# Location-based Web Personalisation

Stergios Morakis: s1047752
Radboud University
Nijmegen, Netherlands

Gabriela Rivas: s1047701
Radboud University
Nijmegen, Netherlands

Shaohan Jia: s1040904
Radboud University
Nijmegen, Netherlands

## ABSTRACT

Over the years, state-of-the-art web search engines have evolved at a tremendous level, with queries offering billions of results to individuals, ordered in a personalised manner. A non-exhaustive list of ingredients for such convoluted ranking systems include the user's location, past click-behaviour and search history when determining the appropriate ranking of the retrieved information. Advanced information retrieval systems on the web determine such rankings based on a variety of autonomous decisions, i.e. whether a user would be interested in a specific web page. This has sparked a heated debate over the complications of filter bubbles, a term used to describe a state of intellectual isolation [5]. This paper focuses on analysing the extent to which the Google search engine personalises results based on a user's location within the European Union (EU), considering the existence of various regulations and directives followed by each Member State. We investigate the concept of personalisation as the extent of the disparities in results obtained from the Google search engine and the impact of carry-over effects originating from previously posed queries, based on three different countries.

## KEYWORDS

Search Personalisation, Filter Bubbles, Location Tracking

## 1 PROJECT DESCRIPTION

### 1.1 Introduction

Initially, the World Wide Web (WWW) was a different environment. Lacking accounts and tracking tools, the first widely used query systems relied on keywords. These engines offered identical collections of pages to users posing the same search query, regardless of their interests or past search behaviour. As technology advanced and access to the internet became more widespread, the amount of online resources exponentially grew, along with the desire for more efficient and accurate procedures of information retrieval. In the late 90's, the introduction of the Page Rank system for web search revolutionised the process of gathering data, by applying a probabilistic model based on the number and quality of the pages' links to other websites [1]. Search engine businesses had to respond effectively to the unique needs of the public and thus, the concept of adjusting web results to the users' specific needs made its first appearance.

At this point, data tracking came in a wide range of varieties. Search history, time spent on a particular website, as well as the frequency of a web page's visits, would all play a major role in an attempt to personalise the individuals' information seeking [2], as they would assist the systems in profiling their users. User profiling entails features such as the user's background, hobbies and location. Thus, we assume a level of responsibility to fall on the search engine's part. Big companies, such as Google and Bing, should be aware of their products' misguidance, when falling under a non-trivial ethical spectrum, such as threat-related [3] and self-harm [6] search patters. The reasons stated above, provoked our interest in investigating the noticeable side effects of information segregation, also known as filter bubbles, in today's society.

This paper focuses on the concept of personalisation based on location tracking for query systems. The concept of measuring location-based web customisation stems from the study by Chloe Kliman-Silver et al. [7]. Our project devolves as a follow-up to this experiment, where we focus our attention on web customisation amongst different countries instead of a single nation and the possibility of the appearance of national filter bubbles [8]. Our team is primarily interested in studying the degree of carry-over effects, inspired by Hannák et al. [4], which made use of Jaccard measure and Kendall Tau [10] to determine the similarities amongst lists of results obtained between queries.

We will motivate our review by reporting our findings in section 3. The project was developed by recording and analysing the retrieved ranked lists corresponding to the first three pages of Google's search engine, from three European Member States. Prior to conducting the described research, queries were designated for a set of 20 topics. The time allocated between queries to regulate carry-over was determined based on the level of dissimilarity between secondary and primary queries. The analysis resulted in the discovery of significant differences between specific search topics, along with great disparities in the levels of carry over effects between countries.

### 1.2 Aim of the Project

This paper dives into the subject of personalised search based on the user's current location and aims to investigate the extent to which it affects the top rankings of the generated results of a popular, state-of-the-art web search engines, in this case Google. In order to pinpoint the user's location, Google makes use of countless techniques. A non-exhaustive list of such techniques include the user's search history and Internet Protocol (IP) address. When carrying out this experiment, a system was developed by utilizing the flexibility of Python programming language and Selenium, an automation framework specialized in simulating real user interactions. Furthermore, our team made use of a Virtual Private Network (VPN) in order to disguise our location between our testing phases. By configuring our current location to the desired ones, we were able to gather a significant amount of data, required to conduct our analysis.

