IMT School for Advanced Studies, Lucca

Lucca, Italy

Measuring Web Search Personalization

PhD Program in Computer Science

XXIX Cycle

By

Van Tien, Hoang

2017

The dissertation of Van Tien, Hoang is approved.

Program Coordinator: Prof. Rocco De Nicola,

Supervisor: Prof. Rocco De Nicola,

Supervisor: Ph.d. Marinella Petrocchi, Ph.d. Vittoria Cozza

The dissertation of Van Tien, Hoang has been reviewed by:

firstreviewer, firstreviewerinst

secondreviewer, secondreviewerinst

IMT School for Advanced Studies, Lucca

2017

To my beloved parents.
Dành tặng riêng tới cha mẹ yêu quý.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my deeply-felt gratitude to my advisor Professor Rocco De Nicola, and co-advisors Marinella Petrocchi and Vittoria Cozza for the continuous support of my PhD study and related researches, for their patience and motivation.

My first thanks go to Prof. Rocco De Nicola. He enlightened the first glance of research. He guided me through all the difficult times of research and writing of this thesis. Without him, I would not be able to finish studying and writing the thesis.

My sincere thanks also go to Marinella Petrocchi. She has led me to become a researcher on every step from the beginning. She is my sample of a researcher, especially for the attitudes towards working. Without her patient support, I would not have been able to start the research from the very first day.

My special thanks go to Vittoria Cozza. She is not only an advisor but also a friend. I would like to thank her for all the inspiration, motivation and confidence she brought to me during the PhD study, particularly in the tough time of research. Without her valuable support, it would not have been possible to conduct multiple studies.

I am happy to acknowledge my debt to professors Francesco Tiezzi and Angelo Spognardi for the works we did together, especially for their helpful advising in the initial parts of my research.

I thank the staffs in the School and in the residence building. There are a lot of people that I even do not know in person,

but without their help throughout the years, it would have been impossible for me to do research.

I thank my fellow colleagues for the constructive discussions, the knowledge from other fields and for all the happiness we had together in the last three years.

Last but not the least, I would like to thank my parents, my brothers and sisters and my significant friend for supporting me spiritually throughout the writing of the thesis and my life in general.

Declaration

Most of the material in this thesis has been published. In particular, Chapter 3, 4, 5 and parts of Chapter 6 are based on (CHPD16; HST+15; CHP16) and (CHPS16), co-authored with Vittoria Cozza, Marinella Petrocchi, Angelo Spognardi, Francesco Tiezzi and Rocco De Nicola. Finally, Section 6.2 of Chapter 6 has been developed in collaboration with Prof. Rocco De Nicola, Marinella Petrocchi and Vittoria Cozza.

Vita

2008 Bachelor of Science
University of Science,
Ho Chi Minh, Vietnam

2012 Master of Science
University of Trento,
Trento, Italy

Publications

1. Van Tien Hoang, Angelo Spognardi, Francesco Tiezzi, Marinella Petrocchi, and Rocco De Nicola. Domain-specific queries and web search personalization: some investigations. In*Proceedings 11th International Workshop on Automated Specification and Verification of Web Systems, WWV 2015, Oslo, Norway*, pages 51–58, 2015.
2. Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi, Angelo Spognardi. Experimental measures of news personalization in Google News. In *SoWeMine 2016: 2nd International Workshop on Mining the Social Web*, 2016.
3. Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi. Google Web Searches and Wikipedia Results: a measurement study. In *IIR 2015 - 6th Italian Information Retrieval Workshop*, 2016.
4. Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi, Rocco De Nicola. Online User Behavioural Modeling with Applications to Price Steering. In *Proceedings of FinRec 2016: 2nd International Workshop on Personalization and Recommender Systems in Financial Services*, Volume 1606, page 16–21, June 2016.
5. Van Tien Hoang, Vittoria Cozza Marinella Petrocchi, Rocco De Nicola. A geo-economics study on price steering in Google Shopping. To be submitted.

Abstract

Personalization in online services is the practice of tailoring data contents for customers according to what is supposed to be their expectation. By design, personalization has been considered an important tool to help users to find the most interesting relevant data. By doing that, personalization also filters the web contents and potentially narrows the view of users. In this context, our work aims to measure personalization levels of web search results.

First, we study personalization for web search engines with two case studies about Google Search. The results show a remarkable level of Google personalized search results based on sets of keywords, and we show that a specific website appears as prevalent in the results of web searches.

Second, we measure the personalization degrees of an online news aggregator to provide a wider view of the problems. It also shows that compelling evidence such as “suggested for you" heavily depends on past users’ activities.

Finally, we study personalization of search results on an online shopping platform by measuring price steering phenomenon. Particularly, we investigate the impacts of online behaviours, locations and economic performance factors, and we observe that price steering is based on user’s behaviours but it is also influenced by the geographic location of users.

Chapter 1

Introduction

The World Wide Web is a huge repository of web pages and the amount of information it contains is increasing at high speed. Cisco Visual Networking Index estimates that global IP traffic will grow nearly three folds in the next four years (2020) and it would cost more than 5 million years to watch the amount of video that will cross networks each month at that time (CIS15).

As the Internet grows, the amount of information and products available on website increases exponentially. Recently, Google has indexed over 30 trillion individual web pages (Ven13), Amazon has more than 400 million products selling (Gre15) in 2015. Each second, there are more than 700 images uploaded to Instagram and more than 7000 tweets are posted online each second (Int16). The more content is published, the harder it is for users to find the data they are looking for. To deal with this issue, most of the available services build a query-based interface for finding information. Users enter keywords and retrieve a list of results. Due to the vast amount of data, the results could be very huge and could overwhelm users' abilities to process the results. Looking for needed products within one million similar items is nearly impossible. To help users find the most relevant subset of results, online services employ personalization approach. Personalization in on-line services is the practice of tailoring data contents for customers according to what is

supposed to be their expectations. This is possible because on-line applications exploit data about users and are able to draw user profiles, along with their interests and other attributes (e.g., geographical location). The data about a user are automatically collected and generally consist in:

- **Explicit data** that are inserted by the user at registration time such as name, gender, location, and age.

- **Implicit data** that are inferred from the IP address from which the user is surfing; from user on-line behaviour (i.e., frequently visited sites, time spent on a page, most clicked links).

Personalization has been considered to help users to find the most interesting and relevant data. Many on-line services like social networks, search engines, and shopping sites apply this technique. For examples, Google and Bing use geographical redirection which leads users to the locale version of the websites; Facebook recently selects the trending news stories for their users (Fac16a) and prioritizes stories on news feed according to whom they are interacted with and how other people react to the stories (Fac16b). Another service, named Foursquare uses location to suggest nearby places such as restaurants, while Twitter shows tweet trends in the area. Those practices require a lot of information from users, even about their interaction with other websites. For example, Facebook can identify its users and track them from its Facebook Social Plugins, which is embedded on third party sites. Google also unifies users' data from all of their products to a single entity to provide personalization contents better. In this thesis, I consider personalization in the context of results shown to the user, in reply of the users' searches, on web search engines. The study focalizes on search engines because, amongst the many different types of online services, search engines are well-known for using personalization. Nowadays they are the essential part of online users; the Google brand name has even become a verb in dictionary ("to google") to indicate the action of looking for information on the Internet. Major web search engines such as Google, Yahoo, and Yandex are not only used for searching information, but also for data filtering as their

algorithms decide which websites are included in results and shown to the user first.

Because of that, web search engine providers not only have the power to filter which contents are shown on their services, but also have significant influences on how the data of a website should be organized and expressed. They do that by giving priority rankings to websites that follow their web standards. For example, an online vendor that wants its website to be available on 1st result page of Google should invest resources on customizing their pages according to Google guideline (Goo16).

The work is grounded on existing scientific results, which proved how different online service providers apply personalization practices to offer their users/customers personalized services.

While Hannak *et al.* (HSMK+13) showed the existence of personalized search results from Google, Mayer *et al.* (XMD+14) found a convinced evidence that search results from Google are heavily personalized with respect to their geographical location.

Wills *et al.* (WT12) found personalization on Google Ads which are based on user traits (i.e., web activities). Similarly, the authors of (LDL+14) found evidence of targeted advertising on Gmail (a product of Google) and personalized recommendation on Amazon.

On the other hand, Kliman-Silver *et al.* (KSHL+15) also studied the impact of location on personalization, finding that the level of personalization depend relatively on the geographical distance. Both works of Lecuyer *et al.* (LSS+15) and (DTD15) study personalization on Google Ads. Google News personalization is also confirmed in the study of (DDJT15). From an insider perspective, work of (GUH16) on Netflix revealed their own recommendation algorithm to provide personalized suggestions to users.

Thus, grounded on different shades of personalizations studies already carried in the literature, the goal of this thesis is 1) to investigate whether the phenomenon of personalization is quantifiable, or not, concentrating exclusively on specific platforms, and 2) to perform such tasks relying on settings (of, e.g., user profiles) which are different from the ones that have been already investigated in past literature. For exam-

ple, we concentrate only on the search history while the other works like (HSMK+13) analyzes several features such as user agents, genders; for price personalization evidence, we concentrate on a different perspective: economic indicators; for online news, we improve the keyword selection procedure with the help of Name Entity extraction.

The specific platforms considered in the thesis are: Google Search Engine, Google News, and Google Shopping.

The set of experiments presented in the thesis aimed at quantifying possible personalization effects in the following different contexts, as illustrated in the following. Furthermore, the purposes of some of the executed experiments are to differentiate the results retrieved on the personalized version of Google with the ones retrieved on the non- personalized version of Google.

1) First, we analyse if personalization is quantifiable looking at search results in return to queries to Google search engines by different kind of users who are interested in a specific domain (e.g., sport); 2) Secondly, in order to measure personalization of news content, we consider Google News; 3) Third, we consider Wikipedia, to study the ranking of Wikipedia pages as shown in return of queries on Google; 4) Finally, we consider an e-commerce scenario and we investigate the practice of price steering in Google Shopping. To do that, we consider user activities on the web as well as the economic performance of the users' location, in terms of their Gross Domestic Products (which is inferred from the IP address).

Methodologically speaking, we developed a methodology to create synthetic user profiles, based on specific characteristics of the user (e.g., affluent users, sport- interested users).

In addition, we also increase the number of certain experiment parameters, such as the number of keywords to search for (Section 4.2). In order to fully automatise the experiments on a large scale, we use synthetic user profiles for all the experiments, instead of semi-synthetic for human-based ones.

Also, we do not use a trim-downed version of web browsers and web extensions (as done, i.e., in (i.e.,(HSMK+13; LSS+15; XMD+14)) but we

use the full feature browser (i.e., Firefox) to get a better control over the experiments and web compatibilities.

1.1 Pros and Cons of Web Personalization

In his popular work on Filter Bubbles (Par11b), Pariser was one of the first 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, despite the amount of online data is vast and most of those are available to public usage, Internet users tend to utilize social media and search engines to access needed information. Consequently, the social media and search engines track user navigation and filter the search results.

In such a way, search engines and social media become "dangerous intermediaries" (Mor11), with the potential consequence of narrowing the world view. Confined in comfortable micro-arenas, the potential risk is that of losing the communicative potential of the web, in which information management is performed in a bottom-up fashion (Hin09). The practice of filtering and re-ordering is testified by the service providers themselves. As an example, the Google patent on personalization of web search (Law05) mentions the existence of mechanisms linking the ordering of search results to the preferences in user profiles.

In recent years, researchers have spent a significant effort in measuring the level of such personalization, giving raise to seminal work like the one in (HSMK+13) on Google. This work showed that criteria such as geo-location and states of users (logged-in or not logged-in) influence personalization of search results.

Popular e-commerce websites, such as the Amazon Marketplace uses personalization in advertisements. Thanks to the fact that they offer a window to thousands of merchants, they are able to provide advertise-

ments and services to millions of potential buyers. In addition, by using more personal information about customers, vendors can take advantages of dynamic pricing; see also (D'O15). It is thus important to investigate the pricing practices of e-commerce websites. The revenue of online-shopping is 126B$ in 2013 (Sta16) only in the USA and online sale of the UK, Germany, France, The Netherlands, Sweden, Italy, Poland and Spain accounts for 156.28B Euro in 2014 (fRR16). There are 12-24M websites for selling online (2013), according to an estimation (Ref13). They could be private websites or platforms for merchant such as Amazon Market Place and eBay. These practices lead to the booming of online advertising with big players like Google Ads (33.3% revenue of worldwide market share in 2015), Facebook Ads and Bing Ads. The more related ads are shown to users, the more potential buying action. This emerges for online target advertising or personalized advertising phenomena.

In this advertising approach, the specific ad is shown only to online users with a specific profile; location, gender, age, e-shopping history are among the monitored features. In this way, the merchant pays only for ads shown to users matching the ideal buyer of its products.

Although personalised ads have the significant advantage of guiding the customer towards products she likes by matching her profile with appropriate merchants, concerns arise because the ads system could hide to the user other potential interesting products (Par11b) or practice price differentiation on user (MGEL12).

Price steering (online) is one of the practices of price differentiation. It refers to the application of changing the order of search results to highlight specific products prices. So far, there are several papers about this. For example, the study of (MGEL12) shows that personal online traits and location contribute to price steering by doing simulations. For real world data, the authors of (MGEL13) use crowd-sourcing information from Amazon Mechanic Turk users. The data then shows that location has an impact on prices offered to users. A similar result is due to (Ha14), in this works the authors extensively measure both price steering and discrimination. With both real data collected through Amazon Mechanical Turks and synthetic data from controlled experiments using a specific

web browser (Hid16), the authors analyse the prices offered by a plethora of online vendors. They find evidence of price differences by different merchants and retailers: their websites record the history of clicked products to discriminate prices among customers.

1.2 Problems and Possible Solutions

Personalization consists of multiple aspects such as data mining algorithms which are used to analyze users' data and how to present user profile data on the system to analyze later. In this thesis, I measure the level of personalized contents by differentiating the results retrieved on the personalized version with the ones retrieved on the non-personalized version.

In that point of view, personalization has multiple advantages, but it raises several issues which we can address the following research questions that we will answer in the rest of this thesis. Those questions are:

1. **How to detect personalization**: Consider a website which provides contents to a user, we want to know whether such contents are tailored for that user, the kinds of tailoring used and whether they are in control of user explicitly or applied without user's intervention.

2. **Quantify personalization level**: In case of personalization, we want to determine be realized the level in numeric values or similar metrics. In this way, we can evaluate the degrees of personalization of systems.

Obviously, there exist studies that try to answer those questions. For example, on target advertising, Wills *et al.* (WT12) measured personalization about online display ads on websites while (LDL$^+$14) focused on ads inside Gmail and recommendation of Amazon. Another similar study (DTD15) reported that there are connections between user's gender and ad contents on Google. Also, personalization of results by web search engines has been investigated by many authors. For instance, Hannak *et al.* (HSMK$^+$13) proposed methods to detect and measure personalization of Google search engine. Mayer *et al.* (XMD$^+$14) studied

the personalization of search results using crowd-sourcing methods via web browser plug-ins. Kliman-Silver *et al.* (KSHL$^+$15) confirmed a remarkable level of location-based personalized results on Google results. All these results are consistent with the work reported in (FDA14), that shows the geographical parameters of URL has more impact than other elements. Instead of analysing what factors make personalization, Majumder *et al.* (MS13) built a probability model to compute personalization vector from search results. Other studies, like (E$^+$14; TDDW15; EEZ$^+$15) proposed methodologies to carry out experiments on personalization information.

We studied problems by conducting multiple experiments. In order to avoid possible noises due to with compatibility issues (i.e., Javascript execution, web rendering) experimented in previous works (KSHL$^+$15), we utilize a real web browser (Firefox). Besides that, our experiments are able to conduct at long time each session. In one session, synthesized users can mimic real user activities such as selecting links to click, moving mouse, scrolling page, and staying at websites for a fixed number of seconds. Other user behaviours like visited links, clicked links, and query terms can also be modelled for analysis. In order to simulate different locations, we utilize dedicated servers at those places and tunnel our data or redirect commands to them. In our experiments, we are inspired by the methodologies proposed in previous works like (DTD15; TDDW15; E$^+$14; LSS$^+$15; EEZ$^+$15), we studied and adjusted them in our works. We utilize the principles of experiment design and adopt them for our studies. We design the experiments in such a way that measurable features are variables while other ones are kept unchanged; or we select statistical testing based on nature of the experiments. Furthermore, in the case of price personalization (price steering), instead of focusing on specific vendors like in previous work (Ha14; GCF10), we investigate the merchandising platform - Google Shopping. This platform gains access to a wide variety of user's data from many services such as YouTube, Gmail, Google Search and websites in Google Display Network.

1.2.1 Our Contributions

We improve upon previous studies by creating a full automatic experiment that could be run at large scale. For small scales, the user profiles (Google accounts) could be also created manually to store search history and other related data. In the larger experiment, this would increase the difficulties and complexities. Therefore, we magnify the number of user profiles by using ***browser profiles*** storing data and reuse them for different experiment sessions. We also advance the abilities to recover data after web browser crashing. Our studies focus on Google and its services (e.g., Google Shopping, Google News) because of its popularity.

Google is selected as the object to be studied for some of its products, like Google Web Search, Google News, and Google Shopping. For each one, all questions which are mentioned at beginning of this part (1.2) are inspected.

Assuming that no technical details of any systems are available, we consider a service like a black box and perform analysis through experiments. By doing so, we want to determine whether there is some personalization or not. Moreover, we study the data which contribute to the phenomena and measure their impact on personalization. Furthermore, we conduct experiments and analyse them from many viewpoints for better understanding. Thus, we provide the following contributions:

1. **Google domain specific queries**: Experiments have been carried out, consisting of search queries performed by a battery of Google accounts with differently prepared profiles. By matching query results, the level of personalization is quantified, according to topics of the queries and the profile of the accounts. This study is based on a series of experiments that are useful for understanding how the characteristics of a set of accounts (mainly, previous search activities from those accounts) may influence further searches. The obtained results do not show an exceptional level of Google personalization, at least regarding the kind of queries we have chosen to perform.

2. **Ranking of websites on Google search results**: This study mea-

sured the position of a website in search results to evaluate their popularity on search results of the search engine. The outcome provides the existing evidence of such website. For instance, most of the times, there is one Wikipedia page within the first five results of Google search. This study aims at contributing to demonstrate how certain websites (in the case of our study, the Wikipedia pages) tend to appear in the very first positions of the search results, thus actuating a sort of monopoly of the information shown to the users.

