged users, which we previously trained with different searches and visits to websites, for a set of topics and publishers. For the user training, we adopt as the reference dataset the *Signal Media One-Million News Articles* dataset, which is a public collection of articles, to serve the scientific community for research on news retrieval. In particular, we restrict and focus only on those elements in the dataset with source `espn.com` (Entertainment & Sports Programming Network), with more than 7,000 news present in the dataset. We only pick this publisher since it is linked to Google News and it has a large number of news in English. Moreover, `espn.com` is also part of the Google Display Network (GDN), namely

a large set of websites publishing Google advertisements[5], that is publicly known to make use of user profiling to provide targeted advertisements [7].

To train the users, we extract the *relevant entities*, such as organizations, persons, and locations, mentioned in the titles of the `espn.com` news in the dataset and we use these entities to emulate the behaviors of users interested in *Sports*. We, then, exploit such behaviors to investigate either expected and unexpected passive personalization over Google News.

The intuition behind our methodology is that intensively engaging a user over a specific publisher and topic would lead Google News to infer a specific interest of such user for that publisher and topic. Thus, successive search queries of that user could lead the provider to alter the order of the results, e.g., ranking first the news of the specific publisher, with respect to the order of the news provided to a user without past activity.

While we observe expected personalization in the dedicated Suggested for You (SGY) section over Google News, there are not sensitive differences in the news results shown on the main page, between the trained user and a fresh one. This leads to results in line with related work in the area, such as [3], achieved however through a different experimental setting. The current study contributes to 1) evaluate if news are sorted with or without regard to past behaviors of the user, and 2) define the settings within which users are not exposed to a news results order exchange. The last aspect is particularly relevant since studies in the literature have shown how users are deeply influenced by the results shown by search engines in return to their queries, see, e.g., [16].

The rest of the paper is structured as follows. Next section briefly relates on Google News news personalization. In Section 3, we introduce the reference dataset we start from in our experiments. Section 4 describes the techniques used to extract relevant entities from the *Signal Media* news titles. Section 5 presents settings and implementation of the training and the test phase over Google News and gives experimental results. Finally, Section 6 concludes the paper.

2 Google News personalization

Google has provided many details about the mechanisms it uses for personalization. For example, a first kind of personalization exists for logged users with their web history activated. On the specific personalization over Google News, work in [11] describes an enhanced recommender system to offer in the Suggested for You (SGY) section of Google News news, customized such as to be closely inherent to the users' interests.

However, there exist other forms of personalization, based also on "past news browsing information", as explained in the Google support documentation on news personalization, available at `https://support.google.com/news/answer/3010317?hl=en`. The effects of such personalization could be subtle,

[5] `https://support.google.com/adwords/answer/2404190?hl=en` All URLs have been accessed on May, 15, 2016.

since most users are not aware of its existence, and, thus, quite obviously, they do not know how to disable it, if needed.

The personalization based on past news browsing has been subject of the study in [3], where the authors define two properties: expected and unexpected passive personalization, where passive in both cases means that user has not customized a personalized search through the functionalities of Google News. To measure both kinds of personalization, the user searched over Google News specific topics and publishers. Then, the same user connected to Google News again, and both SGY and the home page are analyzed, trying to find evidence of the user previous activity during the training phase. Results of [3] showed effects for the expected passive personalization (the one supposedly affecting the news in the SGY section) and no effect for the unexpected one (possibly affecting the news shown in the main section of the Google News home page).

In the following, we will show design and implementation of a set of experiments to evaluate passive personalization, with however different settings with respect to what done in [3]. Indeed, we analyze results of queries made by the users, rather than analyzing the news shown by default by Google News once connected to the platform.