Moreover, our project investigates the carry-over effects due to personalisation between three member states of the European Union: United Kingdom, Netherlands and France. Our initial motivation for this research stems from the question on whether location tracking could cause filter bubbles as the only personalisation feature in search engines. In addition, we were interested in the effect

of the European law's implementation on Google's search results, for future work. Due to this reason, our research focuses on the Netherlands and France, both based on a civil law system, and UK, which is based on a common law system. The two civil law system based countries were chosen mainly due to their different interpretations of law, with the former taking a direct and strict approach while the latter holding a pragmatic point of view. The paper includes a short but sufficient experiment in order to detect the optimal time interval between consecutive queries, for which the carry-over effects decay at a slower pace.

## 2 METHODOLOGY

### 2.1 Preparation

This section gives a detailed description on the factors taken into consideration when designing the experimental procedure, outlining important issues that emerged during the planning phase and limitations of our approach.

*2.1.1 Topics and Queries.* Search prompts where constructed from a list of 20 topics, designed to incorporate a variety of themes expected to influence the difference in retrieved results up to a certain amount. These themes were decided after experiencing rough initial impacts on personalisation with test data with localisation set to the Netherlands. Subjects related to Relationships, social behaviours, history and childcare were found to be amongst those more strongly influencing the ranks of the retrieved links compared to queries in the themes of education, religion, environmental protection and health. Each topic is comprised of two queries and two sub-queries, the latter aiming to add noise to the former's results.

At first, when designing the experiment, we attempted to initialize six different Google accounts, two for each location, however we were forced into a different approach as verification procedures from Google have been an unforeseen obstacle. This led to the investigation and use of private browsing sessions in the Mozilla browser, as incognito mode is configured to store information related to search history during the lifespan of the browser navigation. This approach opened up the possibility for acquiring a larger load of information, leading to an extensive corpus of queries for a broader view into smaller changes in ranking.

*2.1.2 Time Interval Setup.* Upon constructing our first test queries we became aware of the need to denote a suitable time frame between the search history establishing sub-queries and the main prompts. Google data centers provide constant updating of online data but sometimes, it may occur that the latest information may take a longer time to show up on search systems. This signifies that queries searched a minute after a sub-query is prompted may return slightly different results than a query searched after a longer delay. Timing can greatly impact the retrieved ranking of results, depending on the subject of the query, i.e. pages on current news stories or weather forecasts. As a first level of precaution, our experiment's queries were carefully designed so as to avoid topics prone to changes within short amounts of time.

To determine an appropriate delay duration we carried out an automated retrieval of results by initialising profiles (incognito sessions) on the Google search engine, consisting of two consecutive queries with various time intervals in-between. We experimented

with waiting times of 1, 3, 5, 10, 15, 20, 25 and 30 minutes. After retrieving the ranked results for each time delay, our team compared them pairwise by computing the Jaccard distance scores [9] of the subsequent queries, affected by the carry over effects. Upon comparison, the scores after the 20 minute delay mark reached a fair amount of consistency. A waiting time interval of 30 minutes was established as to allow for some further small updates and thus, a stable page rank.

### 2.2 Our Approach

Two rounds for each country were established for each topic, using two automatic processes namely user 1, for establishing a search history, and user 2, a blank profile. During every phase, the algorithm prompts a search for the process user 1, who poses two sub-queries, then proceeds to wait for a time interval of 30 minutes. User 1 then searches for a third query within the same topic, considered a main query, at the same time as user 2. The results for both are then gathered from the first 3 pages of Google results. This procedure is carried out in an equal manner for all topic categories and all three countries. The page links retrieved were stored in a csv format, ordered as they were displayed in the page rank.