3. **Personalization of Google News**: For this study, we focused on news personalization on Google News, aiming at measuring the level of personalization (claimed by Google itself), under different contexts. We found that depending on the kind and number of interactions a user has on the platform, the personalized section differs both in content and quantity. Furthermore, we found that the training of a specific user over a particular topic leads to a different Google News page sections with respect to the personalization of an anonymous user.

4. **Price steering in Google Shopping**: We studied how user behaviours and geographic locations affect the display of search results, with goals to figure out price steering on Google Shopping. In order to form behavioural traits, we train user profiles by letting them perform sets of designated activities. Results of our experiments show that trained profiles are not always influenced by previous search and click activities. However, profiles with different training types yield distinguishable search results. We then investigated the impact of geographic locations. In this case, in general, the analysis results evidence of distinguishable price of products shown to the users in different places both amongst countries and amongst cities in one country.

From these results above, we could confirm the presence of personalization on web results, both for general web search and for commerce web search. Among many features, user behaviour and geographical location

are major criteria which are taken into account by on-line web services for tailoring contents.

1.2.2 Thesis Layout

The thesis is organized in 6 chapters. Chapter 1 defines open problems and anticipate the proposal solutions. Chapter 2 is dedicated to presenting related works. Chapter 3 discusses methodology and techniques to approach the open problems. Chapter 4, Chapter 5 and Chapter 6 describe our works on web search personalization, online personalization and price steering in search results, respectively. Chapter 7 summarizes all problems, provides considerations and conclude the works.

Chapter 2

State of the Art

In the previous chapter, we have described an overview of web personalization, its issues and possible solutions. In this chapter, we discuss related studies of personalization. First, Section 2.1 discusses how user profiles are constructed and represented. Next, Section 2.2 introduces the systems in which such user profiles are utilized. Section 2.3 presents how personalization is implemented in different systems. Then Section 2.4 goes into works of measuring personalization in specific aspects (e.g., in web search engines, in web ads). It also considers methodologies and techniques used to quantify personalization levels. The last part, Section 2.5 considers personalization of search results of e-commerce websites to evaluate personalization pricing (price steering and price discrimination).

2.1 User profiling

In this part, we discuss user profile specifically for web search personalization. A user profile presents a collection of information associated to a specific user. In other words, a user profile is a digital presentation of a user and addresses characters and descriptions of that user.

In the literature, there are two fundamental methods to retrieve data about a user profile. They are called explicit and implicit (also see Chap-

ter 1) information collecting. For the former, data are directly input by users (i.e., personal data provided at registration time). The drawbacks of this method are that a user has a stationary profile and it is valid until that user alters the previous input information. For the latter, profiling uses the undirected ways to gather user's information and analyze the user behavior (i.e., past web activities) to determine user's interests (PCG03). In practice, there are several techniques to collect explicit user information such as proxy server (PG99), desktop agent[1], search logs (LYM02), etc. One of the common collected data type is browsing history. Also, it is possible to construct a hybrid user profile using both ways.

In other to utilize user profiles, they must be presented in a formal way. These representations could be keyword profiles, semantic network profiles, and concept profiles (ontology-based).

Keyword profiles address the user profiles as sets of keywords. Each keyword can represent a topic of interests or a group of categories which reflect user's interests. Some systems utilize this kind of profiles are (Mou96; MGH98).

Semantic network profiles present a profile as a weighted semantic network in which each node is a concept. Two example works are (KL94; SHM+05).

Concept profiles are similar to semantic-base network profiles in the presentation. However, each node is an abstract concept rather than specific words for sets of words. Works of (MSDR04; KAS05) apply this methods by building concept profiles based on ontology (ontology profiles).

2.2 Personalization algorithms

Fundamentally, there are three types of personalization algorithms. It is also possible to apply a combination of the three approaches. The methodologies are:

[1]Google Desktop - A retired product from Google - `https://en.wikipedia.org/wiki/Google_Desktop`

- Rule-Based Personalization: This is the most basic personalization algorithm. This kind of system relies on decision rules to provide personalized content to a user. The rules could be obtained by asking users questions. One popular application of this system is to provide local content according to geographical location of a user (i.e., Google.com will be redirected and/or change the displayed language according to the IP address of users).

- Content-Based Filtering: This profiling method is rooted on the analysis of previous rated items of a user. It assumes that the users who belong to the same groups would behave similarly and, therefore, they would have similar profiles (KS00; GA05b). IIn terms of web search, rates could be associated with ranks of clicked links or time spent on the related pages. Based on that analysis, a user profile is generated, then it is used to predict which items are within the user?s potential interests. For example, in (SKK00), the authors used Naive Bayes and nearest neighbor approach to suggest potential related items to users. In (VdODS13), the authors proposed an algorithm based on convolutional networks to recommend music to a user.

  The following texts will provide the details about those methods

  - Vector-Space Model: it is a statistical-termed based technique (SD97) widely used in information retrieval. In this model, the user profiles are presented as vector(s) of weighted keywords, associated with the users? interests (KS00). Hence, the weight points to the importance of the keywords (GA05b). The user interests could be addressed by a single vector or by multiple vectors, each of them indicating a specific interest (GA05a). The quality of the model is based on the generalization degree of the vector(s).
  - Latent Semantic Indexing: it is also a statistical term-based approach. It aims at solving the orthogonal problem of the Vector-Space Model, by studying the latent structure of a document and the terms inside. This method could learn the re-

lationships between terms and concepts (SD97). Even though there are no overlapping terms between user profiles and contents, it still suggests the relevant contents to users (SD97).

  - Learning Information Agents: it uses the combination of artificial intelligence and user feedback to update the user profile (SD97). The updating algorithm uses the selected contents and the associated user evaluation (such as a user rating) to update weight of user interests.
  - Neural Network Agents: they are similar to Learning Information Agents, but the user does not need to score the documents as the score is automatically computed by the system when the user accepts or rejects a content (GA05b).

- Collaborative Filtering: This method collects information of many user profiles, then it produces the personalized content for a user based on the similarities between that user with the other ones. In this way, a model is produced from user behavior?s data to predict future actions. Also, there are a number of data mining algorithms that could be used to build such models, including Clustering, Classification, or Markov Models. An example is in (SKKR01), where the authors proposed an item-based collaborative filtering algorithm to recommend personalized contents (recommendation); in (UF98), the k-mean algorithm is used for item and user-based clustering.

  Following, we briefly remind two well-known collaborative filtering techniques: memory-based and model-based techniques.

  - Memory-based and model-based techniques: these techniques let users filtering the contents by the rating feedback of similar minded users in the system (SD97). Hence, they could suggest content that do not previously appear within the user interest sets. These methods are highly effective when the number of ratings is relatively large. The memory-based technique predicts the rating of a particular user by analyzing all the previous ratings given by similar users (LNSU08; AT05; SK09). The

model-based technique anticipates the user?s rating by using a model which is learnt from the collection of ratings (AT05; SK09).

Finally, there are also hybrid approaches, which are the combination of the content-based and collaborative methods, to provide better personalized contents (GA05b; KP06).

2.3 Personalization Techniques Overview

Due to the dominance role of search engines (there are four search engines - Google.com, Baidu.com, Yahoo.com, Yahoo.co.in - in Alexa global top ten (Ale16)), web search personalization has attracted many works in recent years.

One approach that the online service providers use to obtain personalization is studying user's profiles. For example, all the relevant data, such as the user's gender, location and click history are exploited. By doing this way, it is required to store a large amount of data, which may also consist of information not available to everyone, i.e., server logs, personal data of users, and so on. Another way is to utilizes data related to the behavior of the user over time, over both short-term and long-term periods as well as in very short-term (browsing session).

The works of Nanda *et al.* (NOD14), Matthijs and Radlinski (MR11), Yury *et al.* (US13), Ryen *et al.* (WBD10), and Makvana *et al.* (MSS14) are examples of the approach, and demonstrate some possible techniques. Also, there are similar works considered by Hu and Chan (HC08) and (YL10). In addition, the authors of Yue *et al.* (YHH14) and Mikhail and Matthew (BR11) combine multiple techniques for studying personalization.

Nanda *et al.* (NOD14) created an ontology-based profile for users by building a hierarchy of topic trees from a list of websites (e.g., from Open Directory Project [2] and Wikipedia). They then combined it with explicit user interests (users provided bookmark links and some keywords that

[2]http://www.dmoz.org

they associated with a topic). Profiles are then created through collaborative filtering. The authors proposed a technique to re-rank the results from search engines according to their relevance to a user, based on her learned profile. Matthijs and Radlinski (MR11) also used long-term search history to develop models of user's interests and used those models to re-rank Web results.

Yury *et al.* (US13) considered short-term context by exploiting browsing history and first queries of search sessions. A search session is a series of similar or same search goal queries issued to a search engine. The authors predicted which Web result links would be clicked by modelling features from search session context like queries, click-through and browsing. The links were then categorized by a hierarchical ontology structure, based on Open Directory Project. Finally, a re-ranking function based on previous prediction models is applied to create personalized web results.

Ryen *et al.* (WBD10) also studied short-term context, current session and query. They proposed to combine and weight the context of each query to predict short-term interests of users.

Makvana *et al.* (MSS14), as opposed to client-side history, analyzed Web logs from servers. They identified related search terms for a particular user from previous searches history, and used these related terms to clarify her search intent for ambiguous queries. To do that, the queries were expanded by adding other related terms to them. For example, the ambiguous query "apple" was transformed into either "apple fruit" or "apple ipod" depending on user's search history. Moreover, they processed user's search query results and used Vector Space Model (VSM) to generate user interest values on the links, and produced new ranks of links.

Hu and Chan (HC08) proposed a scoring function that uses features from terms and images (descriptions) to score a term that matches users profile, which is learned from users bookmarks. Authors in (YL10) analyzed short-term query context, and user context like clicks and links for future use in personalized search sessions.

Other works combine methods. As an example, Yue *et al.* (YHH14)

used short-term and session-based history, collected in three months, to generate a probabilistic model and a click-through model to customize search results for users. Mikhail and Matthew (BR11) used a similar approach to determine probability measures that customize what should be shown to users.

2.4 Measuring Personalization

In many online commercial services, it is nearly impossible or very hard to know what information is used for personalization algorithms. One of the main reason for these difficulties is the fact that data are parts of company business and are consider confidential. In spite of this, there have been many works that tried to determine the information utilized to obtain personalization.

Wills *et al.* (WT12) studied personalization on ads by investigating what web advertisers actually do with the users' information available to them. In order to provide data for the advertisers, users visit a list of websites which are connected to ads networks on daily sessions for a period of ten days. The other users (controlled users) do the same but must remove all cookies and histories at beginning of the sessions. On Google Ads network, the study found that many ads are induced by geographic location, and by behaviour such as input search terms and visited sites. In addition, it was shown that user's browsing behaviour on non-Facebook sites does not influence the ads shown on Facebook. The influences are observed only when the Facebook like button is used to express interest in the content. One of the drawbacks of this study is that the experiments were performed manually; users visited only a small number of sites (20 in case of Google Ads), and each session lasted between 20 and 30 minutes. Currently, this work also cannot be re-produced as it is because `www.google.com` has changed the "*Ad Preference*" settings.

Another recent related study is that of Hannak *et al.* (HSMK$^+$13), in which the authors propose methods to detect and measure personalization of Google search engine. They evaluated the effects on query results

of different elements of a user profile (i.e., age, gender, income, location), of system related data (i.e., browser, operating system) and server-side technology (i.e., cookies). They found that Google provides a low amount of personalization of results in general (12% of results are personalized), and it is based mainly on being logged with Google account and making requests from different geo-localization. The study is, however, limited to queries within the US to Google US version and uses an outdated browser engine (Hid16) for crawling data, which could affects the results; for example, version 12.0 of Opera web browser renders `www.google.com` differently from version 45.0 of Opera.

In contrast, in previous work which developed methods for synthesized experiments, Mayer *et al.* (XMD+14) studied the personalization of search results by comparing results of real Internet users with results re-produced from their servers. To do that, the authors created a web browser extension, called Bobble, and asked users to install it. They successfully collected more than 75,000 real search queries issued by hundreds of Bobble users over nine months. After analysis, they found that the geographic location counts more than user search history and 98% of search queries show some level of search as differences due to geolocation. Although this study is deployed in large scale with real users, other features such search terms and web activities histories are not considered.

For purpose similar to those of the above-mentioned works, the authors of (LDL+14) came up with XRay - a browser plugins and a system to detect personalization according to input data (i.e., email content). They found evidence of targeted ads in Gmail and personalized recommendation on Amazon.

To better understand the impact of geographic location on personalization web search, Kliman-Silver *et al.* (KSHL+15) studied location-based personalized results on Google search results. By analysing user's searching history from different GPS ordinates in the US, they figured out that level of personalization grows as physical distance increases. Moreover, the nature of search terms (types of queries and query itself) also plays an important role as it triggers different levels of per-

sonalization. However, this work focuses only on the mobile version Google web search and emulated Safari web browser by a different web browser (Hid16); the GPS locations are simulated by using *JavaScript Geolocation API* (Pop14). This could create to issues because that browser can be detected (She16) and GPS locations can be inferred from many relevant sources such as GPS, Cell-ID, and Wi-Fi network (Dev16).

Similarly, the authors of (FDA14) worked on privacy impact over Google search's results. In order to do that, the authors analysed details in parameters of URLs that are sent to Google each time. They found that parameter about location has more influences than any other kind of parameters. Indeed, when the parameter that switches personalization on is turned off, there are still personalization results based on geo-location.

Instead of investigating the factors which influence personalization, in the work from Majumder *et al.* (MS13), the authors built a probabilistic model (Latent Topic Personalization) whose inputs are data with personal information of users and non-personalized data. The model then computes a *personalization* vector. Albeit the authors found evidence of profile-based personalization, the data set is very limited (10 real Google users).

Lecuyer *et al.* (LSS$^+$15) proposed a framework to detect the cause of ad targeting on the web, by focusing on personalized advertisements in Gmail. They placed more than 300 emails containing keywords or phrases to encode a variety of topics and discovered that information about a user's health, race, religious affiliation or religious interest, sexual orientation, or difficult financial situation, all are used to generate targeted advertisements. This work has been considered a step forward to understanding personalization with private data.

Similar to (LSS$^+$15), Englehardt *et al.* (E$^+$14) also proposed methodologies to perform experiments and measure personalization level.

The methodology discussed from targets (websites, cookies,etc.) to variables (non-user-specific, users' demographics, history, etc.) and how to quantify personalization by machine learning and statistic tests. The authors then considered a case study of "filter bubbles"(Par11a) on news personalizing over news sites based on user's history and found the effect

was minimal (less than 11%).

To detect personalization of contents, (TDDW15) introduces a methodology to design, set out experiments and analyse results (hypothesis and testing). An application of (TDDW15) is the work of (DTD15) in which automated measuring personalizing of Google's Ads. In this study, authors use programmed web-browser via Selenium framework (Pro16) to control navigation, they detected that Google ads on high salary jobs are biased toward men instead of being equally distributed over men and women. In addition, the methods and software from this work can be applied for other personalization measurements.

2.5 Measuring Personalization Search Results of E-commerce Websites

In the previous part, we have mainly focused on related works on personalization of web content in general. This section continues relating on literature in the area of personalization of web results, a further step after online target advertising. It focuses on price differentiation which includes price steering and price discrimination. While price steering denotes the practice of showing different products with different prices to different users, discrimination is a similar practice, but related to the same product.

Authors of (MGEL12) tried to detect price discrimination on the Internet by conducting experiments in three categories: system-based, location-based and personal-info-based measurements. The results provide no evidence of price discrimination for system based features such as OSes or browsers, and this is also consistent with results from (HSMK$^+$13). It is found that price differences are mainly based on the geographical location of the customer, but depend on the types of products, highest levels for e-books and video games. One another feature - the origin of request (the *referred* part of HTTP header) from users also affects price discrimination (for some sites, prices are considerably lower if visited via a discount aggregator site). This study provides an overview of price and search discrimination and mainly specifies on certain online vendors

only.

However, the results of the previous study are in contrast with what is reported by *Wall Street Journal*. It is written that Orbitz - an online travel agency, tracked its users and figured out that Mac users were more interested in costly hotels and travelling services than Windows users. Therefore, the most costly results as the first results for Mac users (Mat12).

In order to better understand price discrimination at a larger scale and on real world data, Mikians et. al (MGEL13) studied online price discrimination by collecting data from 340 Internet users from 18 countries. The analysis focused on how the price of the same products, which are offered from a set of retailers, varies from retailer to retailer. The authors developed a web browser extension for these experiments. The outcome shows that geographic location of users is one of the main factors affecting the prices even if its constant influence over all the experiments was not assessed. Although geographical location has a considerable impact on displayed prices, it is not constantly influenced through all of experiments. However, the authors did not consider user behaviours such as past website visited, and purchased history.

To captures user behaviour, work in (GCF10) proposed a methodology to analyse price discrimination by using fictitious users, mimicking a visit to shopping websites, from 6 different locations, for 7 days. Also in this case, results show that user location has an impact on price discrimination.

Another approach to study personalization of e-commerce websites is using data from both real and artificial users to better understanding price differentiation. In (Ha14), the authors extensively measures price steering and discrimination. With both real data collected through Amazon Mechanical Turks and synthetic data from controlled experiments using a non-GUI web browser (Hid16), the authors analyse the prices offered by a plethora of online vendors. The work found evidence of price differences by different merchants and retailers: their websites record the history of clicked products to discriminate prices among customers.

It has been reported that not all retailers employ price discrimination, Vissers *et al.* (VNBJ14) studied this phenomenon particularly on air-

line ticket. They evaluated this issue by analyzing 25 airlines continuously for 3 weeks, with dozens of unique user profiles. The experiments were automated by a headless Webkit browser (similar to that of (Ha14; KSHL$^+$15)). In this case, the results did not reveal the use of any systematic price discrimination. However, along with (MGEL12; CMR$^+$15), the authors also provide a useful way of constructing user profiles based on operating system, web browsers, user behavior and geographical location. These provide useful references for other studies, including ours.