In [3], the authors also train both logged and not logged users on a set of specific publishers (USA Today, Reuters, the Wall Street Journal, the Economist). A novelty of our approach is that, by letting users search keywords from real news published by `espn.com` as they appear on the *Signal Media* dataset, we are pretty sure to find indeed news about that publisher as the results of the searches over Google News, not only during the training phase, but also during the test phase. Instead, by focusing on news in the main page without searching for some particular news, work in [3] suffered from the limitation that the chosen publisher could not have been present, given the intrinsic volatility of news.

3 Reference dataset

For our experiments, we refer to the *Signal Media One-Million News Articles* dataset[6]. It consists of a variety of news (from different sources) collected over a period of one month, from September 1st, 2015. Overall, the dataset counts 1 million articles, mainly in English. The sources for these articles include major web sites, such as `espn` considered in this paper, as well as local news sources and blogs. Each article consists of its unique identifier, the title, the textual content, the name of the article source, the publication date, and the kind of the article (either a news or a blog post).

The dataset counts 93k individual unique sources, 265,512 Blog articles, and 734,488 news articles. Moreover, at the time of our study the first five most recurrent sources in the dataset are `MyInforms` (19,228 occurrences), `espn` (from the main website and affiliated ones: (7,713), `Individual.com` (5,983), `4Traders` (4,438), `NewsR.in` (4,039), and `Reuters` (3,898). We have chosen `espn` since it

[6] `http://research.signalmedia.co/newsir16/signal-dataset.html`

is a very large source in the dataset, it has the articles linked to Google News and, finally, because we know in advance that the user activity on this website is tracked by Google. Indeed, `espn` belongs to the Google Display Network (Figure 1).

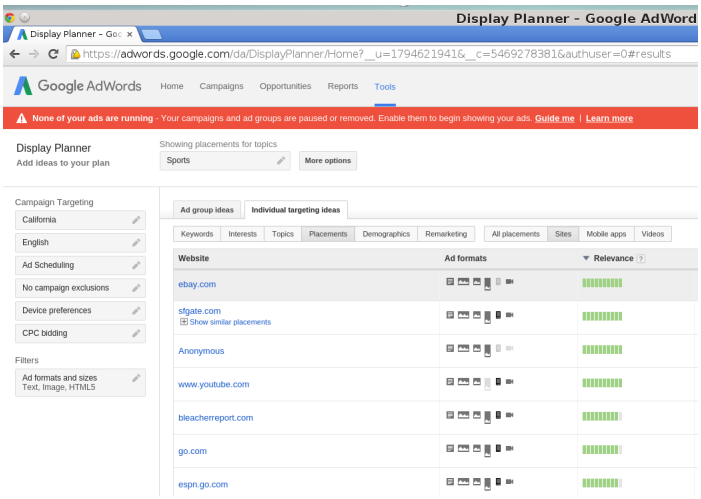

Fig. 1: Google Display Planner: `espn` belongs to the Google Display Network.

4 Named Entity extraction

Named Entity Recognition (NER) is the process of identifying and classifying entities within a text. In the news domain under investigation, common entities are persons, locations and organizations. NER state-of-the-art systems use statistical models (i.e., machine learning) and typically require a large amount of manually annotated training data in combination with a classifier. The best solutions, in terms of classification accuracy, are generally based on conditional random fields (CRF) classifiers [5]. For our experiment, we use the Stanford NER tagger to extract the entities from the news titles in the *Signal Media* dataset. The tagger is indeed implemented through a linear chain CRF sequence labeler classifier and it is part of the Stanford Core NLP[7]. Thus, we exploit the Natural Language Toolkit NLTK[8], which processes natural language through Python programming. NLTK gives the access to the Stanford tagger and to valuable linguistic resources and data models. In detail, as learning model for the classifier, we adopt a ready-to-use model by Stanford, called english.all.3class.distsim.crf.ser.gz, available in NLTK. Figure 2 shows an example of entities extracted from the titles

[7] `http://stanfordnlp.github.io/CoreNLP/`
[8] `http://www.nltk.org/`

of `espn` news, while Figure 3 shows the most recurrent entities under the form of a word cloud.