The following are sample queries relating to a particular topic:

Topic: Global Warming. Queries(Q):

- Q1 (User 1 sub-query): Carbon dioxide
- Q2 (User 1 and 2 sub-queries) : Greenhouse effect
- Q3 (User 1 and 2 main queries) : Glacial Melting
- Q4 (User 2 main query): Antarctic

*2.2.1 Evaluation.* The first stage of the evaluation was focused on the degree of personalization for each topic, for the three countries. The weighted lists consisted in results retrieved by using a profile initialised with a search history and those obtained with a blank profile for each topic. The difference in customisation was determined by using the Kendall Tau metric, which calculates the correlation strength between two lists of results. Kendall scores range from values of -1 to 1, a score of 1 signifying an equal pair of page lists, ranked in the same order. Positive values indicate similarity in the ranks of lists whilst negative values are a consequence of documents ranked in opposite directions. The second stage of the evaluation consisted in creating a visualization for the differences carry over effects by comparing pairs of ranked lists for every combination of two countries and all topics.

*2.2.2 Limitations.* It is worth noting that due to the nature of this research, we attempted to keep the user's location as the sole changing variable and thus, the generated test queries were all in the English language. This approach may introduce inaccuracies in the variation of information retrieved, as a native speaker from the UK should not be compared with a French user searching for information in English.

Due to the long run-time of the software, a non-parallel algorithm was implemented in a way that allowed for all data to be gathered on several executions and thus, another limitation should be stated. The software used to retrieve the data set did not operate on specific hours of the day. Additionally, a single computer was

responsible for gathering the data, increasing the odds of the pages' ranking getting reformed.

## 3 EVALUATION

### 3.1 Degrees of Personalisation

The graphical representations in this section correspond to the comparison of the ranked lists retrieved for two users, one of which has already issued a query related to the same topic.
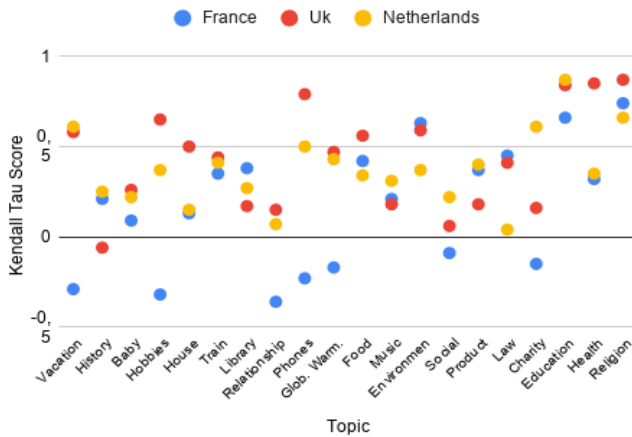


Figure 1: Degree of personalisation per topic

The above plot illustrates the degrees of personalisation for each topic, denoted with a Kendall Tau similarity score, which defines equal resulting rankings with a score of 1. The experiment resulted in an average Kendall correlation score of 0.324, a value that may imply noticeable amounts of customisation in the ranking of the web pages. Detected negative coefficients on certain topics indicated an entirely opposite ordinal association. This may be characterized as a predictable outcome as relevant results emerge at the top of the list.
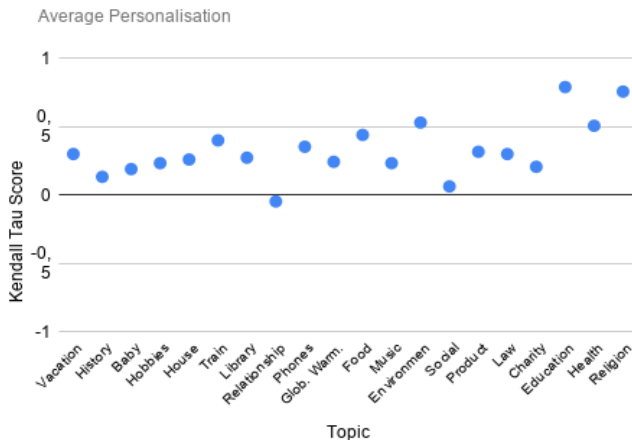


Figure 2: Visual representation of the average personalisation score for the topics tested.

After evaluating the results corresponding to the three specified countries, we were allowed to gain valuable information. The United kingdom averaged a correlation score of 0.43, while the Netherlands a mean score of 0.37. France's Kendall Tau's coefficient

was 0.17, indicating a stronger correlation between search history and consequent results. Despite the civil law system being present in both France and the Netherlands, the data were less personalised on the latter, while signifying a comparable degree of customisation to the UK.