In addition, Table 1 summarizes personalization measurements in different scenarios. The gray row refers to results achieved in state of the art papers, the rest is due to novel experiments and is deeply discussed in this thesis.

Personalization	Services	Based on	Results	Notes
web search	Google Search	user agents, genders, search history, browsing history, search-result-click history	no(HSMK+13)	
web search	Google Search	search history	yes(HST+15)	Sect,4.1
hotel recommendation	Orbitz	operating system	yes(Mat12)	
Google Ads	online news	genders	yes(DTD14)	
web search	Google Shopping	search history	yes(CHPD16)	Sect.6.1
news suggestion(1)	Google News	user settings (active personalization)	yes(CHPS16; DDJT15)	Sect.5
news suggestion(2)	Google News	online behavior based on publisher	no(CHPS16)	Sect.5
news suggestion(3)	Google News	online behavior	no(CHPS16)	Sect.5
price steering	Google Shopping	online user behavior (click links, visit sites)	no(CHPD16)	Sect.6.1
price steering	travel and retails sites	real-world data	yes(HSL+14)	
price steering	Google Shopping	location (cities, countries)	yes	Sect.6

Table 1: Personalization in different scenarios.

Chapter 3

Methodology

This chapter outlines the research approach of the thesis. Section 3.1 describes the principles followed when designing experiments and Section 3.2 provides the implementation details.

3.1 Principles of Experiments

In order to measure personalization levels of web search results, online news and commerce search results, we have developped a number of methodologies and techniques that have been the basis for performing multiple experiments in different scenarios and online services.

Although the online services are publicly accessible, we - the experimenters - have neither control over nor access to the model of the system (see, e.g., Figure 1). Therefore, it could be assumed that

these online services do behave identically for the same requests regardless of the differences in users. Thus, we would expect that a search engine returns the same results to all users for the same query without taking into account their location or their search history.

In order to investigate such an assumption about the online services, Tschantz et al. (TDDW15) and Englehardt et al. (EEZ+15) developed a number of methodologies for observing websites and services in order to detect and quantify the levels of online personalization. The authors of

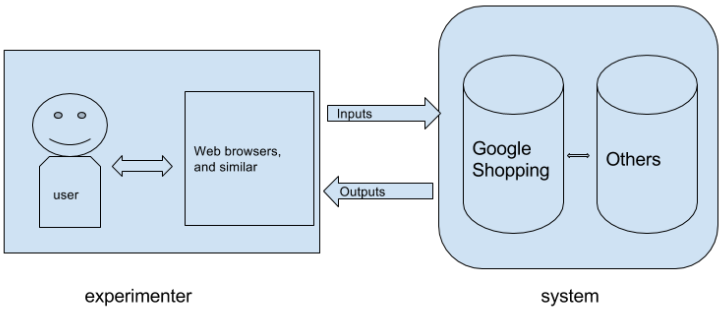

Figure 1: An example of an online system in the view of an experimenter.

the two papers propose various principles of experiment designs; among those, we list below those that could be applied to our specific studies.

Using an appropriate statistical test A statistical test aims at providing us with a technique for obtaining quantitative evaluation of a sample or of samples. Its purpose is to determine if there is enough evidence to "reject" a conjecture or hypothesis about the samples. The conjecture is called the null hypothesis.

In practice, a statistical test of data provides a *p-value*, the probability of seeing results the same as or more extreme as the actual observed data under the assumption that the null hypothesis is true. A small p-value implies that the data is unlikely under the null hypothesis.

To utilize a specific statistical test on data, the data must satisfy some prerequisite conditions (i.e., t-test requires the variable is continuous and has a normal distribution).

In reality, it is difficult or very complicated to force experiments' data following strict conditions for some statistical testing techniques . Therefore, the analyst must select suitable statistical tests based on the experiment's settings.

Being able to use similar experimental web browsers In the experiments, web browsers are run in several parallel sequences. The browsers from the same sequence must have similar settings such as operating system, or software version. By doing this, we do avoid the creation of cause-effect relations by the ordering of each browser inside the series.

Collecting data based on requirements of statistical analysis The data that we gather must satisfy the conditions of the statistical analysis. For example, when using *t*-test, we must make sure that the data are continuous and have normal distribution. Then we must use the analysis to determine the data to be collected. Also, it guides the analyst to design the experiments from the beginning. Hence, this step must be finished before executing the experiments.

Being selective In order to get meaningful results, we should experiment with those online services that could provide stable and consistent personalization effects. For instance, `google.com` provides personalization results for everyone (BH09).

The aforementioned principles are used in several studies. For example, they are applied in (DTD15) to measure personalized Google Ads, and in (DDJT15) to investigate personalized news. In addition, Englehardt et al. (EEZ+15) goes one step further to build a platform to automatize experiments to detect and quantify privacy-impacting behaviors. Their work has been exploited in several papers such as (AEE+14; HG12; SDA+16).

3.2 Implementation of Methodologies

Following the principles outlined in the previous section, we did perform many experiments and analyses. Those are described in details in the next chapters. However, our experiments share a number of features such as:

- methodologies to synthesize user profiles and to mimic their behaviour.
- techniques to prepare system environments and to automate experiments.
- analysis techniques to measure personalization degrees.

Below, we describe such common features that will be instantiated in the following chapters when considering specific experiments.

Creation of users profiles An important step in all our experiments is the ability to synthesize user profiles. In simple words, a user profile contains data to fabricate certain features of a real user of a web browser. We simulate users by first associating them with an account (e.g., Google Accounts), to guarantee that all on-line traits of the profile are stored on the Internet services; and then by associating them with *browser profiles* to store in the local file system the information about users' activity.

In fact, a web browser saves personal information such as bookmarks, histories, and cookies in a set of files, called browser profile, which are kept in a separate folder of the local disk; the browser can have multiple browser profiles, each containing a separate set of information about users.

In our first simulation approach, we manually created the user accounts via the registration page of the online services and used them in the study personalization of web search results in Section 4.1. In our study, to make sure that the profiles did simulate real user behaviours, we performed extra activities such as logging into the accounts, viewing contents (i.e., emails) and modifying user preferences.

However, when the number of accounts increases significantly, manually creating accounts is a time-consuming tasks that is exposed also to human errors, such as typos or entering of wrong settings. It could then be difficult to perform large scale experiments automatically.

In order to resolve the drawbacks of the first method, in the second experiment we automatically created a large quantity of user profiles

by associating browser profiles with user profiles. Due to the nature of browser profile, that number of user profiles is only limited by systems capabilities (i.e., memory size and disk capacity).

In order to enhance the effectiveness of this method, we improved it gradually from study to study. For the study described in Section 4.2, we created multiple user profiles that stay alive for a short time during experiments. In Chapter 5, we describe longer experiments and kept the user profiles active during the whole experiments. In Chapter 6, we report experiments at a larger scale in terms of time and number of profiles, in such a way that user profiles were stored persistently in the local disk all the time and could be restored in case of crashes.

Simulation of users behaviour To make our experiments effective, we need to guarantee that artificially created user profiles do perform activities as similar as possible to those of the real ones. We achieve this, by designing sets of activities, such as clicking and searching, and let each synthesized user perform them. Also, we build user profiles with particular traits by forcing them to perform specific sets of actions. These specific actions are tailored to the specific experiments.

For the study in Section 4.1, we manually log in the online services after registration. We also do some common actions such as reading emails to convince that real people operate those profiles. Besides, we let them search for a list of specific keywords to mimic the real users with particular interests. Those keywords are manually selected by us. In Chapter 5, we go to a step further by letting users choose links from certain sources to click on. In Chapter 6, users can decide on which links to click based on specific values (price values). Artificial users profiles can visit websites and scroll web pages for a specified number of seconds and their interests are built by searching for the keywords associated to specific kinds of information.

Selecting system environment For all experiment, we tend to use computer systems with similar configurations. This allows us to avoid getting noise data resulting from systems differences. Mainly, we use Ubuntu -

a Linux-based operating system- and the most recent version of Firefox web browser.

When we simulate users from various geographical locations, we use different remote computers from these places and control them by tunnelling via proxies. Depending on the experiment, we add extra configurations. For example, in the study in Section 4.1, we modify the host files to match server's IP addresses of a website. In Chapter 5 and Chapter 6, we manage multiple computers simultaneously via SSH tunnelling and SOCKS proxy.

Automatising experiments All experiments are performed automatically by commanding multiple web browsers. Fortunately, several frameworks have been proposed for interacting with web browsers and analyzing results from search engines, Table 2 gives an overview of the most widely used tools. In our studies, we develop tools based on: PhantomJS (Hid16), AdFisher (DTD15) and an OpenWPM (EN16c). The tools help us to control all browsers simultaneously, and to save web pages and other user related data (i.e., cookies). The tools also permit keeping browsers instances completely isolated from each others and thus user profiles fully separated. The tools can redirect Internet traffic to a designated machine via proxy if needed.

In the experiments of Section 4.1, we utilize scripts to manage several PhantomJS web browsers - a special browser for software testing purpose, and let users send queries to a search engine. In Section 4.2 and Chapter 5, we use an Adfisher-based software to control Firefox web browsers and define user behaviours. For example, the user can click on specific links, visit and scroll the web pages for a specific amount of time. In Chapter 6, we develop an OpenWPM-based software to perform experiments. This tool not only controls web browsers and user activities but can also be used to save and restore user profiles from the local disks.

Selecting similarity metrics In order to evaluate the differences between the results obtained by the profiles and to perform further analysis, we employ different metrics. In particular we use the Jaccard Index,

Table 2: Overview of automated tools for controlling web-browsers and analysing collected results from websites.

Tool	Medium	Cons	Pros
Flow experiments code (HSMK+13)	PhantomJS	Not compatible with websites like `bing.com` and `google.com`	Small scripts, run on multiple OSs
AdFisher (DTD14)	Selenium (to control Mozilla Firefox and Google Chrome)	No support to restore and backup browser profiles	Well-designed, easy to extend
Fourth-party (MM12)	Browser extension	No automated web browser	Can be controlled by automated frameworks
Bobble (XMD+14)	Browser extension	No full control over searched terms, time and locations, since it is based on crowd-sourcing	Data from real internet users
OpenWPM (EN16a)	Selenium	No module for analyzing data	Works at large scale

Edit Distance, Kendall Index and NCDG (Normalized Discounted Cumulative Gain) that have been already used in many related studies, like in (HSMK+13), (DBG15), (BIMHL06), (KAH12) and (SGC13). Below, we briefly describe these metrics.

- **Jaccard Index**: measures similarity between contents of two given sets, without taking into account ordering of elements of these sets. The Jaccard index for two sets P and Q is 1 when the two sets are identical and 0 when their intersection is empty. This metric is used for the analysis performed in Chapter 4, Chapter 5 and Chapter 6.

- **Edit Distance**: measures how many operations (insertion, deletion, substitution or swap) are needed to transform one list into another.

For example, if user u_i receives search result set [A, B, C, D] and user u_j gets [D, C] for the same query, then the Edit Distance is 3 (two insertions and one swap). It is used in Chapter 4 and Chapter 5.

- **Kendall Index**: measures ordinal correlation between two ordered lists of results. Intuitively, the Kendall Index between two lists will be high when items of the two lists have a similar rank, and low when they have a dissimilar rank. Given two ranked lists P and Q, the index ranges from -1 to 1, where 0 means no correlation, 1 represents the same order and -1 is inverse order. It is used in Chapter 5 and Chapter 6.

- **NDCG**: measuring the similarity between a given list of results and the ideal list of results. For each result r, there is one gain score g(r), representing its price. For a result page $R = [r_1, r_2, ..r_k]$, we have $DCG(R) = g(r_1) + \sum_{i=2}^{k}(g(r_i)/log_2(i))$. NDCG is $DCG(R)/DCG(R')$, where R' is the ideal result page (a list in which the results are shown from the most expensive to the least one). For instance, in searching for products, we create R' by unionising the results returned, for the same query, to all profiles. Then, we sort such results from the most expensive to the least expensive prices. This metric is used for studies in Chapter 6.

In the next chapters, we will show how to the above described techniques are employed for our specific studies, and will provide further details of their actual implementations.

Chapter 4

Experimenting on Google Web Search

Major search engines deploy personalized Web results to enhance the users' experience, by showing them data, which are supposed to be relevant to their interests. Even if this process may bring benefits to users while browsing, it also raises concerns about the selection of the search results as deeply discussed in Chapter 1. In particular, users may be unknowingly trapped by search engines in protective information bubbles, called "filter bubbles" (Par11b), which can have the undesired effect of separating users from information that does not fit their preferences. This chapter describes early results on quantification of personalization over Google search query results and contains two studies. Each study contributes an analysis from a different perspective of personalization.

The first study aims at sheding light on measuring web search personalization. In order to do this, we have carried out some experiments consisting of search queries performed by a battery of Google accounts with differently prepared profiles. Matching query results, we quantify the level of personalization, according to topics of the queries and the profile of the accounts. This work reports initial results and it is the first step for a more extensive investigation to measure web search personalization in the next chapters.

The second study measures the ranking of Wikipedia pages within a set of search results, given some specific keywords, in order to emphasize the problems of personalization: *is there some specific, recurrent, domain pages, which Internet users are frequently exposed to?* We have chosen Wikipedia for this case study, due to the fact that it is inside the top ten of most visited websites on Alexa worldwide ranking (6th , accessed in February 5th, 2017[1])

The results show that, at about 8 times over ten, Wikipedia pages are in the top five highest positions of the search results page.

This chapter is organized in 2 sections.

Each of them will discuss one study and following the orders above. Section 4.1 explores personalization on Google web search. Section 4.2 describes the ranking of web pages in Google search results, with Wikipedia as a case study.

4.1 Domain Specific Query and Web-search Personalization

To understand how filter bubbles take shape for a given Web user, first it is necessary to assess the level of personalization of results provided to that user by search services. Previous work has evaluated personalization of Web results, exploiting user's features and history information (see, e.g., (MR11; NOD14; MSS14)). In this study, we aim at understanding the level of personalization proposed by a commercial search engine —Google— to its users, in comparison with non-Google users. A Google user is one that has a Google account and logs into her Google account before performing any search activity. A user who accesses Google search service without logging in is called non-Google (or *vanilla*) user.

We check the effects that logging into Google accounts have on search results by building Google user profiles based on search histories and queries' topics. After that, we performed experiments with different Google user accounts settings on local network and then evaluated the

[1] `http://www.alexa.com/topsites`

effects of Web search personalization, comparing to non-Google users.

This study is organized as follows. Section 4.1.1 shows the methodology and the experiments. Section 4.1.2 reports details and comments on experiment results.

4.1.1 Experiments

In this section, we describe the methodology followed for quantifying the current level of personalization of Google Web searches and we illustrate the settings of the conducted experiments.

4.1.1.1 Methodology

In our study, we quantify personalization level by measuring differences in search results between users when they query for the same keywords. The differences can be in the relative ranking of search query results and in the results themselves. We also check if there are differences between a Google user - profiled by means of the queries - and a plain Google user, with a plain profile (or a non-Google user).

In order to compare two search query results, we only considered the normal query results, ignoring the Image and News parts. As evaluation metrics, we employed the Jaccard Index and the Edit Distance, also used by Hannak *et al.* (HSMK$^+$13).

We describe in the following our experiments. First, we created multiple Google accounts with different values in gender, age and location. The accounts were made by normally registering via Google website. Then, we logged in all of them and executed the same queries simultaneously, on a single IP address in different browser sessions. We used a single IP address to eliminate the potential noise due to a difference in geographic locations, which is one of the most significant elements affecting search results, see (HSMK$^+$13). At the same time, we used many other vanilla users to execute exactly the same queries. Once the experiments ended, we compared result pages in terms of Jacquard Index and Edit Distance.

In a second phase, we have built Google accounts with narrower interests, labeling them as profiled Google accounts. These accounts were normally registered from Google and they only searched for words in lists of keywords (the *training keyword list*) about specific and narrow domains. In particular, we have considered the football domain and built a list of keywords composed by football players, coaches, staff people of football clubs (e.g., “AC Milan coach Nereo Rocco", “Fly Emirates" - a sponsor of AC Milan, “football Inter Milan Rodrigo Palacio" - a player of Inter Milan). We also set up another keywords list (the *test keyword list*) containing deliberately vague terms, again related to football teams, such as “next match" and “home stadium". Then, we used each user to execute search queries in the same domain, but with different training keyword lists at the same time. Later, we evaluate whether those differently built accounts received results influenced by their history of queries. In order to do that, we asked them performing queries on the test keywords list. Moreover, we also compared their results with those of users that did not search for the training keywords list.

Summarizing, we performed the following experiments:

- Experiment 1: We have tried to build profiled accounts, letting each of them search in different domains, in a training phase. Then, they searched on a test keyword list, in a domain different from the training ones. By comparing web results on the test queries, it is possible to measure the personalization effect of the training queries.

- Experiment 2: We have tried to build profiled accounts, letting them search for keywords in a narrow domain. Then, they searched on the test keyword list (a less narrow domain) to evaluate if some relevant differences emerge. As before, by comparing web results on the test queries, it is possible to measure the personalization effect of the training queries.

- Experiment 3: We repeated experiment 2 with different query composition, in order to evaluate if and how this affects the user profile and its Google exploitation.

- Experiment 4: We have tested how the test keyword results change when a Google user performs massive queries on narrow topics for a long time.

4.1.1.2 Experiment settings

Our experiment settings and tools are strongly inspired by and adapted from original work of Hannak *et al.* (HSMK+13). All experiments run on PhantomJS[2], a web-browser based on Webkit. Basically, it is a full browser with a Java-script engine as well as a modern web navigator software, but without a user interface. We acknowledge that Google uses many servers with several IP addresses. Since each server could have different index databases according to its location, to eliminate the risk of noise, we fixed one IP address for the Google server in the *host* file of the operating system. In this way, we expected that all search requests are routed to a single server address of Google.

Using PhantomJS, each account is logged in right after its creation, to mimic the behavior of a real user. When performing a search, all accounts operate at the same time, with an interval of 11 minutes between two searches, to avoid carry-over effect (HSMK+13). The carry-over effect is a phenomenon that happens when users perform two sequential queries A and B: the results of query B are influenced by previous search for A. For example, if one - not necessarily profiled - searches for “python" and after searches for “programming language", it is likely to see results relevant to the Python language. Details on the experiments settings and results is available online at `https://sites.google.com/a/imtlucca.it/wwv2015/`.