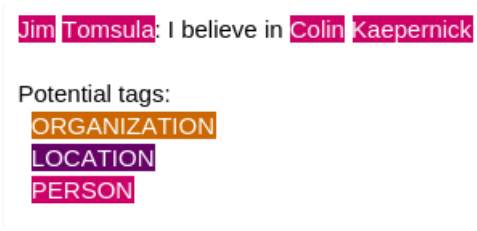

Fig. 2: Stanford Named Entity Tagger: `http://nlp.stanford.edu:8080/ner/`

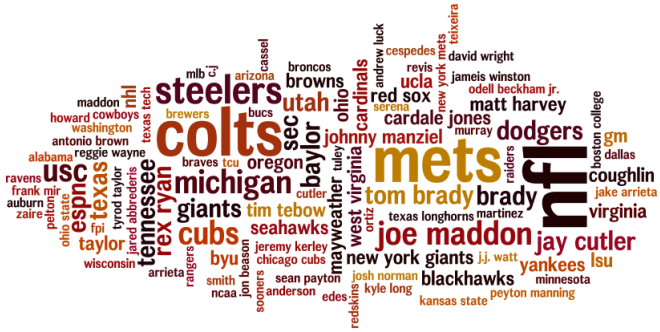

Fig. 3: The most recurrent entities in titles with source `espn`.

On a total of more than 7,000 news titles from `espn` and affiliated sub-publishers, we have considered the ones from the main website (`espn.com`), consisting of 577 titles. We have found 662 entities and 465 unique ones.

5 Experiments

The final goal of the experiments is measuring how users past online behavior affects the news provided by Google News. We investigate two properties, expected (EPP) and unexpected (UPP) passive personalization, as in [3]. The expected version of passive personalization has been disclosed by Google in [11]. Citing Google, "If a user signs in to her Google Account and explicitly enables Web History, the system will record her click history and generate a personalized

section for her, named Suggested for [account], containing stories recommended based on her click history in Google News". Instead, UPP is supposed to have an effect on news shown on the main section of the Google News home page.

Several frameworks have been proposed for analyzing personalized search results and advertisements on Google. The interested reader can refer to [4] for a full survey of measurement tools and their comparison. We have chosen to adopt AdFisher[9] [15], a tool to analyze interactions between online user behaviors, advertisements shown to the user, and advertisements settings. Later, AdFisher has been extended for handling Google searches and news searches and measuring personalization in query results, see, e.g., [3]. Further, a branch version has been created for handling product searches in Google Shopping, to measure price steering [9]. AdFisher has also been used for statistical evaluations, e.g., to measure how users are exposed to Wikipedia results, in return to their Google web searches, see, e.g., [2].

In our work, AdFisher runs browser-based experiments that emulate search queries and basic interactions with the search results, i.e., click only search results satisfying a certain property, e.g., coming from a specific publisher. In particular, AdFisher is automatized with Selenium[10] to efficiently create and manage different users with isolated Firefox client instances, each of them with their own associated cookies, to enable the personalization from the server.

For measuring both EPP and UPP, we emulate two real users, logged into Google, connecting to two separate browser instances, one for each user. Since we are investigating a publisher that is commonly and mostly accessed from US[11], all our experiments are with the users connected from US. We have used the Digital Ocean VPS Service Digital[12] to gain access to machines located in the US.

5.1 Unexpected passive personalization

The following list describes the experiments for training one of the two users and for comparing, in the test phase, the obtained search results with those of the other user.

The user is trained as follows:

- She visits sports-related GDN website pages, including pages from `espn.com`. The websites have been selected with Google Display Planner—a tool providing a series of websites taking part to the GDN and linked to specific topics.
- She issues several queries on Google News, using as keywords the entities extracted from the *Signal Media* dataset (see Section 3).
- She clicks only on the results of the news with source `espn.com` and spends some time on the linked page.