Figure 2 provides insight regarding topics that displayed considerable change in rankings. The queries applied on the topics of education, religion, environmental protection and health lacked signs of customisation. This may be due to the neutral nature of such subjects as opposed to others, such as history and global warming. Based on the experiment's results, we may deduce the importance of certain topics when initializing user profiles. For instance, the topic of 'Baby' included secondary searches related to child care and thus, the search engine would have essential information to aid its user, i.e. the child's age.

Location played an important role in determining the amount of customization for each query. Some topics displayed dissimilar Kendall values amongst countries. By taking a look into the retrieved queries in subject of 'Vacations', our team was able to perceive great differences in the popularity of certain travel websites, where two main sites dominated much of the rankings of the UK and Netherlands. This occurrence was repeated on the subject of hobbies, showing the influence and popularity of particular websites within a country. Further on, on the topic of global warming, the query 'carbon dioxide' initially returned informative results on the gas when searched on a blank profile. If a sub-query related to the greenhouse effect was posed beforehand, results changed to adapt to the subject of emission reduction of harmful gases. France's results in this area were more localised, returning informational articles about carbon emissions and news specific to the country despite the queries being of a generalized nature. This was seen occurring in many of the rest of the topics.

### 3.2 Carry-Over Effects

The following plots are representations of the degree of carry over effects. Kendall Tau correlation was used to compare the queries of all combination pairs of the three countries. Blue bars represent the similarity of two lists of results obtained from blank profiles, from both countries. The yellow bars on the other hand show the comparisons of two lists of results, one for each country, obtained with profiles initialised with a 1 query search history.
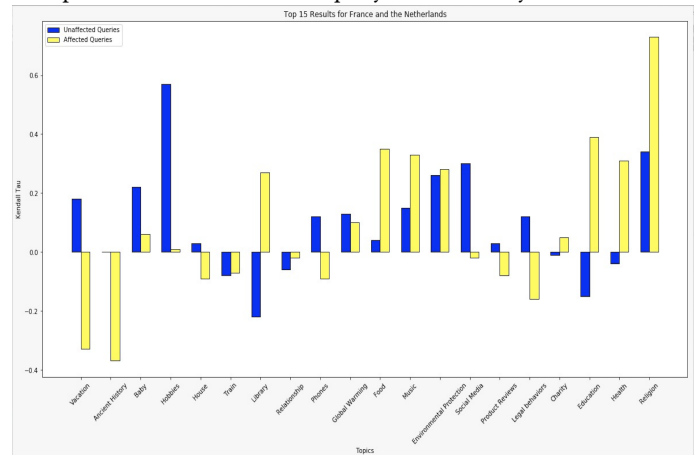
**Figure 3: Top 15 results for comparison on France and the Netherlands**
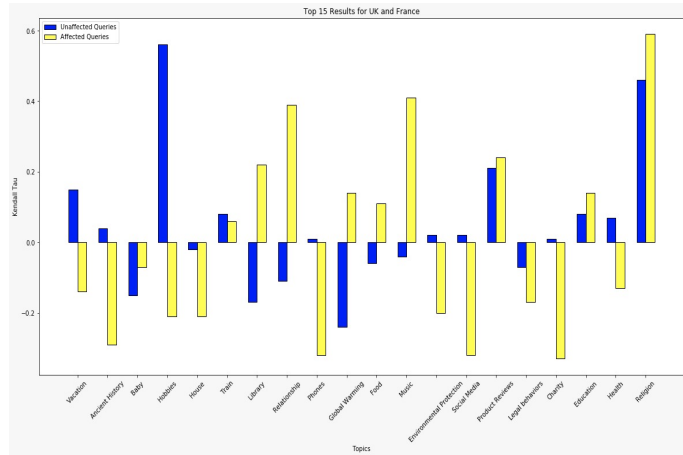