4.1.2 Results

In this section, we comment on the results of the experiments described in Section 4.1.1.

[2]http://www.phantomjs.org

4.1.2.1 Experiment 1

In this experiment, we check how the details in users' profiles influence their search results. In particular, we have created three Google accounts and we have let them search keywords in the following categories:

- Account 1, with training keywords from shoes and baseball categories;
- Account 2, with training keywords from drinks, foods, and retail brands categories;
- Account 3, with training keywords from politics, fashion, and shopping categories.

All Google accounts in our experiments have been manually enrolled from Gmail registration page. The keywords are what the accounts are querying for and the queries are grouped by their semantic meanings. In particular, we have used keywords from Google Trends (August 2014, US), which has categorized popular search queries in corresponding categories. We let all accounts search for those keywords. After this training phase, they search on test keywords list.

Number of search terms	Number of test terms
140	19, running time around 24h

Table 3: Setting for Experiment 1

In the experiment, 4 out of 38 test queries produced different results with respect to different accounts. When we check the natural language meaning of the different ones, we were unable to find a clear connection between them. For example, with test keyword "Plato" searched by Google accounts 1 and 2, Jaccard Index is 0.9, meaning 90% of the results are the same (we checked the first 10 results shown by Google). The only difference between them is 1 web link about "PLATO 2.0 - A space agency" which has no correlated meaning to the previous training searches of both the users (Fig.1). Quite obviously, checking only the first

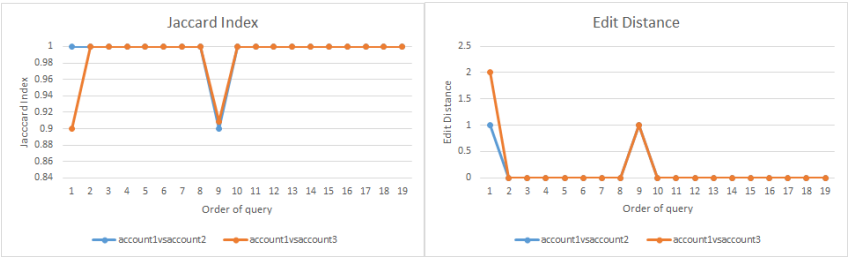

Figure 2: Jaccard Index (Left) and Edit Distance (Right) for Experiment 1.

ten results provided by Google do not give us the capability to conclude with a final claim on the fact that topics in previous queries interfere with results of the test queries. What we can assert is that, for previous queries belonging to the categories in the list shown above, the three accounts under investigation obtain no significant differences in the first ten results on the test queries.

4.1.2.2 Experiment 2

In this experiment, we have tried to train Google accounts to emulate the fact that they have a strong interest in football. We chose two football teams (the Italian AC Milan and Inter Milan) and we collected keywords from their official websites. Those keywords are directly related to the corresponding football clubs, such as history of achievements, captains, coaches, leader boards, management staff and players. All the search query terms exist in the football club official websites.

Number of search terms	Number of test terms
134	26

Table 4: Setting for Experiment 2.

After calculating Jaccard Index and Edit Distance, we found that there were 6 out of 25 test queries yielding different results (see Fig.2). However, those different results were not connected to the two football clubs

under investigation (i.e., when searching for “next match", the different results were not about next match of AC Milan or Inter Milan teams). Again, such an outcome let us to assert that, even if one account has searched for very narrow domains in its past activity over Google, apparently that activity does not influence new results when searching for vague, still related, domains. However, further investigation is needed for more rigorous claims on the matter.

4.1.2.3 Experiment 3

Experiment 3 was a revised version of Experiment 2. In Experiment 3, the search query had the form “football" + [club’s name] + keyword.

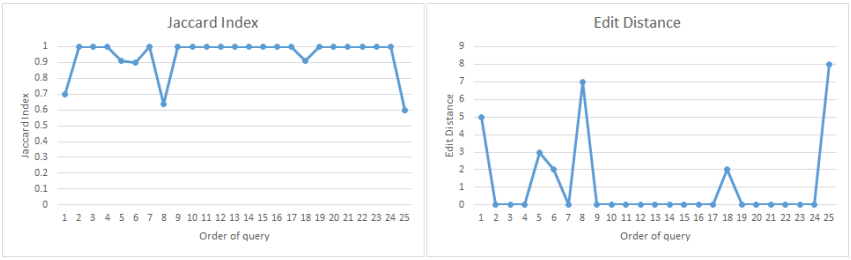

Figure 3: Jaccard Index (Left) and Edit Distance (Right) for Experiment 2.

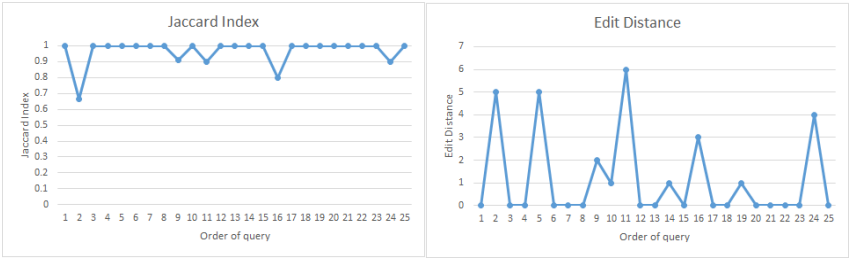

Figure 4: Jaccard Index (Left) and Edit Distance (Right) for Experiment 3.

As shown in Fig. 3, the order of search results among the accounts under investigation shows more differences than the previous experiment.

This paves the way for further investigation.

4.1.2.4 Experiment 4

Search topics	Notes
layers, captains, technical staffs, coaching staffs, management history, administrative staffs, name of championship (i.e. won cups), list of sponsors (technical and/or official), list of captains in history	There are repeated keywords

Table 5: Keyword setting for Experiment 4.

We have selected 400 keywords (topics in Table 3) about one football team and we have created 2 new Google users to continuously perform searches for 72 hours. After a massive number of search queries in a narrow topic, we expected the results from test keywords to be significantly different. Instead, both Jaccard Index and Edit Distance showed a limited variation.

We compare our experiment ideas and setting with a most similar and recent study of Hannak et. al. (HSMK+13). In that study, the authors compare search results of the same queries between users to measure personalization level, which is basically the same as what we did. By employing 200 real internet users from Amazon Mechanic Turk (a platform to hire real user to complete certain tasks with fees), they found that geographical location has a high impact on personalized contents. Compare to our works, there are differences. First, we use a slightly bigger number of queries (140 and 134 in our experiments compare to 120 in the mentioned study). Second, instead of using query from multiple categories, we focus the query in two or three domains with aims to get more specific results. Lastly, we fully control our user activities by synthesizing them in specific conditions. Also, our analysis concentrates on the differences between personalized and non-personalized based on user's behaviors, not based on other factors such as demographic features or computer specification.

4.2 Measuring ranking of websites on Google search

As observed by a recent article of Nature News (Hod15), "Wikipedia is among the most frequently visited websites in the world and one of the most popular places to tap into the world's scientific and medical information". One of the seventh most visited websites[3], the online encyclopaedia is a dominant source of Internet knowledge.

Remarkably, a 2012 study assessed that Wikipedia pages appeared in 96% of the results all the searches through Google UK (SC12). Given that users usually focus on the first few results of their web searches (CG07; HL09), the exclusive privilege of Wikipedia at the very first positions could bias the informative content reachable on the Internet.

In this study, we measure the ranking of Wikipedia pages on Google Italia. To the best of our knowledge, it is the first study of this kind for the Italian language.

The procedure is as follows. For our web searches, we concentrate on Italian keywords (and set of keywords).

To collect the most popular search keywords, they have been chosen from Google Trend, from the Google Display Planner[4], and from the Italian trending words on Twitter. Google Trend gives the most searched terms in a year, as well as trending searches in the past 24 hours[5] and trending searches in the recent past[6]. Google Display Planner is a tool providing a series of websites linked to specific categories. It has also been used, to look for suggested keywords tied to particular categories, like, e.g., Sport and Vehicles. We have performed all the searches on Google Italia, from Italy. To avoid personalised results, we have used newly created browser instances and we have simulated users not logged into Google. The default settings for searches were the Google default settings. Among the obtained results, we considered only organic results and not sponsored one (i.e., Advertisements, Google News, and so on).

[3] `http://www.alexa.com/topsites` (7th March, 2016)
[4] `https://support.google.com/adwords/answer/3056115?hl=en`
[5] `https://www.google.com/trends/home/all/IT`
[6] `https://www.google.com/trends/hottrends#pn=p27`

In the following, we present the experiments and the results.

4.2.1 Experiments

Our reference date is April 7th, 2016. We have collected the top search terms on Google Italia from 2011 to 2015. Also, we have extracted the trending news for the reference date and the hot trends stories of the ten days prior to that date. As a whole, we obtained 1169 unique search terms. For a wider view, we have further gathered the top Italian trending keywords from Twitter (updated three times and within three hours, on April 7th, 2016), leading to 40 unique terms from Twitter.

For the experiments, we have extended AdFisher(DTD14), an automated tool for information flow experiments, freely available at GitHub[7].

In our work, AdFisher runs browser-based experiments that emulate search queries and store the results. AdFisher interacts with Selenium, a web browser automation tool. Selenium allows to run a unique instance of Firefox creating a fresh profile, with new associated cookies.

For each keyword (or keywords set), a new browser instance has been launched and we have searched such keyword(s). The browser instance has been destroyed after saving the query results. New browser instances have been used to avoid the so called carry-over effects (HSMK+13), which would lead to results for the current search being influenced by the previous searches. The time interval between two searches are between 30–60 seconds. The experiment finishes within 12 hours

Examples of keywords that we have searched for are "Elezioni presidenziali negli Stati Uniti d'America del 2016" "Credito Valtellinese", "Una lama di luce". The complete list of keywords (and keywords set) is available from `https://goo.gl/9KasJc`.

4.2.2 Results

Experiments were performed with more than 1200 keywords, spanning 33 categories, with an average of 21 keywords per category. In order to evaluate the search results, we have considered only those keywords for

[7] `https://github.com/tadatitam/info-flow-experiments`

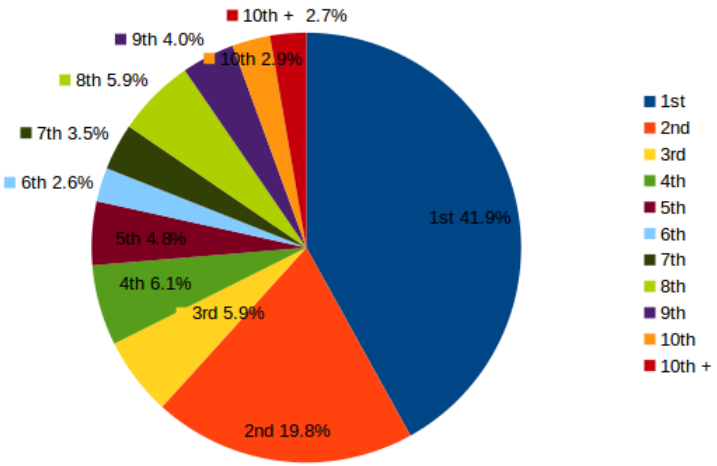

Figure 5: How frequent Wikipedia pages are ranked *i-th*, in return to *google.it* searches

which the corresponding Wikipedia link appeared in the first page of the result list - this corresponds to 708 keywords. Searching for those keywords, we found that Wikipedia pages in the result lists are ranked *1st* 41.1% and *2nd* 19.9%. Overall, they appear in the first five positions 78.8 times over 100. Figure 5 shows the occurrence of Wikipedia pages according to their position.

Searching keywords belonging to “Topic Emergenti" and “Mostre d’-Arte" always yield Wikipedia links as the first result. Searching keywords belonging to “Assicurazioni", “Offerte di Lavoro" and “Economia e Finanza" yield Wikipedia links as first result for less than 10% of such keywords, see Figure 6.

Amongst all, “Mostre d’Arte" is a *top* category: all the keywords belonging to that category always lead to results with a Wikipedia link located within the first three Google Italia results. We only consider top three results because they account for more than 60% of average traffic share for the first 15 results[8] . Examples of keywords are “Picasso -

[8] https://chitika.com/google-positioning-value

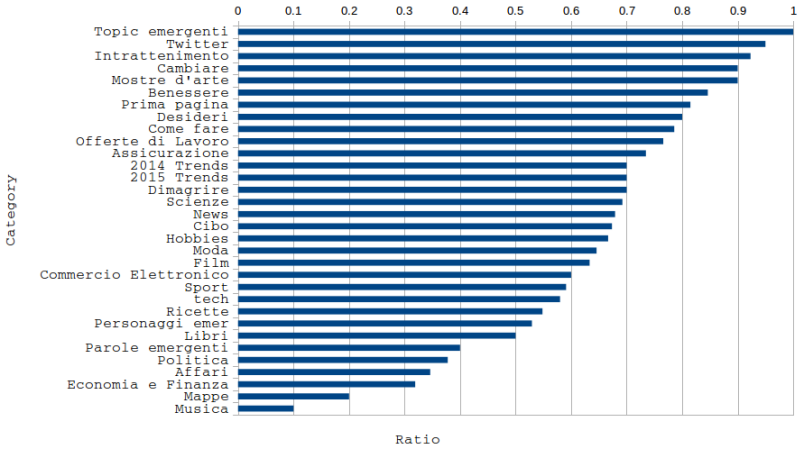

Figure 6: Result pages with a Wikipedia link as 1st result, per category

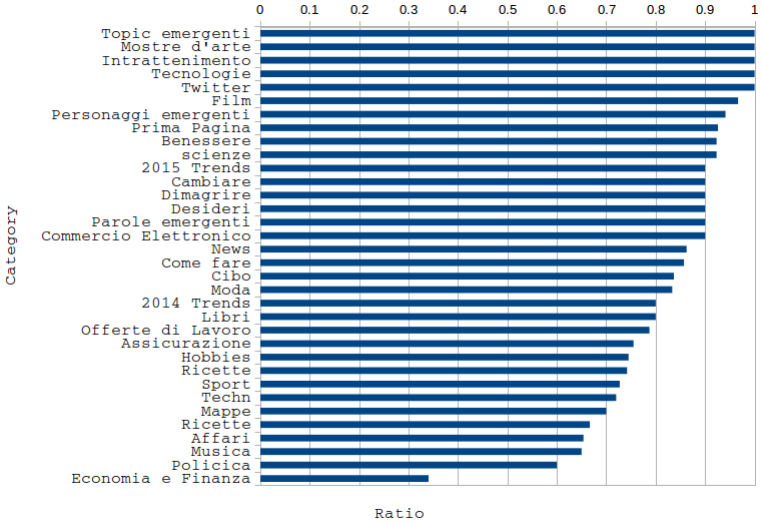

Figure 7: Result pages with a Wikipedia link in the top three results, per category

Milano", “Renoir - Pavia" and “Leonardo - Venaria". “Scienze" and “Benessere" are top categories too. More than 50% of searches of related keywords lead to Wikipedia results in the top three positions, see Figure 7. Instead, only 2% of keywords in the “News" category lead to results in the top three positions, while results for keywords related to “Politica", “Economia e Finanza", and “Hobbies" are ranked in the top three positions less than 10% of times (Figure 7). Figure 6 and Figure 7 show only the categories with the largest number of keywords.

Work in (SC12) performed a similar analysis on Google UK, focusing on encyclopaedic subjects, like scientific and natural sciences. Wikipedia scored extremely well, being its links in the top two result positions. The keywords belonging to the “Scienze" category (the Italian word for “Science") obtained more than 60% of Wikipedia links in the top three result positions.

To date, our work is the only study about measuring the ranking of the links to Italian Wikipedia pages on the Italian version of Google.

Chapter 5

Experimenting on Google News

In this chapter, we focus on news aggregators. The goal is to measure the level of personalization of results returned to different kinds of users when they search over a news dataset. Specifically, the target of our analysis is Google News.

Google News offers a panoramic view on several articles on different subjects, redirecting the users on the publishers' accounts. 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 or on specific topics.

We consider two aspects of news personalization, defined by the related literature as *expected* and *unexpected passive personalization* (DDJT15). 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.

5.1 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 (LDP10) 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, 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 (DDJT15), 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 (DDJT15) 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 (DDJT15). 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 (DDJT15), 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 (DDJT15) suffered from the limitation that the chosen publisher could not have been present, given the intrinsic volatility of news.

5.2 Methodology

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 [1], that is publicly known to make use of user profiling to provide targeted advertisements (HJK12).

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 behaviours of users interested in *Sports*. We, then, exploit such behaviours to investigate either expected and unexpected passive personalization over

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

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 (DDJT15), achieved however through a different experimental setting. The current study contributes to 1) evaluate if news are sorted with or without regard to past behaviours 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., (ZRJT10).

The rest of the work is structured as follows. Next section briefly relates on Google News news personalization. In Section 5.3, we introduce the reference dataset we start from in our experiments. Section 5.4 describes the techniques used to extract relevant entities from the *Signal Media* news titles. Section 5.5 presents settings and implementation of the training and the test phase over Google News and gives experimental results.

5.3 Reference Dataset

For our experiments, we refer to the *Signal Media One-Million News Articles* dataset[2]. 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

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

these articles include major web sites, such as `espn` considered in this work, 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 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 8).

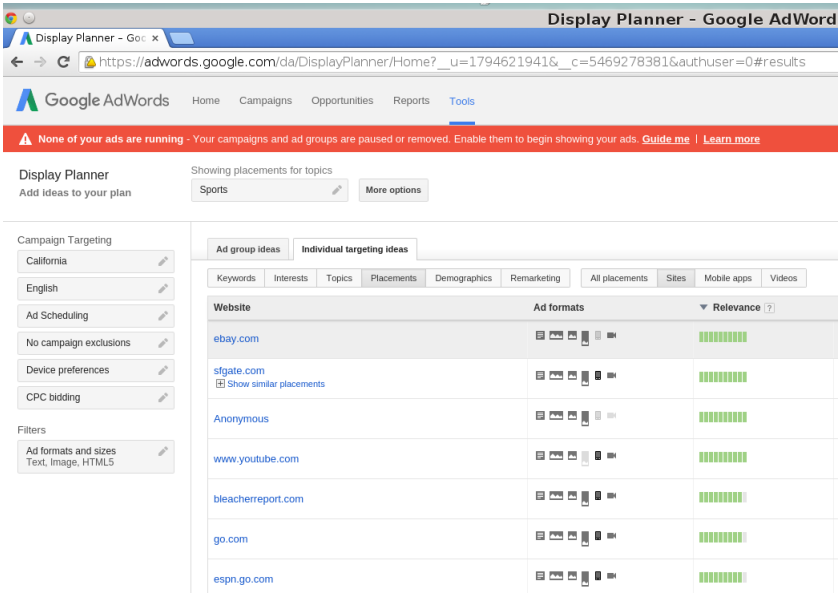

Figure 8: Google Display Planner: `espn` belongs to the Google Display Network.