[9] `https://github.com/tadatitam/info-flow-experiments`
[10] `http://www.seleniumhq.org/`
[11] `http://www.alexa.com/siteinfo/espn.com`
[12] `https://www.digitalocean.com`

The training phase lasted about eight hours. To evaluate UPP for both the users, we have performed searches on Google News, with different keywords with respect to the training phase. Both the users searched for 32 test keywords. We recall that the second user does not undergo any training phase.

As highlighted by the Google News guide[13], a form of personalization exists even for users not logged into a Google account. For these users, the "Google News experience will be personalized based on past news browsing information". We are indeed interested in that kind of personalization, based on the previous online behavior.

Table 1 reports the training details.

Table 1: Training behavior

visited pages	searched keywords	read articles	avg time on website	location
17	464	100	50sec	New York,USA

Noticeably, for all the test queries, the fresh user and the trained user have been shown exactly the same results in the main section of Google News home page. Thus, we were unable, under our experiment context, to reveal personalization based on past news browsing information, contrary to what claimed by Google, in the Google News guide mentioned above.

5.2 Expected passive personalization

To analyze the expected passive personalization, we focus on the Google News SGY section. Given a fresh logged-in user, Google News does not provide the user with such a section. Indeed, the user needs to have formerly interacted with Google (either Google search or Google News engine). We follow two approaches to make that section appear, as described in the following.

1. In the first approach, we try to build two user profiles interested in traveling, letting both the users visit 30 travel-related websites, searching for 327 travel-related keywords on Google News, and clicking on the first result. We have chosen the topic "travel" since we consider it a topic quite disjoint from sport. We have emulated such a behavior until the SGY section appeared. This happened after four iterations of searching keywords and visiting websites. This pre-training phase lasted four days. Remarkably, when the SGY section was populated, it was just with one entry, related to a crime news about a murder in Las Vegas, having nothing in common with traveling. The travel-related keywords and websites were obtained with the support of the Google Display Planner.
2. In the second approach, we try to build two user profiles interested in sports. For both the users, we have trained them according to the training behavior

[13] `https://support.google.com/news/answer/3010317?hl=en`

described in Section 5.1. In this case, the SGY section appears earlier, probably because of a larger interaction of the user with the browser. Indeed, the user clicks on all the news from the `espn` publisher, while, for the travel scenario, she was clicking only on the first result, per search. After one day, using the training settings in Table 1, we obtain a SGY section with ten news. A visual examination helps us to assess that such news are related to Sport. Although a text mining approach would be have more effective in assessing topics of such news, we can argue that the read and click behavior thought for users interested in sports has been appropriately designed.

From the results of this preliminary step, we envisage that the number of SGY news depends on the interactions of the user with Google. Indeed, different interactions yield to different results, both for number and content of news in that section.

The travel topic was not a winning choice. In fact, we observed the lack of timely news related to this topic. This let us argue that the implemented behavior does not lead to a profile of a user really interested in travels. Thus, hereafter, we concentrate the experiments on users trained on sports.

In the following, we consider only two logged users, interested in sports, each with a SGY section. We will further investigate if different behaviors of such users yield to distinct results in that section.

- Training: for the first user, we have repeated the training described in Section 5.1. The second user just waits.
- Test: for both users, we have refreshed the Google News home page every 10 minutes to capture real-time events (default reload time of Google News is 15 minutes[14]) and we have focused on the SGY section only.

Both the pre-training session (to let the SGY section appear) and the training phase lasted eight hours. Thus, we have tried to answer the following: *do two users, both interested in sports, but with different interactions with sport websites and sports news, have different SGY news?*

We have computed standard metrics usually adopted to measure web search personalization (Jaccard and Kendall indexes). Given two sets P and Q, Jaccard Index is 1 when the sets are identical and 0 when their intersection is empty while Kendall index (Kendall tau) quantifies the correlation between P and Q. This index ranges from -1 to 1, where 0 means no correlation, 1 means same order and -1 reverse order.