**Figure 4: Top 15 results for comparison on the UK and France**
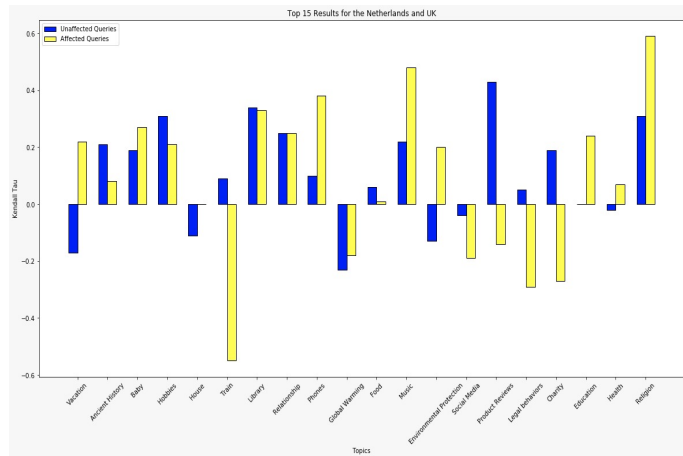


**Figure 5: Top 15 results for comparison on the Netherlands and the UK**

Upon examination of Figures 3-5, we notice a disparity between query scores for both profiles. By taking a closer look into the nature of the most affected topics, we may find out these contain primary queries which would depend on location. For example, the primary query, 'cheap airplane tickets', affected by its secondary query on the topic of vacation, 'vacation destinations', resulted in service sites for local airlines and thus, presented a drop in correlation. This phenomenon is repeated for multiple topics, where the more specific primary query brought local news, blogs and advertisements, in an attempt to retrieve the most relevant websites. The comparison of the Netherlands and the UK in figure 3, shows that citizens within this pair of countries experience the least amount of change between queries. Carry-over effects where most noticeable when comparing to France in figures 3 and 4, demonstrating a greater impact of search personalisation within this country.

## 3.3 Resulting Inferences

Taking into account the outcome of this experiment, we may conclude the user's location plays a crucial role in the personalisation of Google results. However, this feature impacted personalisation within three different countries in different ways, the most prominent being France. The observations deriving from this research lead us to infer that a country's culture and the popularity distribution of specific websites plays as important role in profile customisation as the location of services. Last but not the least, it is very likely the tested countries may be subject to filter bubbles since, even though the tested countries fall within the European Union's jurisdiction, the degrees of carry-over effects were contrasting.

## 3.4 Future research

Determining an appropriate time delay between two queries was considered a preparation step for this project. However, upon discovering the high degree of variances that occur within a 20 minute time frame, our team grew interested in the possibility of disparities in page rank update time across different locations. This could provide an insight into the speed at which filter bubbles may begin to form and how customisation is affected by regular or casual users of a search engine.

## 3.5 Material

Github repsitory: https://github.com/StergiosMorakis/GoogleRetrieval

## REFERENCES

[1] Monica Bianchini, Marco Gori, and Franco Scarselli. "Inside PageRank". In: *ACM Transactions on Internet Technology (TOIT)* 5 (2005), pp. 92–128.
[2] Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. *The Adaptive Web Methods and Strategies of Web Personalization*. 2007.
[3] A. Clement. "NSA Surveillance: Exploring the Geographies of Internet Interception". In: *Proceedings* (2014), pp. 412–425.
[4] Aniko Hannak et al. "Measuring Personalization of Web Search". In: *Proceedings of the 22nd international conference on World Wide Web* (2017), pp. 527–538.
[5] Frederik J. Zuiderveen Borgesius et al. "Should We Worry About Filter Bubbles?" In: *Journal on Internet Regulation* 5 (2016).
[6] Michael J.McCarthy. "Internet monitoring of suicide risk in the population". In: *Journal of Affective Disorders* 122 (2010), pp. 277–279.
[7] Chloe Kliman-Silver et al. "Location, Location, Location: The Impact of Geolocation on Web Search Personalization". In: *Proceeding* (2015), pp. 121–127.
[8] 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* (2015).
[9] Jan Schulz. "Jaccard similarity". In: 5 (2008).
[10] Ronayne Sinead. "Information Point:Kendall's Tau". In: *Journal of Clinical Nursing* 10 (2001), pp. 707–715.