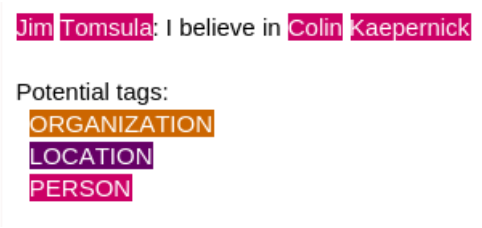

Figure 9: Stanford Named Entity Tagger.[5]

5.4 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 (FGM05). 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[3]. Thus, we exploit the Natural Language Toolkit NLTK[4], 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 9 shows an example of entities extracted from the titles of `espn` news, while Figure 10 shows the most recurrent entities under the form of a word cloud.

On a total of more than 7,000 news titles from `espn` and affiliated sub-

[3] `http://stanfordnlp.github.io/CoreNLP/`
[4] `http://www.nltk.org/`
[5] `http://nlp.stanford.edu:8080/ner/`

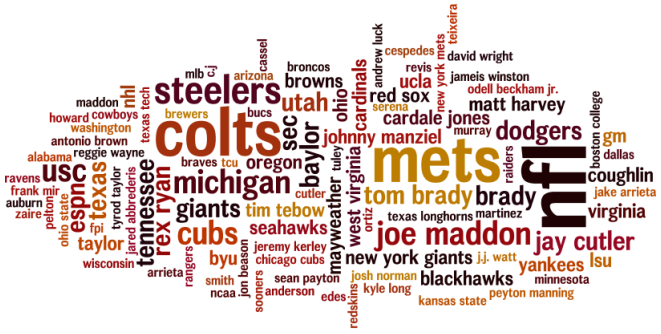

Figure 10: The most recurrent entities in titles with source `espn`.

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.

5.5 Experiments and Results

The final goal of the experiments is measuring how users past online behaviour affects the news provided by Google News. We investigate two properties, expected (EPP) and unexpected (UPP) passive personalization, as in (DDJT15). The expected version of passive personalization has been disclosed by Google in (LDP10). 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. Please refer to (EN16b) for

a full survey of measurement tools and their comparison. We have chosen to adopt AdFisher[6] (TDDW15), a tool to analyze interactions between online user behaviours, 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., (DDJT15). Further, a branch version has been created for handling product searches in Google Shopping, to measure price steering (CHPN16). 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., (CHP16).

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[7] 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[8], all our experiments are with the users connected from US. We have used the Digital Ocean VPS Service Digital[9] to gain access to machines located in the US.

5.5.1 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

[6] https://github.com/tadatitam/info-flow-experiments
[7] http://www.seleniumhq.org/
[8] http://www.alexa.com/siteinfo/espn.com
[9] https://www.digitalocean.com

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 5.3). The time interval between two sequential queries is a random value between 20 and 60 seconds, to mimic real users activities
- She clicks only on the results of the news with source espn.com and spends some time on the linked page.

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 [10], 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 behaviour.

Table 6 reports the training details.

Table 6: Training behaviour

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.

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

5.5.2 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 behaviour until the SGY section appeared. This happened after four iterations of searching keywords and visiting websites. Each iteration is executed in a working time frame (9 a.m. to 5 p.m.) and lasted within 8 hours. Then, we stopped and continued on the next day. 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 behaviour described in Section 5.5.1. 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 6, 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 behaviour 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 behaviour 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 behaviours of such users yield to distinct results in that section.

- Training: for the first user, we have repeated the training described in Section 5.5.1. 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[11]) 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.

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

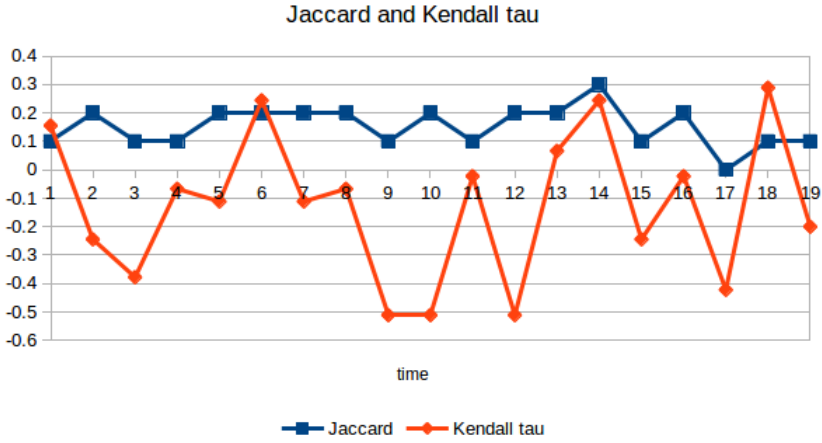

Figure 11: Jaccard and Kendall indexes of the SGY sections.

In Figure 11, 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 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 7, 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.

Figure 12 highlights the SGY section of the two users, at testing time.

Table 7: 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	All
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**

As expected, users have been shown different results.

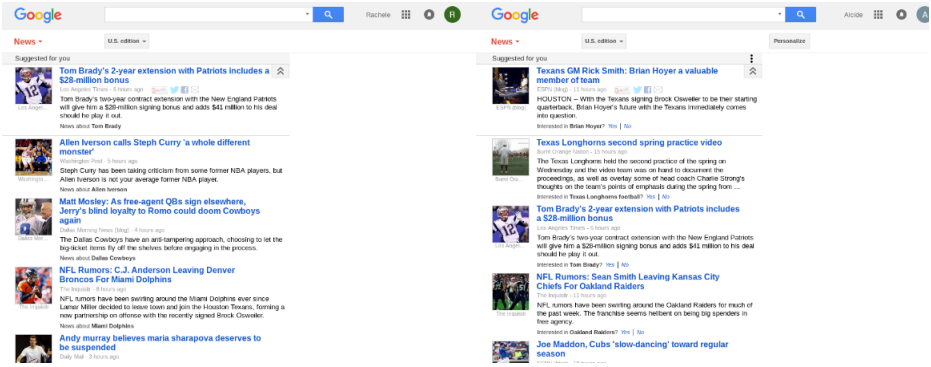

Figure 12: Snapshot of the two SGY sections: differences and similarities.

We compare our work with that in (DDJT15). Technically, we both rely on similar approaches to carry on the experiments. In the mentioned work, the authors also performed experiments to detect unexpected passive personalization and expected passive personalization and they achieved similar conclusions as ours: while EPP exists, there is no evidence of UPP. However, we enhanced the procedures for selecting the keywords, so that to act in a systematic way (see 5.4 and 5.5): this should strengthen the achieved results. We also described in details our evaluation metrics, this has been instead slightly discussed in previous related work. In addition, in our work the personalized results emerge after four days, instead of one week as in (DDJT15).

Chapter 6

Experimenting on Google Shopping

Today, multiple e-commerce websites apply personalization on their content, such as Amazon does for product suggestions, YouTube for recommended videos, and Spotify for music genre recommendation. As an advantage, personalized content may help users to easily reach desirable products and services: as an example, Spotify hints many artists and albums, according to the users' listening history.

However, there exists a kind of harmful personalization, according to which the order of search results is changed, to highlight specific products" and products prices. This is known as the practice of "price steering". In this chapter, we first attempt to quantify the price steering level in search results shown to different kinds of users on Google Shopping. Then, we apply known price steering measurements techniques to measure how prices change, according to various users' locations - distinguishable for their high economic performances.

6.1 Online Behavior Modeling with Application to Price Steering

Price steering refers to the practice of changing the order of search results to highlight specific products prices (MGEL12). In this work, we aim at studying if e-commerce websites rely on user past online behaviour to show her different product prices. In particular, we focus on Google Shopping, to discover if Google shows products of different price based on the *user willingness to pay*.

Google Shopping[1] is a promising platform to study the effect of price personalization. It allows vendors to reach a large number of customers, really interested in specific products, thus showing the right product to the right customer. Google Shopping creates a selling campaign, placing specific products "in front of millions of online shoppers searching on Google.com"[2]. This is possible since Google can access several information on the user search activity, not only including that on Google Shopping. Actually, Google monitors the circle of websites known as Google Display Network (GDN), a large set of websites publishing Google ads[3].

As shown in (HJK12), Google builds *ad user profiles*, monitoring and learning behaviors when the users navigate on the GDN websites. Among the elements considered to build the ad profiles, there are the list of the visited websites, their topics, the time spent on each website, the number of times the user went to the website, the device the user is using for accessing at the platform, geo-localization of the user IP address.

Past work showed that price steering is affected by the user location, see Chapter 2. Our work focuses on how the user behaviour, e.g., visiting a website of luxury goods, clicking on expensive products, affects the price of the items shown as the result of future queries over Google Shopping. We emulate the on-line behaviour of an *affluent* user and we compare her results list with one of a fresh user, which has not searched before on Google Shopping. Preliminary results show that, overall, af-

[1] http://www.google.com/shopping

[2] https://services.google.com/fh/files/misc/product_listing_ads_intro.pdf

[3] https://support.google.com/adwords/answer/2404190?hl=en

fluent profiles have been shown different results with respect to those shown to the fresh control user. However, there is no a fixed rule, leading first to the most expensive products shown to the affluent users. The difference in the results list is however worth to be acknowledged, and calls for further investigation, with a more complex experimental settings and a more extensive evaluation.

The rest of the work is as follows. Section 6.1.1 describes our methodology. In Section 6.1.2, we describe the experiments and we give the results.

6.1.1 Methodology

Our goal is measuring how the user online behaviour affects price steering on Google Shopping. We use the approach which is similar in (E^+14; Ha14; TDDW15). At first, we consider users of two categories: *affluent* and synthesized *control* users. Intuitively, the former feature higher willingness to pay than the latter. Thus, aiming at mimicking their behaviour, we assume that affluent users search and click on more expensive products than control users. We have considered two kinds of behaviours: 1) visiting web pages, and 2) searching for keywords on Google Shopping. Visiting a page means staying on that page for a while and also scrolling the page. We define the affluent user behaviour as follows:

- an affluent user visits websites selling luxury goods;
- an affluent user searches for keywords representing luxury goods on Google Shopping and
  - she visits the three results representing the first three products whose price is above the average (among all the obtained results)
  - she visits those result pages that are the same of a previously visited website selling luxury goods.

The behaviour of a control user is different, she is idling while the affluent is in action.

Our expectation is that, when users query Google Shopping after a training phase where they behave as described above, affluent (resp., control) ones will likely see the highest (lowest) price products ranked first in their list of results.

In order to identify websites and keywords for luxury goods, we have exploited Display Planner tool of Google AdWords. The tool guides helps us to find websites and keywords inherent to specific topics and terms[4].

It is well known that Google monitors the users' behavioural activities over the Internet through tools such as Google Analytics, Google Plus, and the Google ads system[5].

Thus, we train two affluent user profiles according to the specific online behaviours described above. Each profile has a control user profile associated. A control user has the same configuration as the user she is associated to (same browser, same OS).

The difference is that control users are not logged into a Google account. Then, we compare the results obtained searching on Google Shopping the same keywords for both the trained users and the control users.

In detail, the training and test phase are as follows:

- Training step 1: The two affluent users visit a list of websites with topics related to their category. Websites have been chosen using the Google Display Planner.
- Training step 2: The two affluent users search on Google Shopping for keywords related to their user category (keywords have been chosen according to the Display Planner, too). Then, they click on the most expensive product results and on those results coming from websites visited at step 1 (when present).
- Test: All users (trained plus control) search for new products on Google Shopping. They do not interact with the results.

We let the two affluent users repeat the training phase eight times. At the end of each training, we run the test. This is for building a longer

[4] `https://adwords.google.com/da/DisplayPlanner/Home`
[5] `https://www.google.com/policies/privacy/`

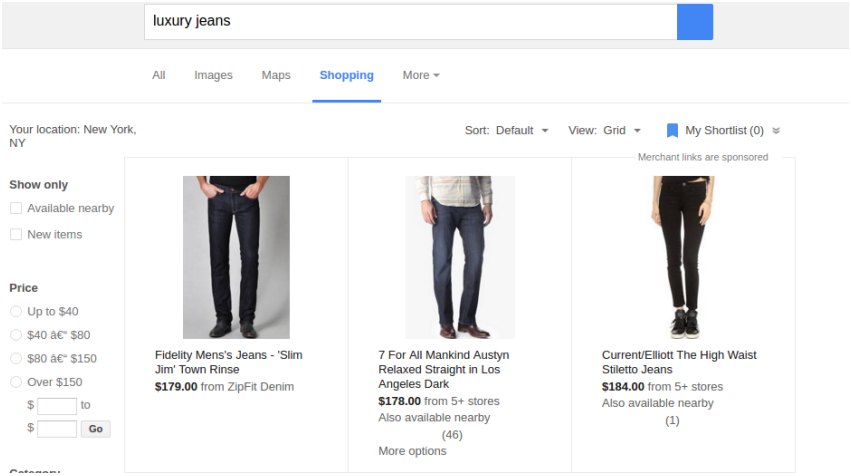

Figure 13: Top 3 results on Google Shopping: control, 3rd test block

behavioral history of the user. Indeed, Google itself states that it uses browsing and search histories to personalise the results and enhance the user experience[6]. This is why we have repeated the training phase eight times, always on the same profile, to emphasize possible personalization aspects. One evidence of the efficacy of this modality is in (BCK+14): Google infers the interests of users only after a certain amount of websites visiting. Another evidence is in (DDJT15), where the authors described that Google took five days of training on a single user to produce personalized news content.

6.1.2 Experiments

For our experiments, we use AdFisher (DTD15). It has been extended for handling Google searches and news searches, to measure personalization in query results, see, e.g., (DDJT15) and applied to novel news search experiments (CHPS16). AdFisher has also been used for statistical evaluations, e.g., to measure how users are exposed to Wikipedia

[6] https://www.google.com/policies/terms/

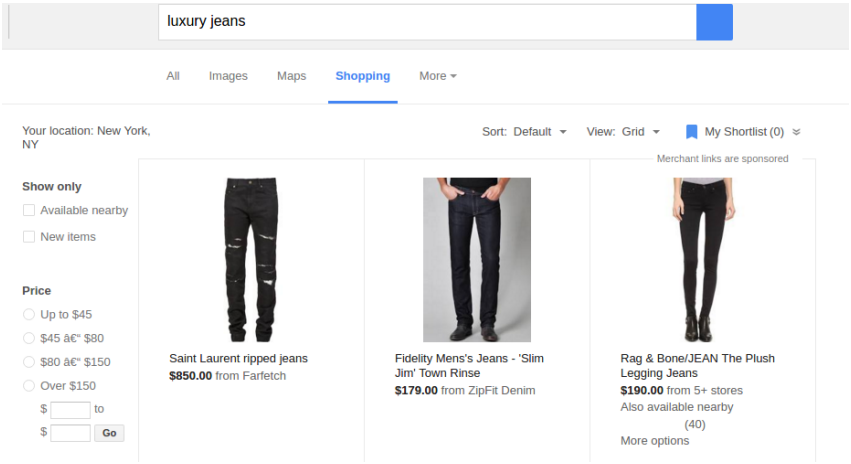

Figure 14: Top 3 results on Google Shopping: affluent, 3rd test block

results, in return to their web searches, see, e.g., (CHP16). The interested reader can refer to (EN16b) for a full survey on tools for measuring and analysing users' interactions with online services (including AdFisher).

In our work, AdFisher runs browser-based experiments that emulate search queries and basic interactions with the search results, e.g., interacting with those search results whose price is above or below the price average on the total of the results, or those results belonging to a list of previously visited websites. Github hosts our extended version[7].

AdFisher interacts with Selenium, a web browser automation tool. Selenium allows to run a unique instance of Firefox creating a fresh profile, with new associated cookies, the so called *Firefox profile*. The Firefox profile that is used is stripped down from what is installed on the machine, to only include the Selenium WebDriver.xpi plugin. Further, we take advantage of a plugin to automatically obtain the Python code for recording actions on web pages (e.g., clicking, typing, etc.), provided by Selenium IDE for Firefox[8]. All the experiments are done on the Fire-

[7] `https://github.com/tienhv/Adfisher_for_GoogleShopping`
[8] `http://www.seleniumhq.org/projects/ide/`

fox web browser version 43.0.4, controlled by Selenium in Python, under XUbuntu 14.04.

To simulate different real users, we browse from different IP addresses, implementing a solution based on SSH tunneling to remote computers. To simulate many computers from one geographic area, we use the VPS services of Digital Ocean[9]. It is well known that, when a user enters a query to Google, the query is unpredictably sent to many distributed servers, to retrieve the results. This could produce noise due to inconsistent data among different servers. To avoid the issue, we query only towards specific Google servers IP addresses, as in (HSMK+13). Finally, we use Jaccard Index and NDCG to evaluate the search results of the test.

6.1.2.1 Settings and results

We have extended the AdFisher functionalities to handle Google Shopping pages and to mimic the behaviours described in Section 6.1.1. We automatically implement the whole experiments for the affluent and control users. The extended AdFisher also stores the query results for further analysis, as the calculation of the NCDG metric.