In Figure 4, we can see the differences in the SGY sections of two users, according to the introduced metrics. The sections have from 1 to 3 identical news, out of ten (Jaccard index is from 0 to 0.3). Most of the time, the order of these news is not correlated (for 13 out of 19 the Kendall index is negative). It is worth noting that the untrained user obtains more real-time news (updated within 60 minutes) than the other one. Even if we have no clear evidence to explain this phenomenon, we can suppose that untrained users have a wider

[14] `https://news.google.com/news/settings`

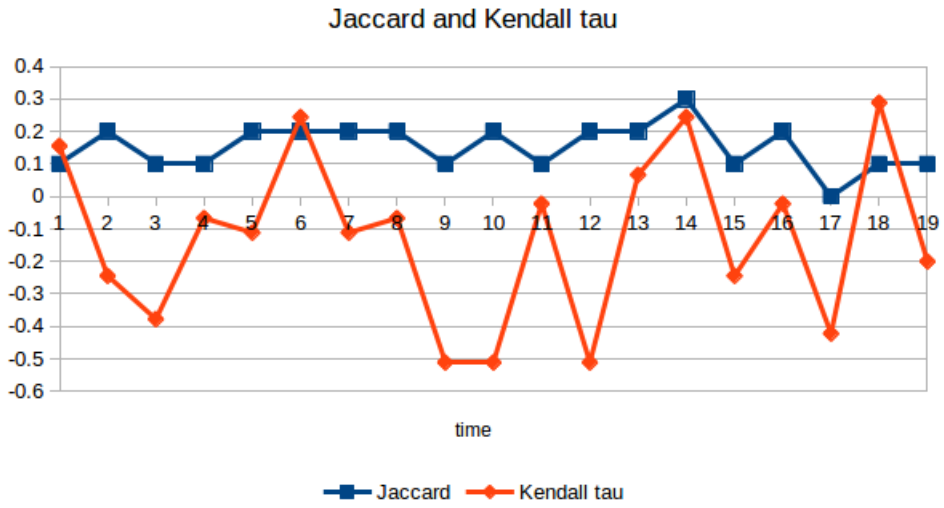

Fig. 4: Jaccard and Kendall indexes of the SGY sections.

interest domain than the trained one, leading to a wider suggestion of news by Google.

As a corollary, we have also considered how many news, among the suggested ones, are from `espn.com`. The results are in Table 2, at each refreshing time of the SGY section. Overall, the trained account has been shown more espn news than the untrained one. Considering the top 10 news shown to each user, we always got 1 to 3 espn news in their SGY section. It is worth however noting that the result is also related to the number of espn news actually published at experimenting time. Indeed, SGY section tends to show first the most recent news.

Table 2: Statistic of news from `espn.com` in SGY section, at each refreshing time

Time slot	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	total
Untrained account	1	1	1	0	1	1	1	1	0	1	0	1	1	3	1	1	1	1	2	**19**
Trained account	2	1	1	1	2	1	2	1	1	2	1	1	1	2	1	1	1	1	2	**25**

Figure 5 highlighs the SGY section of the two users, at testing time. As expected, users have been shown different results.