For the training phase, we emulate two user profiles logged into Google. We consider 80 websites for each type of user profile and we let the profiles visit all of them. Such choice has been driven by (BCK+14), which proved that visiting 50 websites is enough for Google Ads to infer the user interests. When a user visits a website in training step 1, she stays there for around 10 seconds. This, to be sure that the webpage is completed loaded. The website visit time for a user is a random value, however less than 30 seconds (such threshold being estimated following the `alexa.com` statistics). Then, she closes the webpage and stays idle for 5 seconds, before visiting another website. On training step 2, after observing the search results, the user scrolls the webpage to search for the expected products and clicks on the links. Depending from server connections or other external network factors, it could take several seconds (up to more than one minute) to visit one link. The user also scrolls

[9] `https://www.digitalocean.com/`

Table 8: Training phase: An excerpt of websites and keywords

training websites	training keywords	test keywords
nicekicks.com nordstrom.com gucci.com www.toryburch.com, www.luisaviaroma.com	luxury designer shoes, mens boots mens shoes, mens leather boots shoes, luxury shoe, dress mens shoes, mens leather dress shoes	mens dress casual shoes, luxury shoes, dance boots, women trendy boots, luxury jeans, casual jeans

Table 9: NCDG of "luxury shoes" for 8 test sessions

luxury shoes	slot 1	slot 2	slot 3	slot 4	slot 5	slot 6	slot 7	slot 8
Affluent 1	0.45	0.45	0.49	0.40	0.38	0.32	0.37	0.33
Control 1	0.45	0.42	0.48	0.38	0.32	0.30	0.35	0.24
Affluent 2	0.45	0.44	0.47	0.40	0.33	0.28	0.38	0.30
Control 2	0.45	0.39	0.48	0.36	0.32	0.28	0.34	0.25

the webpage like a real user. The time between two searches follows a normal distribution with mean equal to 30 seconds. This has been set to avoid being blocked by Google, due to their counter measures against multiple automatic queries by the same user. Furthermore, each profile has been trained with 15 training keywords, and 3 were the resulting links to be clicked, associated to the top 3 most expensive products.

Table 8 shows an excerpt of the visited websites and the searched keywords, selected with the help of the Google Display Planner.

For the test phase, we still consider the two trained profiles, plus two control ones. All feature the same behaviour, which consists of querying Google Shopping with the same test keywords. Then, we collect the results. Figure 13 and Figure 14 show the first results showed to affluent and control users, for a specific test query. We extract links and prices of all the results to calculate the metrics listed in Section 6.1.1.

The training and the test phases are repeated eight times per user. Each session lasts around 90 minutes.

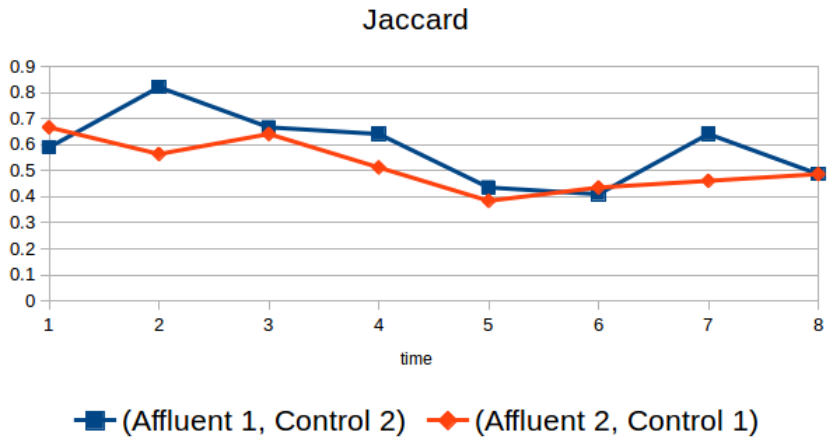

Figure 15: Jaccard Index for “luxury shoes"

Table 9 shows NCDG values for each test session (results are for query “luxury shoes"), for each user.

Figures 15 and 16 plot, resp., the Jaccard index and the Kendall index, for the eight test sessions about “luxury shoes". The blue lines represent are calculated over the results pages of *Affluent 1* and *Control 2* users, while red lines are calculated over the result pages of *Affluent 2* and *Control 1* users. This is to consider two users trained in same ways and connected from two different machines. Jaccard index shows evidence of results customization, while Kendall index says that, most of the times, the results for affluent and control profiles have a level of agreement (featuring a positive values for that index).

Figure 17 plots the values obtained calculating the average NCDG of the two affluent users and the two control users, over the eight test sessions. The test query is “luxury shoes". Average NCDG indicates that even if affluent and control users follow the same pattern of pricing ordering, the former is closer to the ideal list results (where the most expensive products are in the first positions).

Figure 18 shows the average NCDG value for all the test queries (av-

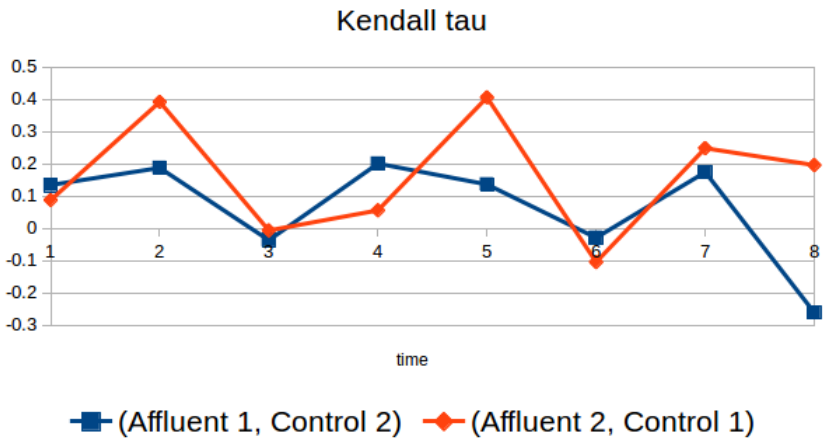

Figure 16: Kendall Index for “luxury shoes"

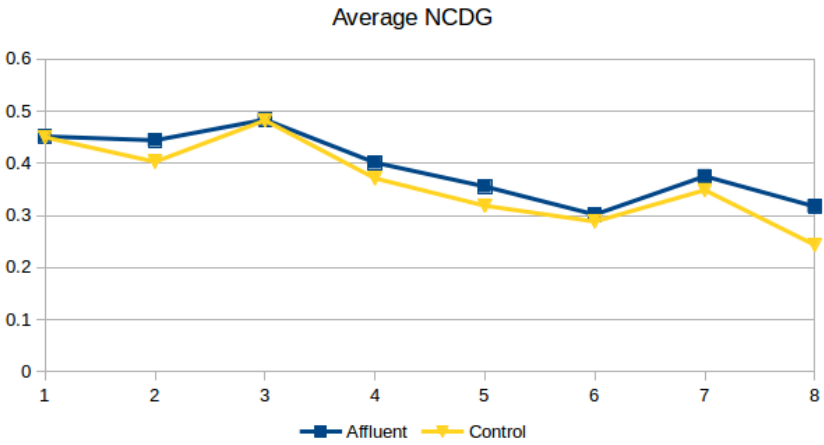

Figure 17: Average NCDG for “luxury shoes"

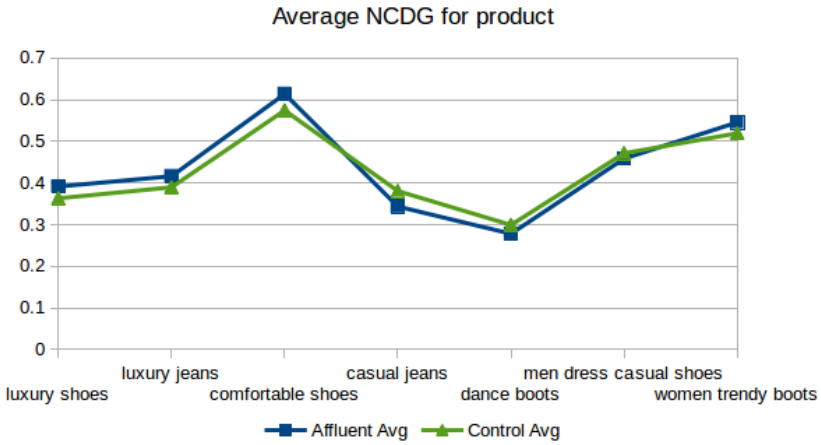

Figure 18: NCDG values averaged over profiles and sessions, per product

eraged both per profile and over the eight sessions). The Figure shows, at a glance, how, in some cases, NCDG values are higher for affluent users than for control. The control users almost always obtain NCDG values lower than the affluent (or comparable in those cases with very similar values for the two kind of profiles). However, we argue that results are also affected by the specific search query over Google Shopping. As an example, Figure 19 shows the NCDG values for the test query “luxury jeans" for affluent 1 and the average of results of the two controls, over the eight sessions. The picture clearly indicates NCDG values that are greater for the affluent user. Instead, “mens dress casual shoes" provides a higher value for the control which is an opposite result with respect to our expectations. While these initial results are promising, there is the need of further evaluation, where more, and more generic keywords, should be tested, at a larger scale.

It is worth noting that, as introduced and motivated at the end of Section 6.1.1, the two affluent users (as well as the two control ones) are identical in terms of behaviour. Further, browser and OS settings are the same, while the IP address from which they browse is different (differ-

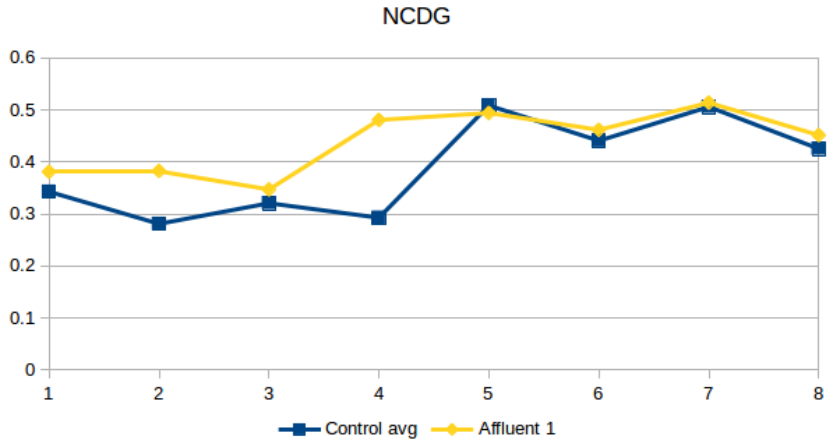

Figure 19: NCDG per session, for “luxury jeans"

ent devices, from the same geographical area). We have not compared directly the two affluents, since their results pages could be different for uncontrollable effects, such as timeouts or network delays (we are indeed emulating different IP addresses). The existence of sources of noise is also the reason why we have chosen to show, for some experiments, the average of the results of the two users.

We compared our results with those in (MGEL12), in which the authors tried to detect price and discrimination on the Internet by accessing online retails websites (i.e., `amazon.com`, `stapple.com`) and `google.com`. The authors found the existence of both price and search discrimination on the Internet. Interestingly, we both found out that users with affluent traits are discriminated against the non-affluent ones. However, we enhance the investigation by using well-known metrics (Jaccard and NDCG) instead of using descriptive statistics about absolute values of prices (min, max, standard deviation) like in (MGEL12).

Despite the use of artificial users, our results are aligned with the conclusions of a similar study – (Ha14). There, the authors hired real users via crowdsourcing of Amazon Mechanic Turks. Those users then

manually searched for products at retail websites like bestbuy.com, expedia.com, cdw.com. Despite of the differences in the experiment settings, we both discover evidence of price personalization.

6.2 A geo-economics study on price steering in Google Shopping

In this part, we consider the prices of the products provided as the result of web search queries. In particular, we measure the differences in prices of the results shown to the users searching from different cities of different countries. Recent work attempted to measure price steering on Amazon, Google Shopping, E-bay and other players, along different dimensions, such as the user location and the user behaviour on the analyzed platforms (Ha14; CHPN16). In our previous work (CHPN16), we presented an initial investigation to analyze at large scale the level of price steering of search results shown to different users on Google Shopping[10].

In this work, we consider price steering in relation to the geographical location of the users. Particularly, we aim at identifying how a possible price steering is related to economical aspects of countries and cities. We consider three countries, characterized by different Gross Domestic Product (GDP) values - US, Philippines, and India. To minimize possible noises caused by different factors, in our experiments we do not consider features such as the users' behavior: search history, purchased history, and clicked links.

Also, instead of providing artificial ad hoc user profiles, we consider the IP addresses of the users. Indeed, it has been shown that, in some countries (e.g., in US), the manual insertion of the user location (in the form, e.g., of a postal code) in the user profile do unfairly affect the price results (VHLY15). Furthermore, locations and postal codes can be easily altered during the profile registration phase (LMA15). Our analysis is launched on Google Shopping. Indeed, Google is well known for tracking users and providing them personalised services (e.g., target advertis-

[10] `https://www.google.com/shopping`

ing)[11].

The experiments collected the list of search results from different users in different countries, namely US, India and Philippines. We limited our investigation considering the products shown in the first page of Google shopping, which correspond to 40 items. Statistics on the results are provided considering the all 40 items, but also the top 3 and the top 10. It's well known indeed the user check only the first results of a web search.

An extensive real data experimental analysis, shows that Google Shopping US, when surfing from US, tends to show to the users products from the highest expensive to the lowest, this with respect to India and Philippines. Looking at different products categories, less costly electronic products are sold in US, but from all the available products, users are going to buy the most expensive products. Looking at the category Apparel & Body Care, in Philippines the most expensive products are shown first, but also overall they sell the most expensive products with respect to the other countries, when looking at the average prices.

We also present a series of tests to evaluate their statistical significance.

The work is organised as follows. Next section presents related work in the area. Section 6.2.1 describes our experimental methodology. Then, Section 6.2.2 provides the description of the experiments and of their outcome.

6.2.1 Methodology

We quantify the differences in product prices, on search results in return to users queries on Google Shopping, US version. In particular, we consider two different scenarios:

- users search from different countries, characterised by different Gross Domestic Product;
- different cities within a single country, where the cities are characterised by different Gross Domestic Product.

[11] https://www.google.com/retail/shopping-campaigns/

All the experiments were conducted by collecting results during July, 2016. The prices we consider are the retail ones displayed by Google Shopping in US dollars (thus, excluding shipping fees).

Emulating users To mimic real users, our synthetic users can browse, scroll pages, stay on a page, and click on links. We use a fully-fledged web browser to get the correct desktop version of the website under investigation. This is because websites could be designed to behave according to user agents, as witnessed by the differences between the mobile and desktop versions of the same website.

There are several measurement tools for interacting with websites (see 3.2). The interested reader can refer to (EN16a) for a complete survey. However, we considered tools able to run browser-based experiments to emulate search queries and basic interactions with the search result, and that could be easily extended with new user behaviours and new evaluation metrics. Both AdFisher[12] (DTD14) and OpenWPM (EN16a) meet these constraints. In past work (CHP16; CHPS16; CHPN16), we experimented issues with AdFisher, related to the recovery of browsers data after crashing. Thus, in this work we choose OpenWPM, since it features a recovery mechanism after crashes (our experiments running, on average, 24 hours). OpenWPM is automatized with Selenium[13] to efficiently create and manage different users with isolated Firefox and Chrome client instances, each of them with their own associated cookies. In all the experiments, the software is run on our local server, but the browser's traffic is redirected to the designated remote servers (i.e., to India), via tunnelling in SOCKS proxies. By operating in this way, we guarantee that all the commands are simultaneously distributed over all the proxies. We use a Mozilla Firefox browser (version 45.0) for the web browsing tasks and we run experiments under Ubuntu 14.04. Also, for each query, we consider the first result page; such page shows 40 products.

[12] `https://github.com/tadatitam/info-flow-experiments`
[13] `http://www.seleniumhq.org/`

Metrics To evaluate the result pages, so that to quantify the differences in prices of the search results, we use a metric widely adopted in the information retrieval research area. The Normalized Discounted Cumulative Gain - **NDCG** - metric measures the similarity between a given list of results and an ideal list of results. In our work, the ideal list of results is a list in which the products are listed from the most expensive to the least expensive one. NDCG, originally introduced in (JK02) in its non-normalised version DCG, has been already adopted in (Ha14) for measuring price steering. For each search result r, there is one gain score g(r), representing its price. In a page with k results, we let $R = [(r_1), (r_2), ..(r_k)]$ and $R' = [(r'_1), (r'_2), ..(r'_k)]$, where r'_1 is the most expensive result and r'_k is the least expensive one. Thus, R' is our defined ideal list of results.

$$DCG(R) = g(r_1) + \sum_{i=2}^{k} (g(r_i)/log_2(i))$$

$$NDCG = DCG(R)/DCG(R').$$

We create R' by first unionising the results returned, for the same query, to all the profiles under investigation, and, then, sorting such results from the most expensive to the least expensive one.

For each query, we calculate the NDCG obtained from the corresponding result page. After a profile successfully executes x queries, we obtain one NDCG vector with x elements for it, where each element corresponds to the NDCG value of a single query. This is the *NDCG vector for a single profile*.

The *NDCG vector of a location* is instead obtained by unionizing all the NDCG vectors of the profiles querying from that location. As an example, with five profiles querying from the same location, there are five NDCG vectors associated to those profiles. The five vectors are then unionized into a single one (with length equal to $5 * x$ elements).

6.2.2 Experiments and Results

We selected three countries, with two countries featuring a significantly different Gross Domestic Product with respect to the third one:

Philippines GDP per capita 7,846.463$[14];

India GDP per capita equal to 6,664.020$[15];

US GDP per capita equal to 57,765.512$[16].

All the values come from the International Monetary Fund, they refer to 2016 and they represent the estimated "Gross domestic product based on purchasing-power-parity (PPP) per capita GDP". For all our experiments, we used English keywords. The three countries have English speaking population. Indeed English is the official language in US, and one of the official languages in India and Philippines. Within these countries, we further selected three relevant cities, from where we emulated users searching over Google Shopping: Manila, New Delhi, and Los Angeles.

For the keywords, we extracted the lists of product categories from `Amazon.com` (due to its richer categorization, when compared with categories from Google Shopping). In details, we considered 130 terms, belonging to various product categories, including body care products, apparel, bags, shoes, car accessories, house accessories (like lightning devices and home appliances), and electronics devices (such as personal audio devices and computer-related products). For each product category, we selected common nouns of products (such as luggages, telephones, books) and we discarded specific brands.

The experiment settings were the following:

- Number of locations: 3 (Los Angeles, New Delhi, Manila);
- Number of keywords: 130;
- Number of browser profiles for each location: 5;
- Web browser: Firefox 45.0;
- OS: Ubuntu 14.04.

[14] `https://goo.gl/bxn4dN`
[15] `https://goo.gl/IxQLpm`
[16] `https://goo.gl/O6Rwtf`

Each browser profile contains its full web history and cookies.

Moreover, it is kept isolated from the other profiles and is deleted right after finishing the search. By design, all browsers work simultaneously and send the same query to Google Shopping.

The user visit lasts 15 seconds, the interval time between one search and one other is a random number between 15 to 30 second due to exponential distribution. For each country, we collected the prices associated with the items in the first page of results, for all the 130 keywords, and for all the profiles searching from that country.

To measure the presence of possible price steering, we consider the NDCG metric, as described in Section 6.2.1. This requires to compute an ideal list composed of all the results shown to the users, sorted from the most to the least expensive. For each country, we combine the first 10 products shown to each of the different users, by storing the 10 most expensive ones. We consider 10 elements since users searching from the same city have been shown very similar results. Being basically the difference only in the order, the list of distinct elements tend to be short. Similarly, we do the same with 3 items only. In fact, it is well known that users usually consider only the first items in their list of results, thus, focusing on 3 items could lead significative results. The two ideal lists are then used to compute, respectively, ndcg@10 and ndcg@3.

Figure 20 represents a screenshot of the results, for 3 profiles from different cities, searching for *pant*.

For each keyword search issued by the users from one country, we compute the average ndcg@3 and ndcg@10. Table 12 reports a snapshot of the results. From `https://goo.gl/u3wpnf` you can access a list of all the ncdg for all the keywords from those we computed the average.

For Philippines, the three top highest ndcg@3 are those related to *briefcase* (0.863), *boot* (0.826) and *blankets* (0.783); For India, *dress* (0.905), *briefcase* (0.88) and *boot* (0.85). For US, *portable DVD player* get 0.859, *desktop* 0.813, and *blankets* 0.8.

Regarding ndcg@10 , for Philippines we obtain, in the first positions, *moisturizer*, *boot* and *ceiling fan*, with ndcg of 0.755, 0.746 and 0.736, respectively. For India: *fuel system* 0.830, *helmet* 0.812 and *dress* 0.778. For

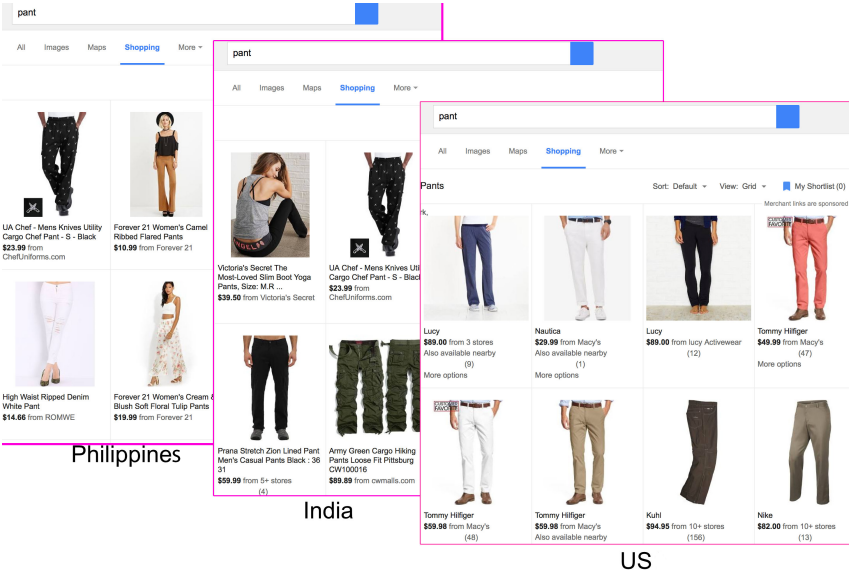

Figure 20: Example of results in return to query *pant*, per city in different countries

US: *portable DVD player* 0.870, *wheel* 0.793, *desktop* 0.783. Each searched product, in each country, features average NDCG values less than 1, meaning that, in all the cases, the search results are not ordered from the most to the least expensive one.

Figure 21 reports mean and standard deviation of the ndcg@3 values of all the profiles, per country. Similarly, Figure 22 is related to ndcg@10. Overall, US reported the highest results in terms of mean, both for ndcg@3 and ndcg@10, while Philippines the lowest ones.

Since keywords can be grouped by categories, we compute the average ndcg per categories. Table reports the results for the categories Apparel (29 keywords), Electronic Devices (31 keywords), and Body Products (10 keywords).

Table 11 shows that category matters. Overall, the three countries have similar average values (last line in the table). Interestingly, for elec-

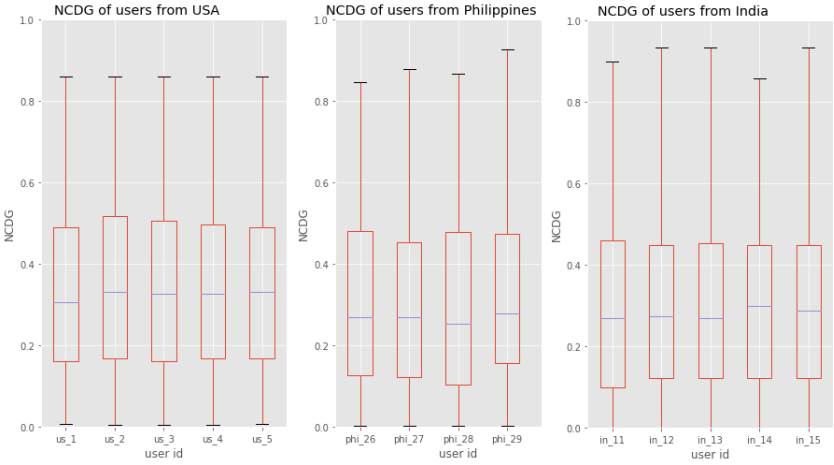

Figure 21: ndcg@3 for countries

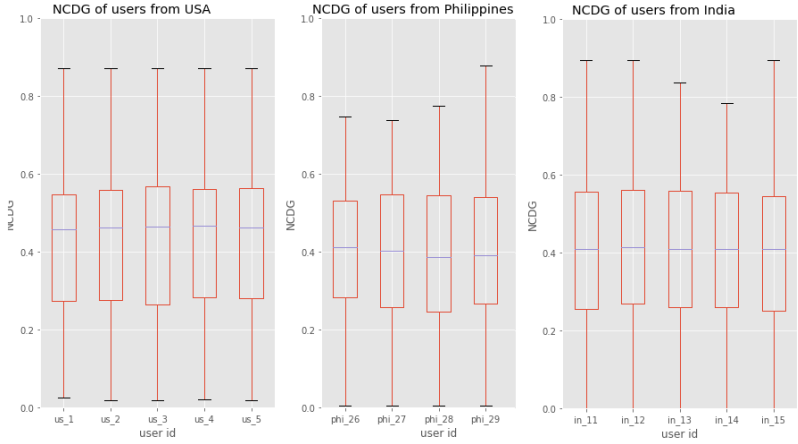

Figure 22: ndcg@10 for countries

Table 10: Average ndcg shown to the five profiles, per country and per sample product

	ndcg@3			ndcg@10		
Product	Philippines	India	US	Philippines	India	US
Boot	0.826	0.85	0.18	0.746	0.755	0.252
Ceiling fan	0.524	0.527	0.245	0.736	0.692	0.475
Desktop	0.114	0.062	0.813	0.219	0.178	0.783
Dress	0.628	0.905	0.282	0.678	0.778	0.371
Helmet	0.592	0.786	0.639	0.639	0.812	0.654
LightScribe	0.138	0.387	0.174	0.27	0.588	0.432
Moisturizer	0.775	0.524	0.48	0.755	0.549	0.523
MP3 Player	0.164	0.111	0.165	0.305	0.289	0.312
Pant	0.228	0.271	0.291	0.305	0.446	0.405
Pocket video camera	0.019	0.079	0.574	0.91	0.229	0.28
Portable CD player	0.129	0.097	0.163	0.342	0.352	0.475
Portable DVD player	0.338	0.318	0.859	0.402	0.426	0.87
Television	0.222	0.282	0.542	0.422	0.461	0.592
Universal remote	0.554	0.161	0.597	0.506	0.194	0596
Wheel	0.177	0.202	0.715	0.281	0.311	0.793

tronic devices US has a much higher value than the other two countries; India features a greater value, compared to US, for apparel; Philippines gets a bit more higher value for body products, with respect to the other two countries.

To have a glue of what really happens, it could be interesting to analyze the average prices of products, as for Table 12 and Table 13. From Table 12, it can be noticed that even if electronic product most expensive are shown first in US, anyay overall they sell products with lower prices respect for example India. As an example, from Table 13, the average price for keyword desktop is 717$, 887$ and 368$, respectively searching from Philippines, India and US. This means that even if electronic product most expensive are shown first in US, anyay overall they sell products with lower prices respect for example India.

Table 11: Average ndcg on the five profiles, per country and per the most relevant category of products

	ndcg@10		
Category	Philippines	India	US
Apparel	0.389	**0.4**	0.31
Electronic Devices	0.195	0.202	**0.4**
Body Products	**0.48**	0.455	0.446
All	0.397	0.406	**0.42**

6.2.3 Statistical significance of the experiments

To give statistical significance to our experiments, we run permutation tests (NH02), following the approach proposed in (EN16a; TDDW15). A permutation test on a set of data (like, in our case, the NDCG values of the location) provides a value, the *p-value*, that represents the probability that a so called null hypothesis is true. We chose this test because it does not require any assumption on input data.

Considering different countries (Sect 6.2.2), we define the first null hypothesis as follows: the obtained prices of all the investigated products, for India and Philippines, are not distinguishable from the obtained prices for US. The second null hypothesis is similar, considering however the distinguished categories of products.

In Table 14, we report the p-values. Looking at the row with p-value equal to 0.050, the test outcome is that the probability of distinguishing which prices are from Philippines or US is 0.95, with a false rate of 0.05. We can notice that all the obtained p-values of Philippines/India vs US are significantly low.

Table 12: Average and median of prices shown to the five profiles, per country and per product category (considering, at most, top 40 results)

Country (City)	**Average price (US $)**	**Median price (US $)**
US (Los Angeles)	221	80
India (New Delhi)	173	74
Philippines (Manila)	221	78
US (LA) Apparel & Body Care	68	47
India (New Delhi) Apparel & Body Care	35	35
Philippines (Manila) Apparel & Body Care	185	39
US (LA) Electronic devices	166	92
India (New Delhi) Electronic devices	283	156
Philippines (Manila) Electronic devices	288	189

Table 13: Average prices shown to the five profiles, per country and per sample products (considering, at most. the top 10 results)

	Average prices (US $)			Median prices (US $)		
Product	Philippines	India	US	Philippines	India	US
boot	136	136	122	90	90	120
ceiling fan	360	340	211	359	359	120
desktop	717	887	368	629	629	399
dress	110	100	86	108	108	92
helmet	299	378	315	400	400	390
LightScribe	67	65	128	65	59	58
moisturizer	22	18	14	14	13	12
pant	36	51	73	22	24	82
Pocket video camera	151	117	73	61	66	149
Portable CD player	50	52	72	25	22	38
Portable DVD player	66	57	64	58	59	60
Television	455	519	625	328	348	499
Universal remote	41	43	95	11	11	60
wheel	262	238	178	175	159	180

Table 14: p-values obtained by running permutation tests over location NDCG values. Different countries.

Location	Category	p-value
Philippines vs US	All	0.050
India vs US	All	0.050
Philippines vs India	All	0.89
Philippines vs US	Accessories	0.669
India vs US	Accessories	0.525
Philippines vs India	Accessories	0.819
Philippines vs US	Apparel	0.136
India vs US	Apparel	0.056
Philippines vs India	Apparel	0.629
Philippines vs US	Body Care	0.394
India vs US	Body Care	0.609
Philippines vs India	Body Care	0.711
Philippines vs US	Electronic devices	0.013
Philippines vs India	Electronic devices	0.90
India vs US	Electronic devices	0.056
Philippines vs US	For House	0.096
India vs US	For House	0.193
Philippines vs India	For House	0.740
Philippines vs US	Others	0.449
India vs US	Others	0.351
Philippines vs India	Others	0.601

Chapter 7

Conclusions

Personalization of online contents can help Internet users to easily find information they are looking for. However, at the same time, it could create the so-called “filter bubble effect". Users could be trapped inside their personalization content bubbles. In this thesis, we measured the level of web search personalization along different contexts, including searching for online news and for shopping. First, after surveying related work in the area, we discussed open problems and proposed our solutions. Then, we carried on a set of experimental studies, aimed at measuring

the personalization level with quantifiable results. In the following paragraphs, we provide additional considerations on each of the three studies considered in this thesis.

Personalization of web search results For this topic, we started from previous work and continued to investigate the level of web search personalization on Google. We proposed a series of experimental settings that are useful and inspiring for understanding how the previous search activities of a set of accounts may influence the results of further searches. Despite the research efforts, the obtained results did not provide evidence of a well defined personalization methodology applied by Google, at least for the kind of queries we have chosen to perform and the number of pages results we have analyzed. However, we consider the methodol-

ogy and the preliminary results a valid initial step for further experiments. Indeed, we found a procedure to build artificial users, and an automatic way to simulate many computers within a single local network and to mimic real web users.

Based on the preliminary results, we further evaluated whether particular domains appear as prevalent in the results of web searches.

In order to do that, we studied the ranks of a popular website on Google search results. Our study measured the position of the Wikipedia links in the search results returned by Google Italia for a set of keywords. The obtained results provide evidence that Wikipedia is dominant, as for the Google Italia search results ranking: there is a link to a Wikipedia page within the first five search results in more than 78% of times. A closer look to the search results shows that keywords related to "Mostre d'Arte" category always have an associated Wikipedia page in the top three results, while those related to "News" are less than 10%. We focused on quantifying the phenomenon, without investigating the semantics motivation behind the difference rankings for categories. We leave this for future work.

There could be multiple possible explanations for the presence of Wikipedia links in the first positions of search results. For example, Wikipedia is the leader of online encyclopedia market. It accounts for 97% of the market share[1]. Hence, it is almost likely that for any general term, there is one entry at Wikipedia to be display first. Its web page structures harmonize with search engine ranking algorithms, such structures present rich internal links, and header tags[2]. However, we did not explore in depth the motivations, and we give here simple conjectures.

In fact, the study is not intended to be directly related to personalization topics. It rings a bell about power of web search providers. They are not only possible to decide which links are displayed but also could decide a de facto standard of web page structure. When a website has similar structure as they suggested, it would get a higher priority to be

[1] https://econsultancy.com/blog/3185-wikipedia-has-97-of-the-encyclopedia-market-online

[2] https://www.shoutmeloud.com/wikipedia-seo-strategies.html

indexed and displayed[3].

Personalization of online news To further investigate the personalization phenomena, we considered personalization in Google News. We measured the level of personalization (claimed by Google itself) in different contexts: logged users, expected personalization (looking at the Suggested for You - SGY - sections) and unexpected personalization (looking at Google News home). From the study, it has been shown that it takes time to produce personalized content. Specifically, in our experiments, it takes roughly four days.

We differ from related work in the area, mainly because we observed the results obtained in return to specific user queries. Again, at least for the chosen experiments configuration, we did not observe high differences in the results obtained by a trained user and by a fresh one. However, we made a further step: since the SGY section is not automatically shown to non logged users, nor to freshly logged ones, we performed 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 for 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). Again, this work confirms the existence of personalization practices on the Internet. The personalization is realized by exploiting the users data, included the ones about the users activities and behaviors.

Personalization over e-commerce websites In this study, we designed and implemented a methodology to train and test *user behaviours* on Google Shopping, for evaluating a potential price steering, based on the *willingness to pay* attitude of the users. We have analysed the results list of two types of users - with different online behaviors - so called affluent and control users. The results lists were obtained over eight test sessions, one at the end of each training session. The outcome of the experiments was

[3] `https://support.google.com/webmasters/answer/35769?hl=en`

that, for most of the test queries, the result list of the affluent user is biased towards more expensive products than the one of the control user. Indeed, the experiments results pave the way for further investigation. Even if carried on over Google Shopping, we highlight that our experimental approach is general enough to be applicable to other e-commerce websites, like, e.g., *Amazon.com* and *eBay.com*.

To better quantify the presence of price steering applied to search results shown to the user on e-commerce websites, we investigated the impact of searching from specific locations on the order of price results.

In particular, we simulated the search of common products over Google Shopping. Differently from previous work in the area, we considered geographical locations based on one indicator of the economic performance of the locations, i.e., the Gross Domestic Product. As locations, we considered cities in three different countries, and a set of cities located in US. From a qualitative point of view, considering the price values of the first 40 results returned by Google Shopping, it is possible to notice differences both in average and median prices, per country.

To better investigate the differences, we considered the NDCG metric, which, in our context, measures the differences between the list of price results, as obtained by considering the results of our queries, and an ideal list of results, defined as an ordered list, where the products with the highest prices are first shown to the user. The results of our investigations are that the geographical location does have impact on the results given by Google Shopping.

In fact, regarding the analyses carried out for cities in different countries, the permutation tests to assess the statistical significance of the experiments gave a positive answer: aiming at comparing cities in countries characterised by really different GDP, the obtained prices of all the investigated products, and for distinct categories of products, from India, are distinguishable from those obtained from US. The same holds considering Philippines and US.

From the aforementioned remarks, we confirm the existence of different forms of personalization practices on the Internet. The personalized content is provided by the web search engine, news search engine and

shopping search engine.

From the aforementioned remarks, we confirm the existence of some forms of personalization practices on the Internet. The personalized contents are provided by web search engines, news search engine and even shopping search engine.

Also, our studies highlight the following issues, which calls for further investigations.

Personalization on social network platforms. Twitter, Facebook and VKontact have undeniable advantages over a web search engine. They not only have data from users but also data from the users' connections and possibly by real-life relationship. These data can be used by these platforms to provide users with more relevant information or better personalization contents.

Larger scale experiments Although we perform several experiments, the collected data are relatively small compared to the real world data. The study could provide more accurate results if experiments would be larger. For example, one could consider more user profiles, more accounts, more geographical locations and a longer training time. This extension could lead to other advantages, such as utilizing different analyses methods (i.e., machine learning analysis is likely better with more data).