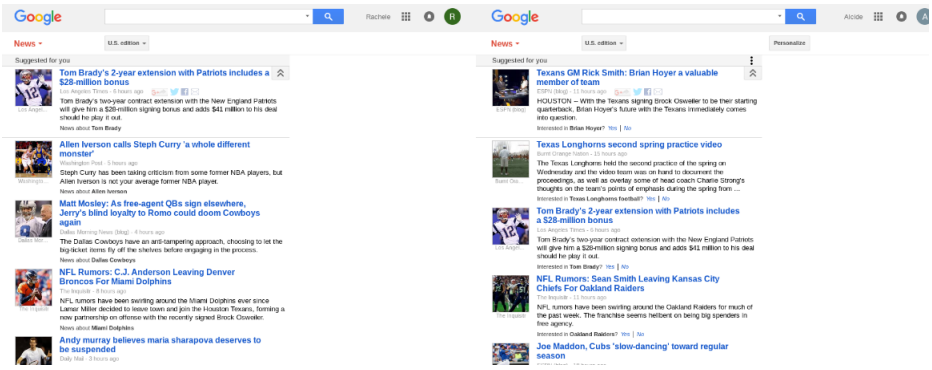

Fig. 5: Snapshot of the two SGY sections: differences and similarities.

6 Conclusions

In this paper, we focused on news personalization on Google News, aiming at measuring the level of personalization (claimed by Google itself), under different contexts: logged users, expected (in SGY sections) and unexpected (in Google News home) personalization. We differ from related work in the literature mainly because we observe the results obtained in return to specific user queries. However, at least for our experiments configuration, we did not observe particular differences in the results obtained by a trained user and a fresh one. Instead, we found interesting results when we searched for expected personalization, looking specifically at the SGY section of Google News. Since the section is not automatically shown to non logged users, nor to freshly logged ones, we carried on experiments to let the section appear on the users pages. Our approach showed that, depending on the kind and number of interactions a user has on the platform, the SGY section differs both in content and number of the shown news. Furthermore, results after training a specific user over a particular topic leads to a different SGY section with respect to SGY of the non trained user (confirming, in this case, previous related results).

References

1. M. Conti, V. Cozza, M. Petrocchi, and A. Spognardi. TRAP: using targeted ads to unveil google personal profiles. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2015.
2. V. Cozza, V. T. Hoang, and M. Petrocchi. Google web searches and Wikipedia results: a measurement study. In *Proc. of 7th Italian Workshop on Information Retrieval (IIR) 2016*. CEUR Workshop Proceedings, 2016.
3. A. Datta, A. Datta, S. Jana, and M. C. Tschantz. Poster: Information flow experiments to study news personalization. In *Computer Security Foundations Symposium (CSF), 2015 IEEE 28th*. IEEE, 2015.
4. S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. [Technical Report], May 2016.

5. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
6. A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring personalization of web search. In *22nd International Conference on World Wide Web*, WWW '13, pages 527–538, 2013.
7. T. Haveliwala, G. Jeh, and S. Kamvar. Targeted advertisements based on user profiles and page profile, Nov. 27 2012. US Patent 8,321,278.
8. M. Hindman. *The Mith of Digital Democracy*. Princeton University Press, Princeton, NJ, USA, 2009.
9. V. T. Hoang, V. Cozza, M. Petrocchi, and R. De Nicola. Online user behavioural modeling with applications to price steering. In *FinRec 2016: 2nd International Workshop on Personalization and Recommender Systems in Financial Services*. CEUR Workshop Proceedings, 2016.
10. S. Lawrence. Personalization of web search, Mar 2005. US Patent App. 10/676,711.
11. J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 31–40, New York, NY, USA, 2010. ACM.
12. E. Morozov. *The Net Delusion: The Dark Side of Internet Freedom*. Perseus Books, Cambridge, MA, USA, 2011.
13. E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011.
14. E. Pariser. *The Filter Bubble: What the Internet is hiding from you*. Penguin UK, 2011.
15. M. Tschantz, A. Datta, A. Datta, and J. Wing. A methodology for information flow experiments. In *Computer Security Foundations Symposium (CSF), 2015 IEEE 28th*, pages 554–568, July 2015.
16. M. Zanker, F. Ricci, D. Jannach, and L. Terveen. Measuring the impact of personalization and recommendation on user behaviour. *International Journal of Human-Computer Studies*, 68(8):469 – 471, 2010. Measuring the Impact of Personalization and Recommendation on User Behaviour.