Considering data from real Internet users We have considered synthetic data, which may be not of the same quality of data by real users. The latter could significantly improve the validity of the results. One possibility would be adopting crowd-sourcing methods by, e.g., distributing a web browser plug-in to users around the world.

New analysis methods Although the used metrics (i.e., Jaccard Index) are still useful, we could get other interesting results by resorting to other methods based on machine learning. For example unsupervised learning (i.e., deep learning) could introduce a new way to detect the presence of personalization.

References

[AEE+14] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, pages 674–689, New York, New York, USA, nov 2014. ACM Press. 27

[Ale16] Alexa. The top 500 sites on the web. `http://www.alexa.com/topsites`, 2016. [Online; accessed 5-Nov-2016]. 16

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005. 15, 16

[BCK+14] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 597–608. ACM, 2014. 64, 66

[BH09] Matthew Kulick Bryan Horling. *Personalized Search for everyone*, 2009. `http://googleblog.blogspot.it/2009/12/personalized-search-for-everyone.html`. 27

[BIMHL06] Judit Bar-Ilan, Mazlita Mat-Hassan, and Mark Levene. Methods for comparing rankings of search engine results. *Comput. Netw.*, 50(10):1448–1463, July 2006. 31

[BR11] Mikhail Bilenko and Matthew Richardson. Predictive client-side profiles for personalized advertising. In *Proceedings of the ACM SIGKDD'11*. ACM, 2011. 16, 18

[CG07] Edward Cutrell and Zhiwei Guan. What are you looking for?: An eye-tracking study of information usage in web search. In *Human*

Factors in Computing Systems, CHI '07, pages 407–416. ACM, 2007. 42

[CHP16] Vittoria Cozza, Van Tien Hoang, and Marinella 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. xv, 54, 65, 74

[CHPD16] Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi, and Rocco De Nicola. Online user behavioural modeling with applications to price steering. In *Proceedings of FinRec 2016: 2nd International Workshop on Personalization and Recommender Systems in Financial Services*, volume 1606 of *CEUR Workshop online Proceedings*, pages 16–21. V. Cozza, V. T. Hoang, M. Petrocchi, R. De Nicola. In FinRec, 2016,, June 2016. online: http://ceur-ws.org/Vol-1606/. xv, 24

[CHPN16] Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi, and Rocco De Nicola. Online user behavioural modeling with applications to price steering. In *FINREC 2016*. CEUR Workshop Proceedings, 2016. 54, 72, 74

[CHPS16] Vittoria Cozza Cozza, Van Tien Hoang, Marinella Petrocchi, and Angelo Spognardi. Experimental measures of news personalization in google news. In *SoWeMine 2016: 2nd International Workshop on Mining the Social Web*. Springer, 2016. xv, 24, 64, 74

[CIS15] CISCO. The Zettabyte Era—Trends and Analysis. `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html`, 2015. [Online; accessed 30-Oct-2016]. 1

[CMR+15] Juan Miguel Carrascosa, Jakub Mikians, Rubén Cuevas Rumín, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: *measuring online behavioural advertising*. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT 2015, Heidelberg, Germany, December 1-4, 2015*, pages 13:1–13:13, 2015. 23

[DBG15] Tawanna R. Dillahunt, Christopher A. Brooks, and Samarth Gulati. Detecting and visualizing filter bubbles in google and bing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1851–1856, New York, NY, USA, 2015. ACM. 31

[DDJT15] Amit Datta, Anupam Datta, Suman Jana, and Michael Carl Tschantz. Poster: Information flow experiments to study news personalization. In *Computer Security Foundations Symposium (CSF), 2015 IEEE 28th*. IEEE, 2015. 3, 24, 27, 47, 48, 49, 50, 53, 54, 59, 64

[Dev16] Android Developer. Location Strategies. `https://developer.android.com/guide/topics/location/strategies.html`, 2016. [Online; accessed 5-Nov-2016]. 20

[D'O15] Jillian D'Onfro. The Clever Way Amazon Gets Away With Not Always Offering The Lowest Prices . `http://uk.businessinsider.com/how-amazon-adjusts-its-prices-2015-1`, 2015. [Online; accessed 1-Oct-2016]. 6

[DTD14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014. 24, 31, 43, 74

[DTD15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015. 3, 7, 8, 21, 27, 30, 64

[E+14] Steven Englehardt et al. Web privacy measurement: Scientific principles, engineering platform, and new results. *Manuscript posted at `http://randomwalker.info/publications/WebPrivacyMeasurement.pdf`*, 2014. 8, 20, 62

[EEZ+15] Steven Englehardt, Christian Eubank, Peter Zimmerman, Dillon Reisman, and Arvind Narayanan. OpenWPM: An Automated Platform for Web Privacy Measurement. Manuscript, 2015. 8, 25, 27

[EN16a] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. [Technical Report], May 2016. 31, 74, 81

[EN16b] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. [Technical Report], May 2016. 53, 65

[EN16c] Steven Englehardt and Arvind Narayanan. OpenWPM. `https://github.com/citp/OpenWPM`, 2016. [Online; accessed 5-Nov-2016]. 30

[Fac16a] Facebook. Here's Facebook's guide to how its 'Trending Topics' tool works. `http://www.recode.net/2016/5/12/11665360/report-leaked-documents-facebook-trending-headlines`, 2016. [Online; accessed 3-Sept-2016]. 2

[Fac16b] Facebook. How Facebook News Feed Works . https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/, 2016. [Online; accessed 30-Oct-2016]. 2

[FDA14] Anisha T. J. Fernando, Jia Tina Du, and Helen Ashman. Personalisation of Web Search: Exploring Search Query Parameters and User Information Privacy Implications-The Case of Google. *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (PIR 2014)*, pages 31–36, 2014. 8, 20

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher 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. 52

[fRR16] Center for Retail Research. Online Retailing: Britain, Europe, US and Canada 2016 . http://www.retailresearch.org/onlineretailing.php, 2016. [Online; accessed 3-Sept-2016]. 6

[GA05a] Daniela Godoy and Analia Amandi. User profiling for web page filtering. *IEEE Internet computing*, 9(4):56–64, 2005. 14

[GA05b] Daniela Godoy and Analia Amandi. User profiling in personal information agents: a survey. *The Knowledge Engineering Review*, 20(4):329–361, 2005. 14, 15, 16

[GCF10] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. *Internet Measurement Conference*, pages 81–87, 2010. 8, 22

[Goo16] Inc Google. Webmaster Guidelines. https://support.google.com/webmasters/answer/35769, 2016. [Online; accessed 30-Oct-2016]. 3

[Gre15] Paul Grey. (2015) How Many Products Does Amazon Sell? https://export-x.com/2015/12/11/how-many-products-does-amazon-sell-2015/, 2015. [Online; accessed 30-Oct-2016]. 1

[GUH16] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13, 2016. 3

[Ha14] Aniko Hannak and al. Measuring price discrimination and steering on e-commerce web sites. In *IMC*, pages 305–318. ACM, 2014. 6, 8, 22, 23, 62, 71, 72, 75

[HC08] J. Hu and P. Chan. Personalized Web Search by Using Learned User Profiles in Re-ranking. In *Workshop on Knowledge Discovery on the Web, KDD conf*, pages 84–97, 2008. 16, 17

[HG12] Chris Jay Hoofnagle and Nathan Good. Web privacy census. *Available at SSRN 2460547*, 2012. 27

[Hid16] Ariya Hidayat. PhantomJS. `http://phantomjs.org`, 2016. [Online; accessed 30-Oct-2016]. 7, 19, 20, 22, 30

[Hin09] M. Hindman. *The Mith of Digital Democracy*. Princeton University Press, Princeton, NJ, USA, 2009. 5

[HJK12] T.H. Haveliwala, G.M. Jeh, and S.D. Kamvar. Targeted advertisements based on user profiles and page profile, November 27 2012. US Patent 8,321,278. 49, 61

[HL09] Nadine Höchstötter and Dirk Lewandowski. What users see - structures in search engine results pages. *Inf. Sci.*, 179(12):1796–1812, 2009. 42

[Hod15] Richard Hodson. Wikipedians reach out to academics. *Nature News*, Sept. 2015. 42

[HSL+14] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 305–318, New York, NY, USA, 2014. ACM. 24

[HSMK+13] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 527–538, New York, NY, USA, 2013. ACM. 3, 4, 5, 7, 18, 21, 24, 31, 35, 37, 41, 43, 66

[HST+15] Van Tien Hoang, Angelo Spognardi, Francesco Tiezzi, Marinella Petrocchi, and Rocco De Nicola. Domain-specific queries and web search personalization: some investigations. In *Proceedings 11th International Workshop on Automated Specification and Verification of Web Systems, WWV 2015, Oslo, Norway, 23rd June 2015.*, pages 51–58, 2015. xv, 24

[Int16] Internetlivestats.com. 1 Second - Internet Live Stats. `http://www.internetlivestats.com/one-second/`, 2016. [Online; accessed 30-Oct-2016]. 1

[JK02] Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. 75

[KAH12] Abhijith Kashyap, Reza Amini, and Vagelis Hristidis. SonetRank: leveraging social networks to personalize search. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, pages 2045–2049, 2012. 31

[KAS05] Patricia Kearney, Sarabjot Singh Anand, and Mary Shapcott. Employing a domain ontology to gain insights into user behaviour. In *Working Notes of the IJCAI Workshop on Intelligent Techniques for Web Personalization*, pages 25–32, 2005. 13

[KL94] Kevin Knight and Steve K Luk. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778, 1994. 13

[KP06] Mehdi Khosrow-Pour. *Encyclopedia of e-commerce, e-government, and mobile commerce*. IGI Global, 2006. 16

[KS00] Tsvi Kuflik and Peretz Shoval. Generation of user profiles for information filtering?research agenda (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 313–315. ACM, 2000. 14

[KSHL+15] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, IMC '15, pages 121–127, New York, NY, USA, 2015. ACM. 3, 8, 19, 23

[Law05] S. Lawrence. Personalization of web search, Mar 2005. US Patent App. 10/676,711. 5

[LDL+14] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. XRay: Enhancing the Web's Transparency with Differential Correlation. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 49–64, San Diego, CA, August 2014. USENIX Association. 3, 7, 19

[LDP10] Jiahui Liu, Peter Dolan, and Elin Rønby 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. 48, 53

[LMA15] J. Larson, S. Mattu, and J. Angwin. Unintended Consequences of Geographic Targeting. *Technology Science*, 2015. 72

[LNSU08] Heng Luo, Changyong Niu, Ruimin Shen, and Carsten Ullrich. A collaborative filtering framework based on both local user similarity and global user similarity. *Machine Learning*, 72(3):231–245, 2008. 15

[LSS+15] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 554–566. ACM, 2015. 3, 4, 8, 20

[LYM02] Fang Liu, Clement Yu, and Weiyi Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 558–565. ACM, 2002. 13

[Mat12] D Mattioli. On Orbitz, Mac users steered to pricier hotels. *Wall Street Journal*, 2012. 22, 24

[MGEL12] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting price and search discrimination on the Internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, HotNets-XI, pages 79–84. ACM, 2012. 6, 21, 23, 61, 71

[MGEL13] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Crowd-assisted search for price discrimination in e-commerce: First results. *Proc. of CoNEXT*, pages 1–6, 2013. 6, 22

[MGH98] M Montebello, WA Gray, and S Hurley. A personal evolvable advisor for www knowledge-based systems. *based Information*, page 59, 1998. 13

[MM12] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 413–427, Washington, DC, USA, 2012. IEEE Computer Society. 31

[Mor11] Evgeny Morozov. *The Net Delusion: The Dark Side of Internet Freedom*. Perseus Books, Cambridge, MA, USA, 2011. 5

[Mou96] Alexandros Moukas. Information discovery and filtering using a multiagent evolving ecosystem. 1996. 13

[MR11] Nicolaas Matthijs and Filip Radlinski. Personalizing web search using long term browsing history. *Wsdm 2011*, page 25, 2011. 16, 17, 34

[MS13] Anirban Majumder and Nisheeth Shrivastava. Know your personalization: Learning topic level personalization in online services. *22nd international conference on World Wide Web*, pages 873–884, 2013. 8, 20

[MSDR04] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004. 13

[MSS14] Kamlesh Makvana, Pinal Shah, and Parth Shah. A novel approach to personalize web search through user profiling and query reformulation. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pages 1–10. IEEE, September 2014. 16, 17, 34

[NH02] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002. 81

[NOD14] Ashish Nanda, Rohit Omanwar, and Bharat Deshpande. Implicitly Learning a User Interest Profile for Personalization of Web Search Using Collaborative Filtering. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 54–62. IEEE, August 2014. 16, 34

[Par11a] Eli Pariser. Beware online filter bubbles. `http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles`, 2011. 20

[Par11b] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011. 5, 6, 33

[PCG03] Danny Poo, Brian Chng, and Jie-Mein Goh. A hybrid approach for user profiling. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 9–pp. IEEE, 2003. 13

[PG99] Alexander Pretschner and Susan Gauch. Ontology based personalized search. In *Tools with artificial intelligence, 1999. proceedings. 11th ieee international conference on*, pages 391–398. IEEE, 1999. 13

[Pop14] Andrei Popescu. Geolocation api specification - w3c editors draft 11 july 2014. `https://dev.w3.org/geo/api/spec-source.html`, 2014. 20

[Pro16] Selenium Projects. Selenium WebDriver. `http://www.seleniumhq.org/`, 2016. [Online; accessed 30-Oct-2016]. 21

[Ref13] Referralcandy.com. How Many Online Stores are there in the U.S. (2013 edition)? `http://www.referralcandy.com/blog/how-many-online-stores-are-there-in-the-us-2013-edition`, 2013. [Online; accessed 3-Sept-2016]. 6

[SC12] Sam Silverwood-Cope. Wikipedia: Page one of Google UK for 99% of searches. *pi-datametrics.com Blog*, 2012. 42, 46

[SD97] Scott Stewart and John Davies. User profiling techniques: a critical review. In *Proceedings of the 19th Annual BCS-IRSG conference on Information Retrieval Research*, pages 10–10. British Computer Society, 1997. 14, 15

[SDA+16] Oleksii Starov, Johannes Dahse, Syed Sharique Ahmad, Thorsten Holz, and Nick Nikiforakis. No honor among thieves: A large-scale analysis of malicious web shells. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1021–1032, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. 27

[SGC13] Amit Sharma, Mevlana Gemici, and Dan Cosley. Friends, Strangers, and the Value of Ego Networks for Recommendation. *ICWSM*, April 2013. 31

[She16] Sergey Shekyan. Detecting PhantomJS Based Visitors. `http://engineering.shapesecurity.com/2015/01/detecting-phantomjs-based-visitors.html`, 2016. [Online; accessed 30-Oct-2016]. 20

[SHM+05] Lubomira Stoilova, Todd Holloway, Ben Markines, Ana G Maguitman, and Filippo Menczer. Givealink: mining a semantic network of bookmarks for web search and recommendation. In *Proceedings of the 3rd international workshop on Link discovery*, pages 66–73. ACM, 2005. 13

[SK09] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009. 15, 16

[SKK00] Ingo Schwab, Alfred Kobsa, and Ivan Koychev. Learning about users from observation. In *AAAI 2000 Spring Symposium: Adaptive User Interface*, pages 241–247, 2000. 14

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001. 15

[Sta16] Statista. Statistics and facts about global e-commerce . `http://www.statista.com/topics/871/online-shopping`, 2016. [Online; accessed 1-Oct-2016]. 6

[TDDW15] Michael Carl Tschantz, Amit Datta, Anupam Datta, and Jeannette M Wing. A methodology for information flow experiments. In *2015 IEEE 28th Computer Security Foundations Symposium*, pages 554–568. IEEE, 2015. 8, 21, 25, 54, 62, 81

[UF98] Lyle H Ungar and Dean P Foster. Clustering methods for collaborative filtering. In *AAAI workshop on recommendation systems*, volume 1, pages 114–129, 1998. 15

[US13] Yury Ustinovskiy and Pavel Serdyukov. Personalization of web-search using short-term browsing context. *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pages 1979–1988, 2013. 16, 17

[VdODS13] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013. 14

[Ven13] VentureBeat. How Google searches 30 trillion web pages, 100 billion times a month. `http://venturebeat.com/2013/03/01/how-google-searches-30-trillion-web-pages-100-billion-times-a-month/`, 2013. [Online; accessed 30-Oct-2016]. 1

[VHLY15] Keyon Vafa, Christian Haigh, Alvin Leung, and Noah Yonack. Price discrimination in the Princeton review's online SAT tutoring service. https://techscience.org/a/2015090102/, 2015. Online; accessed Apr-21-2017. 72

[VNBJ14] Thomas Vissers, Nick Nikiforakis, Nataliia Bielova, and Wouter Joosen. Crying Wolf ? On the Price Discrimination of Online Airline Tickets. *7th Workshop on Hot Topics in Privacy Enhancing Technologies*, page 12, 2014. 22

[WBD10] Ryen W. White, Paul N. Bennett, and Susan T. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10. ACM, 2010. 16, 17

[WT12] Ce Wills and C Tatar. Understanding What They Do with What They Know (Short Paper). *Proceedings of the 2012 ACM workshop on Privacy in the Electronic Society*, pages 13–18, 2012. 3, 7, 18

[XMD+14] Xinyu Xing, Wei Meng, Dan Doozan, Nick Feamster, Wenke Lee, and Alex C. Snoeren. Exposing inconsistent web search results with bobble. In *Proceedings of the 15th International Conference on Passive and Active Measurement - Volume 8362*, PAM 2014, pages 131–140, New York, NY, USA, 2014. Springer-Verlag New York, Inc. 3, 4, 7, 19, 31

[YHH14] Zhen Yue, Shuguang Han, and Daqing He. Modeling search processes using hidden states in collaborative exploratory web search. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*, CSCW '14, pages 820–830, New York, NY, USA, 2014. ACM. 16, 17

[YL10] Jie Yu and Fangfang Liu. Mining user context based on interactive computing for personalized web search. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, volume 2, pages V2–209. IEEE, 2010. 16, 17

[ZRJT10] Markus Zanker, Francesco Ricci, Dietmar Jannach, and Loren 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. 50